Uniwersytet Warszawski
Wydział Matematyki, Informatyki i Mechaniki

Joanna Giemza, Katarzyna Zwierzchowska
Nr albumu: 281605, 277838

Wprowadzenie do modelu regresji logistycznej wraz z przykładem zastosowania w pakiecie statystycznym R do danych o pacjentach po przeszczepie nerki

Praca licencjacka
na kierunku MATEMATYKA
w zakresie JEDNOCZESNYCH STUDIÓW
EKONOMICZNO–MATEMATYCZNYCH

Praca wykonana pod kierunkiem
dra inż. Przemysława Biecka
Instytut Matematyki Stosowanej i Mechaniki

Lipiec 2011

Oświadczenie kierującego pracą

Potwierdzam, że niniejsza praca została przygotowana pod moim kierunkiem i kwalifikuje się do przedstawienia jej w postępowaniu o nadanie tytułu zawodowego.

Data Podpis kierującego pracą

Oświadczenie autora (autorów) pracy

Świadom odpowiedzialności prawnej oświadczam, że niniejsza praca dyplomowa została napisana przeze mnie samodzielnie i nie zawiera treści uzyskanych w sposób niezgodny z obowiązującymi przepisami.

Oświadczam również, że przedstawiona praca nie była wcześniej przedmiotem procedur związanych z uzyskaniem tytułu zawodowego w wyższej uczelni.

Oświadczam ponadto, że niniejsza wersja pracy jest identyczna z załączoną wersją elektroniczną.

Data Podpis autora (autorów) pracy

Streszczenie

W pracy przedstawione jest wprowadzenie do modelu regresji logistycznej i przykład zastosowania do rzeczywistych danych medycznych. Praca składa się z trzech części. Pierwsza jest częścią teoretyczną, w drugiej przedstawione są funkcje pakietu R służące do budowy modelu regresji logistycznej, a w trzeciej przy użyciu danych rzeczywistych modelowane jest prawdopodobieństwo prawidłowego funkcjonowania nerki u pacjentów po 24 miesiącach od przeszczepu.

Słowa kluczowe

uogólniony model liniowy, regresja liniowa, regresja logistyczna, statystyka Walda, szansa, iloraz szans

Dziedzina pracy (kody wg programu Socrates-Erasmus)

11.2 Statystyka

Klasyfikacja tematyczna

62J12, 92C50

Tytuł pracy w języku angielskim

Introduction to logistic regression with an application in statistical package R to data on kidney transplant results

Spis treści

Spis rysunków

Spis tabel

Udział w przygotowaniu pracy

Niniejsza praca została napisana przez dwie osoby: Joannę Giemzę oraz Katarzynę Zwierzchowską. Praca była tworzona wspólnie, w związku z czym przypisanie poszczególnym fragmentom treści tylko jednej z autorek nie jest możliwe. Można jednak wskazać fragmenty, na których ostateczną postać jedna z osób miała większy wpływ. Dla Katarzyny Zwierzchowskiej są to: podrozdział 1.1, punkty 1.2.1, 1.3.1, podrozdziały 3.5, 3.6, 3.7; dla Joanny Giemzy: punkty 1.2.2, 1.3.2, 1.3.3, 1.3.4, 1.3.5, podrozdziały 3.1, 3.2, 3.3. Wprowadzenie, rozdział 2, podsumowanie oraz dodatek A były pisane wspólnie.

Wprowadzenie

Opis zależności między czynnikami chorobotwórczymi zyskuje coraz większą popularność w środowisku medycznym. Lekarze coraz częściej sięgają po badania naukowe, oparte właśnie na statystycznym modelowaniu takich zależności. Sprzyja temu zarówno rozwój metod statystycznych, jak i nauk medycznych, oferujących narzędzia, dzięki którym przeprowadzanie takiej oceny jest możliwe. Jednocześnie rozwój chorób cywilizacyjnych skłania do coraz głębszej analizy czynników je powodujących. Celem niniejszej pracy jest przybliżenie zagadnienia modelu regresji logistycznej, często wykorzystywanego w pracach medycznych. Jest to bardzo przydatne narzędzie do modelowania zależności między binarną zmienną objaśnianą a zmiennymi objaśniającymi. Z uwagi na łatwą interpretację współczynników modelu coraz częściej stosowane nie tylko w badaniach epidemiologicznych, ale również ekologicznych, finansowych czy biologicznych.

Praca składa się z trzech części. W rozdziale pierwszym przedstawiona zostanie teoria modeli regresji logistycznej. Na początku zdefiniujemy inny model regresji – regresję liniową, następnie przybliżymy krótko teorię uogólnionych modeli liniowych (klasy modeli, która zawiera m.in. regresję liniową i logistyczną). Dużo uwagi poświęcimy regresji logistycznej, podając szczegóły estymacji, wnioskowania statystycznego oraz interpretacji wyników modelu w terminach szans. Drugi rozdział zawierać będzie opis podstawowych i najbardziej przydatnych funkcji w pakiecie R służących do analizy danych przy pomocy uogólnionych modeli liniowych. W ostatnim rozdziale przedstawimy analizę danych rzeczywistych dotyczących pacjentów po przeszczepie nerki. Modelowanym zagadnieniem będzie prawdopodobieństwo powodzenia przeszczepu.

Rozdział 1

Teoria

1.1. Model regresji

Ogólna postać modelu, którym będziemy się zajmować, dana jest wzorem:

$$Y|X \sim F(\theta),$$

$$E(Y|X) = f(X, \beta).$$

Przez Y oznaczamy zmienną objaśnianą (zależną), a $X = ((X_1, \ldots, X_k)$ to wektor zmiennych objaśniających (niezależnych). Regresja jest metodą pozwalającą na opisanie związku pomiędzy wielkościami występującymi w danych. Wykorzystując tę wiedzę, możemy przewidywać nieznane wartości zmiennych na podstawie znanych wartości innych. Problemem, który będzie nas interesował, jest opisanie wartości oczekiwanej zmiennej Y za pomocą zmiennych X, co sprowadza się do wyznaczenia parametrów β danego modelu.

1.1.1. Regresja liniowa

W modelu regresji liniowej wartość oczekiwana zmiennej Y opisana jest jako liniowa kombinacja zmiennych X_i.

Definicja 1.1.1 *Regresja liniowa dana jest równaniem:*

$$E(Y|X) = \beta_0 + \beta_1 X_1 + ... + \beta_n X_n + \epsilon = X^T \beta.$$

ϵ jest to błąd losowy o rozkładzie $N(0, \sigma^2)$.

Model liniowy stosuje się, gdy spełnione są następujące założenia:

- zmienne objaśniające są liniowo niezależne,
- zmienne ϵ są niezależne o rozkładzie $N(0, \sigma^2)$,
- wartość oczekiwana Y wyraża się przez liniową kombinację zmiennych X.

W przeciwnym przypadku standardowe metody wnioskowania statystycznego dla regresji liniowej mogą być niepoprawne. Ponadto model liniowy może być stosowany jedynie dla zmiennej Y, która jest zmienną ilościową i ma rozkład normalny.

W kolejnym rozdziale na podstawie [2] opiszemy szerszą klasę modeli, uogólnionych modeli liniowych, które można stosować nie tylko do zmiennych objaśnianych o rozkładzie normalnym, ale także na przykład do zmiennych jakościowych lub binarnych.

1.2. Uogólniony model liniowy

1.2.1. Definicja uogólnionego modelu liniowego

Uogólniony model liniowy jest definiowany poprzez określenie dwóch elementów:

- rozkładu zmiennej objaśnianej, należącego do naturalnej wykładniczej rodziny rozkładów,
- funkcji łączącej, opisującej związek wartości oczekiwanej zmiennej objaśnianej i kombinacji liniowej zmiennych objaśniających.

W niniejszym podrozdziale definiujemy pojęcia naturalnej rodziny wykładniczej [1] i funkcji łączącej, do każdego podając kilka przykładów.

Definicja 1.2.1 *Rodzina rozkładów prawdopodobieństwa nazywa się naturalną rodziną wykładniczą, jeżeli każdy należący do niej rozkład ma gęstość postaci*

$$f(y|\theta,\phi) = \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right], \tag{1.1}$$

gdzie:
θ *to parametr kanoniczny,*
ϕ *- parametr dyspersji.*

Poszczególne elementy naturalnej rodziny wykładniczej wyróżnia się poprzez sprecyzowanie funkcji a, b i c.

Oto przykłady rozkładów należących do naturalnej rodziny wykładniczej:

1. Rozkład normalny.

   Wstawiając:

$$\begin{cases} \theta & = & \mu, \\ \phi & = & \sigma^2, \\ a(\phi) & = & \phi, \\ b(\theta) & = & \frac{\theta^2}{2}, \\ c(y,\phi) & = & -\frac{\frac{y^2}{\phi} - \log(2\pi\phi)}{2} \end{cases}$$

   do wzoru (1.1), otrzymujemy gęstość rozkładu $N(\mu,\sigma^2)$.

$$\begin{aligned} f(y|\theta,\phi) &= exp\left[\frac{y\mu - \frac{\theta^2}{2}}{\sigma^2} - \left(\frac{y^2}{\phi} - \log(2\pi\phi)\right)/2\right] \\ &= \frac{1}{2\pi\sigma} exp\left[-\left(\frac{y-\mu}{\sqrt{2}\sigma}\right)^2\right]. \end{aligned} \tag{1.2}$$

[1] Naturalna rodzina wykładnicza jest szczególnym przypadkiem szerszej klasy rozkładów – rodziny wykładniczej.

2. Rozkład Poissona uzyskujemy dla:

$$\begin{cases} \theta & = & \log(\mu), \\ \phi & \equiv & 1, \\ a(\phi) & = & 1, \\ b(\theta) & = & exp(\theta), \\ c(y,\phi) & = & -\log(y!). \end{cases}$$

Wówczas

$$f(y|\theta,\phi) = \exp[y\log(\mu) - \mu - \log(y!)] = \frac{e^{-\mu}\mu^y}{y!}. \tag{1.3}$$

3. Rozkład dwumianowy uzyskujemy dla:

$$\begin{cases} \theta & = & \log(\frac{\mu}{1-\mu}), \\ \phi & \equiv & 1, \\ a(\phi) & = & 1, \\ b(\theta) & = & -n\log(1-\mu) \quad = \quad n\log(1+\exp(\theta)), \\ c(y|\theta) & = & log\binom{n}{y}. \end{cases}$$

Wówczas

$$\begin{aligned} f(y|\theta,\phi) &= \exp\left[y\log(\mu) + (n-y)\log(1-\mu) + \log\binom{n}{y}\right] \\ &= \binom{n}{y}\mu^y(1-\mu)^{n-y}. \end{aligned} \tag{1.4}$$

Kolejnym pojęciem wymagającym zdefiniowania jest *funkcja łącząca*.

Definicja 1.2.2 *Funkcja łącząca $g(\mu)$ opisuje związek między wartością oczekiwaną zmiennej objaśnianej $EY = \mu$ a modelem liniowym:*

$$g(\mu) = x^T\beta.$$

Przykłady:

- Dla rozkładu normalnego popularną funkcją łaczącą jest identyczność $g(\mu) = \mu$.
- Dla rozkładu Poissona standardowym wyborem funkcji łączącej jest $e^{g(\mu)} = \mu$. Wtedy $g(\mu) = \log(\mu)$, co gwarantuje, że $\mu > 0$.
- Dla rozkładu dwumianowego z prawdopodobieństwem sukcesu p, gdzie $0 < p < 1$, funkcja łącząca musi być monotoniczna i spełniać warunek $0 \leqslant g^{-1}(\mu) \leqslant 1$. Często wybierane są:
  - logit: $g(\mu) = \log(\frac{\mu}{1-\mu})$,
  - probit: $g(\mu) = \Phi^{-1}(\mu)$, gdzie Φ jest dystrybuantą rozkładu $N(0,1)$,
  - *complementary log-log*: $g(\mu) = \log(-\log(1-\mu))$.

Często naturalnym wyborem funkcji łączącej jest *kanoniczna funkcja łącząca*.

Definicja 1.2.3 *Kanoniczna funkcja łącząca to funkcja g spełniająca:*

$$g(\mu) = \theta,$$

gdzie θ jest kanonicznym parametrem rodziny wykładniczej.

Przykłady kanonicznych funkcji łączących przedstawione są w tabeli 1.1.

rozkład zmiennej Y	funkcja łącząca
Normalny	$g(\mu) = \mu$
Poissona	$g(\mu) = \log \mu$
Dwumianowy	$g(\mu) = \log(\frac{\mu}{1-\mu})$

Tabela 1.1: Kanoniczne funkcje łączące dla wybranych uogólnionych modeli liniowych

1.2.2. Estymacja parametrów w uogólnionym modelu liniowym

Parametry β_i możemy estymować używając metody największej wiarygodności. W przypadku modelu liniowego problem maksymalizacji funkcji wiarygodności ma rozwiązanie analityczne:

$$\hat{\beta} = (X^T X)^{-1} X^T Y.$$

Jednak dla innych przypadków uogólnionego modelu liniowego rozwiązanie analityczne nie musi istnieć. Przybliżone rozwiązanie uzyskuje się za pomocą algorytmów numerycznych, np. stosowanego przez pakiet R algorytmu *Fisher Scoring* lub algorytmu *ridge stabilized Newton-Raphson*. Opis obu tych algorytmów znajduje się w [9, str. 620-621], ponadto w [2, str. 129-131] algorytm *Fisher Scoring* jest przedstawiony jako analogiczny do iteracyjnie ważonej metody najmniejszych kwadratów (IRWLS) dla regresji liniowej.

1.3. Regresja logistyczna

Odtąd ograniczymy rozważania teoretyczne do szczególnego przypadku uogólnionych modeli liniowych - modelu regresji logistycznej. Opieramy się na [2] oraz [3].

Definicja 1.3.1 *Ogólna postać* regresji logistycznej *jest następująca:*

$$Y \sim B(1, p), \tag{1.5}$$

$$p = E(Y|X) = \frac{\exp(\beta X)}{1 + \exp(\beta X)}, \tag{1.6}$$

gdzie $B(1, p)$ jest to rozkład dwumianowy z prawdopodobieństwem sukcesu p.

Ostatnia równość zakłada wybór kanonicznej funkcji łączącej - logitu. Modelowanie p za pomocą logitu umożliwia wygodną interpretację wyników regresji logistycznej w terminach szans, którym poświęcamy następny podrozdział.

1.3.1. Szansa

Szansa (ang. *odds*) jest to funkcja prawdopodobieństwa. Zamiast wyliczania klasycznego prawdopodobieństwa, czyli stosunku liczby sukcesów do liczby wszystkich prób, wyliczamy stosunek prawdopodobieństwa sukcesu do prawdopodobieństwa porażki. Niech o oznacza szansę oraz p prawdopodobieństwo sukcesu. Wtedy:

$$o = \frac{p}{1-p} \tag{1.7}$$

oraz

$$p = \frac{o}{1+o}. \tag{1.8}$$

Prawdopodobieństwo zdarzenia $p \in (0,1)$, więc szansa przyjmuje wartości z zakresu $(0,\infty)$, a jej logarytm- wartości z zakresu $(-\infty,\infty)$.

Na przykład, rozważmy hipotetyczne zdarzenia: A – w próbce 100 osób palących paierosy 90 zachorowało na nowotwór i B- spośród 100 osób niepalących zachorowało 20. Wtedy $o(A) = \frac{90}{10} = 9$, $o(B) = \frac{20}{80} = 0.25$ Oznacza to, że prawdopodobieństwo pojawienia się zdarzenia A jest dziewięć razy większe niż prawdopodobieństwo jego niepojawienia się wśród osób palących. Możemy też powiedzieć, że szansa wystąpienia przypadku A wynosi 9:1 (analogicznie interpretujemy zdarzenie B).

Regresja logistyczna opiera się właśnie na takim sposobie wyrażania prawdopodobieństwa. W modelu regresji logistycznej dla jednej zmiennej objaśniającej X_1 szansa wynosi:

$$\frac{P(X)}{1-P(X)} = \exp(\beta_0 + \beta_1 X_1). \tag{1.9}$$

Natomiast logarytm szansy wynosi:

$$\log \frac{P(X)}{1-P(X)} = \beta_0 + \beta_1 X_1. \tag{1.10}$$

Logarytm szansy jest liniowo zależny od zmiennej objaśniającej X_1, dzięki czemu β_1 możemy łatwo interpretować. Współczynnik ten mówi nam o zmianie wartości logarytmu szansy związanej ze zmianą o jednostkę czynnika opisanego przez X_1. Przechodząc z logarytmu szansy na terminy szansy, e^{β_1} to relatywna zmiana możliwości wystąpienia zdarzenia pod wpływem czynnika opisanego przez zmienną X_1.

- Jeżeli $e^{\beta_1} > 1$, to czynnik opisywany przez zmienną X_1 ma stymulujący wpływ na wystąpienia badanego zjawiska.
- Jeżeli $e^{\beta_1} < 1$, to dany czynnik działa ograniczająco.
- Jeżeli $e^{\beta_1} = 1$, to czynnik nie ma wpływu na opisywane zdarzenie.

Iloraz szans (ang. *odds ratio*) stosuje się w przypadku porównywania dwóch klas obserwacji. Jest to iloraz szans, że dane zdarzenie zajdzie w pierwszej grupie elementów, oraz że zajdzie ono również w drugiej. Opisane jest wzorem:

$$OR = \frac{p_1}{1-p_1}\frac{1-p_2}{p_2} = \frac{p_1(1-p_2)}{p_2(1-p_1)}, \tag{1.11}$$

gdzie p_i oznacza prawdopodobieństwo zdarzenia w i-tej klasie obserwacji. Interpretujemy je następująco:

- jeżeli $OR > 1$, to w pierwszej grupie zajście zdarzenia jest bardziej prawdopodobne.
- jeżeli $OR < 1$, to w drugiej grupie zajście zdarzenia jest bardziej prawdopodobne.
- jeżeli $OR = 1$, to w obu klasach obserwacji zdarzenie jest tak samo prawdopodobne.

Na podstawie danych z poprzedniego przykładu obliczymy iloraz szans grupy A do grupy B. Otrzymujemy: $OR = \frac{9}{0.25} = 36$. Zatem szansa rozwoju nowotworu u palaczy jest 36-krotnie większa niż u niepalących osób.

1.3.2. Założenia

Model regresji logistycznej nie wymaga niektórych założeń koniecznych dla regresji liniowej. Wektor zmiennych objaśniających i reszty nie muszą mieć rozkładu normalnego, dopuszczalna jest heteroskedastyczność. Jednak konieczne jest spełnienie kilku innych warunków:

- Zależność między logarytmem szans a wektorem zmiennych objaśniających musi być liniowa[2] (zgodnie z równaniem (1.10)).
- Zmienna objaśniana musi być binarna, gdzie poziom zakodowany jako "1" reprezentuje pożądany wynik (sukces).
- Obserwacje muszą być niezależne – korzystamy z tego wyprowadzając postać funkcji wiarygodności.
- Model musi być dobrze dopasowany, to znaczy zawierać tylko te zmienne objaśniające, które mają wpływ na zmienną objaśnianą, oraz nie pomijać żadnej takiej zmiennej.
- W danych nie może występować silna współliniowość – jest ona źródłem problemów numerycznych.

Ostatnie dwa warunki mają bardziej charakter wskazówek niż założeń. Nie korzystamy z nich do wyprowadzenia teorii regresji logistycznej, jednak model, który ich nie spełnia, może prowadzić do niepoprawnych wniosków.

1.3.3. Funkcja wiarygodności

Wyprowadzimy postać funkcji wiarygodności L dla regresji logistycznej. Zmienna objaśniana Y jest binarna i dla pojedynczej obserwacji i zachodzi:

$$Y_i|X_i = \begin{cases} 1 & \text{z prawdopodobieństwem } p(X_i) \\ 0 & \text{z prawdopodobieństwem } 1 - p(X_i). \end{cases}$$

Stąd

$$L(X_i, \beta) = P(Y_i = 1|X_i)^{Y_i} \cdot P(Y_i = 0|X_i)^{1-Y_i} = p(X_i)^{Y_i}[1 - p(X_i)]^{1-Y_i}, \tag{1.12}$$

gdzie wektor estymowanych parametrów β jest uwikłany w funkcji p, zgodnie ze wzorem (1.6).

Funkcja wiarygodności dla n obserwacji, pod założeniem ich niezależności, jest produktem funkcji wiarygodności dla pojedynczych obserwacji (1.12):

$$L(X_1, \ldots, X_n, \beta) = \prod_{i=1}^{n} p(X_i)^{Y_i}[1 - p(X_i)]^{1-Y_i}. \tag{1.13}$$

Funkcję wiarygodności wykorzystuje się do estymacji parametrów β metodą największej wiarygodności[3] oraz do testowania hipotez statystycznych. Często funkcję wiarygodności zastępuje się jej logarytmem[4], z uwagi na łatwiejszą obliczeniowo postać:

$$\log L(X_1, \ldots, X_n, \beta) = \sum_{i=1}^{n} (Y_i \log p(X_i) + (1 - Y_i) \log(1 - p(X_i))). \tag{1.14}$$

[2] Analogicznym założeniem w regresji liniowej jest liniowa zależność między zmienną objaśnianą a zmiennymi objaśniającymi

[3] Model regresji logistycznej jest jednym z tych uogólnionych modeli liniowych, dla których problem maksymalizacji funkcji wiarygodności nie ma rozwiązania analitycznego.

[4] Nie zmienia to wyników estymacji MNW, gdyż logarytm jest funkcją monotoniczną.

1.3.4. Testowanie hipotez w modelu regresji logistycznej

W tym punkcie pokażemy, w jaki sposób testowane są hipotezy w modelu regresji logistycznej. W szczególności hipotezy o istotności statystycznej zmiennych, czyli hipotezy dotyczące kwestii, czy model, która zawiera pewną zmienną, dostarcza istotnie więcej informacji o zmiennej objaśnianej od modelu bez tej zmiennej. Testowanie takiej hipotezy opiera się na porównaniu wartości zaobserwowanych zmiennej objaśnianej Y z jej wartościami dopasowanymi $\hat{Y}$ przez dwa modele, jeden z interesującą nas zmienną objaśniającą, drugi bez niej.

Wprowadzimy pojęcie *modelu nasyconego*. Dotyczy ono nie tylko regresji logistycznej, ale całej klasy uogólnionych modeli liniowych.

Definicja 1.3.2 *Model nasycony (ang. saturated model) to model o liczbie parametrów równej liczbie obserwacji.*

Przykładowo, dla zbioru danych zawierającego dwie obserwacje modelem pełnym jest $g(EY) = \beta_0 + \beta_1 X_1$. Pojęcie modelu pełnego umożliwia inną interpretację wartości zaobserwowanych zmiennej objaśnianej - jako wartości dopasowanych z modelu pełnego dla danego zbioru danych.

Testowanie istotności zmiennej objaśniającej korzysta ze statystyki dewiancji *(ang. deviance)* D:

$$D = -2\log\left[\frac{\text{wartość funkcji wiarygodności estymowanego modelu}}{\text{wartość funkcji wiarygodności modelu pełnego}}\right]. \tag{1.15}$$

Pomnożony przez (-2) logarytm ilorazu wiarygodności ma znany rozkład, nadaje się więc do testowania hipotez statystycznych. Oparte na nim testy to *testy ilorazu wiarygodności*.

Wstawiając do postaci D z (1.15) wyrażenia z (1.14) otrzymujemy:

$$D = -2\sum_{i=1}^{n}\left[Y_i \log\left(\frac{p(X_i)}{Y_i}\right) + (1-Y_i)\log\left(\frac{1-p(X_i)}{1-Y_i}\right)\right]. \tag{1.16}$$

W modelu regresji logistycznej wartość funkcji wiarygodności dla modelu pełnego wynosi 1, co można pokazać, wstawiając $p(X_i) = Y_i$ (własność modelu pełnego) do (1.13):

$$L(\text{model pełny}) = \prod_{i=1}^{n} Y_i^{Y_i}[1-Y_i)]^{1-Y_i} = 1. \tag{1.17}$$

Stąd i z (1.15) otrzymujemy:

$$D = -2\log(\text{wartość funkcji wiarygodności estymowanego modelu}). \tag{1.18}$$

Ocena istotności zmiennej objaśnianej przeprowadzana jest na podstawie statystyki G, mierzącej zmianę wartości dewiancji w wyniku dodania do modelu zmiennej, której istotność badamy (oznaczmy ją przez X_*):

$$\begin{aligned} G &= D(\text{model bez zmiennej } X_*) - D(\text{model ze zmienną } X_*) \\ &= -2\log L(\text{model bez zmiennej } X_*) + 2\log L(\text{model ze zmienną } X_*) \\ &= -2\log\left[\frac{L(\text{model bez zmiennej } X_*)}{L(\text{model ze zmienną } X_*)}\right]. \end{aligned} \tag{1.19}$$

Dla dostatecznie dużej liczby obserwacji n i przy założeniu hipotezy zerowej o nieistotności zmiennej X_* ($\beta_* = 0$), statystyka G ma rozkład χ^2 z jednym stopniem swobody.

Istnieje uogólnienie powyższej metody na przypadek testowania istotności wielu zmiennych objaśniających. Niech $X_* = (X_{*1}, \ldots, X_{*k})$ oznacza wektor k zmiennych objaśniających, których istotność testujemy. Statystyka G jest analogiczna do (1.19):

$$G = -2 \log \left[\frac{L(\text{model bez zmiennych } X_*)}{L(\text{model zawierający zmienne } X_*)} \right], \tag{1.20}$$

oraz ma rozkład χ^2 z k stopniami swobody.

Warto zauważyć, że statystyki D i G mają swoje odpowiedniki w regresji liniowej. Dewiancja w regresji logistycznej pełni tę samą rolę, co suma kwadratów reszt w modelu liniowym, natomiast statystyka G odpowiada statystyce testu F.

Oprócz testu ilorazu wiarygodności, istnieją dwie alternatywne metody testowania istotności zmiennych objaśniających: test Walda i test Score, które poniżej tylko krótko przedstawimy. Tak jak test ilorazu wiarygodności, wymagają założenia dostatecznie dużej liczby obserwacji n. Test Walda otrzymujemy dzieląc oszacowanie parametru przy zmiennej X_* przez błąd standardowy[5] tego oszacowania (oznaczany przez SE):

$$W = \frac{\hat{\beta}_*}{SE(\hat{\beta}_*)}. \tag{1.21}$$

Przy założeniu hipotezy zerowej ($\beta_* = 0$) W ma asymptotycznie rozkład $N(0, 1)$. Test Score opiera się natomiast na statystykach otrzymanych z pochodnych logarytmu funkcji wiarygodności i nie wymaga obliczania estymatorów MNW parametrów β. Terminem *score* określa się pochodną logarytmu funkcji wiarygodności:

$$U(\beta) = \frac{\partial \log L(\beta|x)}{\partial \beta}. \tag{1.22}$$

Statystyka testowa dla hipotezy zerowej $\beta_* = 0$ ma asymptotycznie rozkład $\chi^2(1)$ i wynosi:

$$S = U(0)^2 I(0)^{-1}, \tag{1.23}$$

gdzie $U(0)$ to *score* względem parametru β_*, natomiast I jest informacją Fishera.

1.3.5. Wyznaczanie przedziałów ufności

Przedziały ufności dla oszacowań współczynników w modelu regresji logistycznej konstruuje się na podstawie statystyki testu Walda. Korzystając z faktu, że W ma asymptotyczny rozkład standardowy normalny:

$$W = \frac{\hat{\beta}}{SE(\hat{\beta})} \sim N(0, 1),$$

krańce przedziału ufności na poziomie istotności $(100 - \alpha)\%$ dla oszacowania parametru β wynoszą:

$$\hat{\beta} \pm z_{1-\alpha/2} SE(\hat{\beta}),$$

gdzie $z_{1-\alpha/2}$ jest kwantylem rozkładu $N(0, 1)$ rzędu $1 - \alpha/2$.

Na tym zakończymy rozdział teoretyczny. Zawarłyśmy w nim kwestie uznane przez nas za najistotniejsze w temacie teorii regresji logistycznej. Informacje te mogą jednak być niewystarczające do zastosowania metody w praktyce. W kolejnych rozdziałach, gdzie przedstawimy

[5] Błąd standardowy to pierwiastek z wariancji. Szczegóły estymacji wariancji oszacowań współczynników w modelu regresji logistycznej można znaleźć w [3].

możliwości pakietu R w zakresie uogólnionych modeli liniowych i przykładowe zastosowanie metody regresji logistycznej do rzeczywistych danych, pojawi się kilka zagadnień, których nie poruszałyśmy w tym rozdziale. Będą one dotyczyły budowy i diagnostyki modelu. Ich dokładny opis oraz głębsze opracowanie problemu zastosowania regresji logistycznej można znaleźć w [3], [4] i [6].

Rozdział 2

Model regresji logistycznej w pakiecie R

W rozdziale tym przedstawimy funkcje służące do budowy i analizy modelu regresji logistycznej w pakiecie R. Źródłem informacji były dla nas [1] oraz [8]. Do przykładowych komend użyjemy danych opisanych i analizowanych w rozdziale 3.

2.1. Funkcja glm()

Funkcja `glm()` służy do dopasowania uogólnionych modeli liniowych, w szczególności modelu regresji logistycznej. Wyznacza m.in. oszacowania współczynników β oraz wylicza wartości reszt. Określamy pełną implementację:

```
glm(formula, family = gaussian, data, weights, subset,
    na.action, start = NULL, etastart, mustart, offset,
    control = list(...), model = TRUE, method = "glm.fit",
    x = FALSE, y = TRUE, contrasts = NULL, ...)
```

`formula`	opis zależności między zmienną objaśnianą a zmiennymi objaśniającymi
`family`	wskazanie rozkładu zmiennej objaśnianej
`data`	wskazanie ramki danych zawierającej zmienne modelu
`weights`	wskazanie wektora wag dla obserwacji
`subset`	możliwość określenia podzbioru obserwacji
`na.action`	określenie, co zrobić, gdy brakuje wartości w danych
`start`	wskazanie początkowych wartości parametrów β
`etastart`	wskazanie początkowych wartości wektora zmiennych objaśniających
`mustart`	wskazanie początkowych wartości wektora średnich
`offset`	podanie znanych wartości współczynników β
`control`	parametry służące do kontroli procesu dopasowywania
`model`	czy w wyniku ma być informacja o użytej formule
`method`	czy w wyniku ma być informacja o użytej metodzie estymacji
`x,y`	czy w wyniku ma być informacja o wartościach X, Y
`contrasts`	wskazanie sposobu kodowania zmiennych jakościowych

Tabela 2.1: Opis argumentów funkcji glm()

W tabeli 2.1 przedstawiamy opis argumentów funkcji `glm()`.Dalsze możliwości funkcji `glm()` będziemy ilustrować określonym dwoma argumentami (`formula` i `family`) przykładem modelu regresji logistycznej. Ogólny schemat komendy `glm()`dla takiego modelu jest następujący:

```
nazwa_modelu <- glm(zmienna_objaśniana ~ zmienna_objaśniająca_1 +
zmienna_objaśniająca_2 + ... + zmienna_objaśniająca_n, family = "binomial"),
```

natomiast dla przykładowych danych:

```
model1 <- glm(MDRD24 ~ wiek.biorcy + wiek.dawcy + czas.zimnego.niedokrwienia,
family = binomial()).
```

Wpisując `print(nazwa_modelu)` lub po prostu nazwę modelu, otrzymujemy wyniki.

```
model1

Call:  glm(formula = MDRD24 ~ wiek.biorcy + wiek.dawcy
+ czas.zimnego.niedokrwienia,      family = binomial())

Coefficients:
               (Intercept)                  wiek.biorcy
                  1.231068                     0.005433
                wiek.dawcy  czas.zimnego.niedokrwienia
                 -0.069015                     0.027310

Degrees of Freedom: 224 Total (i.e. Null);  221 Residual
Null Deviance:      282.1
Residual Deviance: 249.2        AIC: 257.2
```

Więcej informacji o modelu możemy otrzymać używając funkcji `summary(nazwa_modelu)`.

```
summary(model1)

Call:
glm(formula = MDRD24 ~ wiek.biorcy + wiek.dawcy
+ czas.zimnego.niedokrwienia,    family = binomial())

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5929  -0.8159  -0.6118   0.9775   2.1915

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)                1.231068   0.824530   1.493    0.135
wiek.biorcy                0.005433   0.014400   0.377    0.706
wiek.dawcy                -0.069015   0.013210  -5.225 1.75e-07 ***
czas.zimnego.niedokrwienia 0.027310   0.019849   1.376    0.169
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 282.09  on 224  degrees of freedom
Residual deviance: 249.24  on 221  degrees of freedom
AIC: 257.24

Number of Fisher Scoring iterations: 3
```

Otrzymujemy m.in. tabelę `Coefficients`, której pierwsza kolumna stanowi informacje o wyestymowanych współczynnikach modelu. W drugiej kolumnie znajduje się odchylenie standardowe tych współczynników, w następnej wartość statystyki testu Walda. W ostatniej kolumnie znajdziemy p-wartość testu. Hipotezą zerową tego testu jest nieistotność danej zmiennej objaśniającej. Wiersze tabeli odpowiadają zmiennym objaśniającym i wyrazowi wolnemu dopasowanego modelu.

Na obiekcie klasy `summary` możemy zastosować kilka funkcji, niektóre przedstawimy poniżej. Chcąc wyświetlić jedynie tabelę oszacowań współczynników modelu używamy funkcji `$coef` lub `$coefficients`. Stosując funkcję `$cov.unscaled` dostaniemy macierz kowariancji dla oszacowań współczynników.

```
 summary(model1)$coefficients
                              Estimate Std. Error    z value     Pr(>|z|)
(Intercept)                1.231067904 0.82453006  1.4930540 1.354231e-01
wiek.biorcy                0.005433414 0.01439977  0.3773264 7.059310e-01
wiek.dawcy                -0.069015487 0.01320983 -5.2245558 1.745737e-07
czas.zimnego.niedokrwienia 0.027310355 0.01984903  1.3759036 1.688515e-01
```

Następnie opiszemy funkcje, które działają na obiekcie klasy `glm`. Wektor oszacowań współczynników uzyskamy poprzez `$coef` lub `$coefficients`.

```
model1$coefficients

(Intercept)  wiek.biorcy   wiek.dawcy   czas.zimnego.niedokrwienia
1.231067904  0.005433414  -0.069015487      0.027310355
```

Wektor reszt możemy otrzymać używając funkcji `$residuals`. Domyślnie są to reszty dewiancji, aby wskazać inny typ reszt używamy `residuals(nazwa_modelu,type=...)`. Oddzielne funkcje służą do uzyskania reszt standaryzowanych lub studentyzowanych, odpowiednio: `$rstandard` i `$rstudent`. Komenda `$df.residual` wyświetla liczbę stopni swobody dla reszt.

```
model1$df.resid
[1] 221
```

Kolejną funkcją, którą możemy zastosować jest `$fitted.values`, która wypisuje dopasowane wartości, odpowiadające estymowanym prawdopodobieństwom $\hat{p}$. Natomiast funkcja `linear.predictors` wylicza oszacowania $ln(\frac{\hat{p}}{1-\hat{p}})$. Funkcją `$iter` sprawdzamy, ile dokonano iteracji w algorytmie Fishera, natomiast `$converged` określa, dlaczego algorytm zakończył iterowanie (TRUE – jeśli znalazł maksimum, FALSE – jeśli wykonał więcej iteracji niż podano w założeniach). Funkcją `$family` sprawdzimy, którą funkcję łączącą użyto.

```
model1$iter
[1] 3
```

```
glm$converged
[1] TRUE

glm$family
Family: binomial
Link function: logit
```

2.2. Inne funkcje przydatne w analizie danych

W tym podrozdziale zaprezentujemy kilka funkcji użytecznych w analizie danych, które nie są funkcjami na obiektach `glm` lub `summary`, choć ich argumentem zazwyczaj jest nazwa modelu. Przedziały ufności dla oszacowań współczynników otrzymamy wpisując polecenie `confint(nazwa_modelu)`.

```
confint(model1)

Waiting for profiling to be done...
                                 2.5 %      97.5 %
(Intercept)                -0.37408042  2.87166905
wiek.biorcy                -0.02268880  0.03400662
wiek.dawcy                 -0.09593431 -0.04394255
czas.zimnego.niedokrwienia -0.01143037  0.06670495
```

W celu dobrania odpowiednich zmiennych objaśniających modelu możemy posłużyć się funkcją

```
step(nazwa_modelu, direction = c("both", "backward", "forward"), steps =
     1000, k = 2)
```

Funkcja ta znajduje najlepiej dopasowany model do naszych danych. Stosowana metoda budowy modelu jest określona w zależności od wyboru parametru `direction`. Jego możliwe wartości to `backward`, `forward`, `both`. Metoda `backward` oznacza, że usuwane są najmniej istotne zmienne z modelu zawierającego wszystkie zmienne objaśniające dopóki wszystkie zmienne w modelu będą istotne. `Forward` określa metodę dodawania najbardziej istotnych zmiennych do modelu zawierającego tylko wyraz wolny. Natomiast `both` oznacza metodę, która do modelu zawierajacego tylko wyraz wolny, dodajemy zmienną istotną posiadającą najmniejszą p-value, a następnie usuwamy zmienną nieistotną z największą p-value. Kroki te są powtarzane aż model przestaje ulegać zmianie. Istotność zmiennych określana jest na podstawie jednego z kryteriów, które możemy wybrać za pomocą parametru `k`. Domyślnym kryterium jest Akaike (AIC), natomiast wybierając `k=log(n)` zmieniamy kryterium na kryterium Schwartza (BIC), gdzie `n` oznacza liczbę obserwacji. Parametrem `steps` określamy maksymalną liczbę kroków.

```
step(model1, direction="backward")

Start:  AIC=257.24
MDRD24 ~ wiek.biorcy + wiek.dawcy + czas.zimnego.niedokrwienia

                            Df Deviance    AIC
- wiek.biorcy                1   249.38 255.38
```

```
- czas.zimnego.niedokrwienia  1   251.15 257.15
<none>                            249.24 257.24
- wiek.dawcy                  1   280.77 286.77

Step:  AIC=255.38
MDRD24 ~ wiek.dawcy + czas.zimnego.niedokrwienia

                             Df Deviance    AIC
- czas.zimnego.niedokrwienia  1   251.26 255.26
<none>                            249.38 255.38
- wiek.dawcy                  1   281.56 285.56

Step:  AIC=255.26
MDRD24 ~ wiek.dawcy

             Df Deviance    AIC
<none>            251.26 255.26
- wiek.dawcy  1   282.09 284.09

Call:  glm(formula = MDRD24 ~ wiek.dawcy, family = binomial())

Coefficients:
(Intercept)   wiek.dawcy
    1.95395     -0.06562

Degrees of Freedom: 224 Total (i.e. Null);  223 Residual
Null Deviance:      282.1
Residual Deviance: 251.3        AIC: 255.3
```

Mając dobrany model do danych warto sprawdzić, czy nie występuje problem współliniowości. Funkcja `vif(nazwa_modelu)` wyświetla wektor wartośći współczynnika VIF dla każdej zmiennej objaśniającej. Wymaga ona biblioteki `Design`.

```
library(Design)
vif(model1)
```

```
wiek.biorcy               wiek.dawcy      czas.zimnego.niedokrwienia
1.046534                   1.077083                1.030825
```

W następnym rozdziale, większość z opisanych w tym rozdziale funkcji, wykorzystamy do analizy rzeczywistych danych.

Rozdział 3

Analiza danych rzeczywistych

3.1. Kontekst modelu

Zastosowanie metody regresji logistycznej zilustrujemy na przykładzie rzeczywistych danych medycznych, dotyczących przeszczepu nerki. Transplantacja nerki jest uważana za najlepszą metodę leczenia schyłkowej niewydolności nerek, wykonywana jest jednak stosunkowo rzadko. Jednym z powodów jest niedobór narządów do przeszczepu. Drugim jest ryzyko. Przeszczepiony narząd może zostać odrzucony, dlatego pacjenci muszą stale otrzymywać leki immunosupresyjne. Leki te osłabiają jednak system odpornościowy, utrudniając zwalczanie zakażeń. Niełatwe decyzje o zakwalifikowaniu pacjenta na przeszczep lub oddaniu własnej nerki bliskiej osobie mogą być wspomagane przez modele statystyczne. Zaproponowany w tym rozdziale model regresji logistycznej będzie szacował prawdopodobieństwo powodzenia przeszczepu nerki.

3.2. Opis danych

Dane pochodzą z Kliniki Nefrologii Wrocławskiego Uniwersytetu Medycznego. Zawierają 334 obserwacje i 16 zmiennych. Każda obserwacja dotyczy pojedynczego pacjenta, poddanego transplantacji nerki, i zawiera informacje o czynnikach, które mogły mieć wpływ na funkcjonowanie nerki po przeszczepie, oraz wyniki badań wydolności nerek (współczynnik filtracji kłębuszkowej, obliczany metodą MDRD) w różnych odstępach czasu od zabiegu transplantacji. Mamy do czynienia z danymi ilościowymi (np. wiek dawcy i biorcy, liczba leków na ciśnienie) oraz jakościowymi (schemat terapii, występowanie cukrzycy).

3.3. Modyfikacje danych

Nie możemy wykorzystać wszystkich danych do konstrukcji i estymacji modelu regresji logistycznej. Jednym z powodów jest obecność nieokreślonych wartości, uniemożliwiająca oszacowanie parametrów modelu. Obserwacje z brakującymi danymi trzeba zatem usunąć. Z innych powodów pozbywamy się niektórych zmiennych. Ze względu na cel modelu: szacowanie prawdopodobieństwa powodzenia przeszczepu, zmienną objaśnianą konstruujemy na podstawie jednej ze zmiennych określających wydolność nerek. Wybrałyśmy zmienną MDRD24, określającą współczynnik filtracji kłębuszkowej po 24 miesiącach od przeszczepu, z uwagi na stosunkową niewielką liczbę brakujących wartości (dla 22 pacjentów) oraz długość czasu, który upłynął od chwili transplantacji - stan zdrowia pacjenta po dwóch latach wydaje się być dobrym wskaźnikiem powodzenia zabiegu. Siedem zmiennych dotyczących wyników badania

wydolności nerek w innych niż dwa lata odstępach czasu nie będzie wykorzystanych. Dalszej analizie poddamy zbiór ograniczony do 9 zmiennych i 225 obserwacji. Tabela 3.1 przedstawia fragment zbioru danych, a tabele 3.2 i 3.3 podstawowe statystyki dla każdej zmiennej.

wiek biorcy	wiek dawcy	czas zimnego niedokrwienia	niezgod-ności AB	niezgod-ności DR	schemat terapii	czy cukrzyca	leki na ciśnienie	MDRD 24
50	50	20	2	1	cm	0	2	46
67	18	26	1	1	ca	0	1	79
56	66	23	2	0	ca	0	2	41
48	52	42	3	0	cm	0	1	75
53	51	31	3	0	ca	0	2	47

Tabela 3.1: Fragment analizowanych danych – informacje o pacjentach po przeszczepie nerki.

	wiek biorcy	wiek dawcy	czas zimnego niedokrwienia	MDRD24
Minimum	15.00	14.00	0.50	14.0
1. Kwantyl	34.00	34.00	18.00	42.0
Mediana	43.00	45.00	23.00	51.0
Średnia	41.99	42.88	23.13	52.32
3. Kwantyl	50.00	52.00	28.00	63.0
Maximum	67.00	66.00	44.00	99.0

Tabela 3.2: Kwantyle i średnie dla zmiennych ciągłych modelu.

liczba niezgodności AB	liczba niezgodności DR	czy cukrzyca	liczba leków na ciśnienie	schemat terapii
0: 6	0: 63	0: 172	0: 15	ca: 101
1: 33	1: 152	1: 53	1: 46	cm: 51
2: 99	2: 10		2: 75	tc: 73
3: 72			3: 58	
4: 15			4: 28	
			5: 2	
			6: 1	

Tabela 3.3: Częstości zmiennych dyskretnych modelu.

Prognozowane przez model zdarzenie przeszczepu nerki może zakończyć się "sukcesem" (dobrym stanem zdrowia pacjenta) bądź "porażką" (kiepskim stanem zdrowia pacjenta lub odrzutem przeszczepu). Zmienna objaśniana Y będzie zatem zmienną binarną. Konstruujemy ją ze zmiennej MDRD24 w następujący sposób:

$$Y_i = \begin{cases} 1 & \text{jeżeli } MDRD24_i > k \quad \text{(sukces)}, \\ 0 & \text{jeżeli } MDRD24_i \leqslant k \quad \text{(porażka)}. \end{cases}$$

gdzie k jest pewną wartością krytyczną współczynnika filtracji kłębuszkowej. Wartość tę należy ustalić korzystając ze specjalistycznej wiedzy medycznej. Przyjmujemy $k = 60$, ponieważ

wyniki poniżej tej wartości świadczą o co najmniej umiarkowanej niewydolności nerek [7]. W tabeli 3.4 znajdują się częstości zmiennej Y.

MDRD24	Częstość
0	153
1	72

Tabela 3.4: Częstości Y.

Modyfikacji ulega również zmienna jakościowa *schemat terapii*. Przyjmuje ona trzy wartości $\{ca, cm, tc\}$. Do estymacji modelu program R traktuje ją automatycznie jako dwie zmienne binarne:

$$schemat\ terapii\ cm_i = \begin{cases} 1 & \text{jeżeli } schemat\ terapii_i = cm, \\ 0 & \text{w przeciwnym przypadku.} \end{cases}$$

$$schemat\ terapii\ tc_i = \begin{cases} 1 & \text{jeżeli } schemat\ terapii_i = tc, \\ 0 & \text{w przeciwnym przypadku.} \end{cases}$$

Wartość *ca* jest poziomem bazowym. Utworzenie zmiennych binarnych w miejsce zmiennej jakościowej[1] jest niezbędne do poprawnej konstrukcji modelu.

Konieczna jest także binaryzacja zmiennych *liczba leków na ciśnienie*, *liczba niezgodności AB*, *liczba niezgodności DR*. Pozostawiając je w obecnej formie, zakładałybyśmy, że np. logarytm szans pacjenta, przyjmującego 2 leki na nadciśnienie wynosi dwa razy tyle, co logarytm szans pacjenta przyjmującego 1 lek. Niekoniecznie musi to być prawdą. Nieprawidłowe byłoby także utworzenie zmiennej binarnej dla każdego poziomu (w celu uniknięcia dokładnej współliniowości pomijając poziom bazowy), ponieważ niektóre wartości występują zbyt rzadko w zbiorze danych. Na przykład, tylko sześciu pacjentów charakteryzuje liczba niezgodności AB równa 0 (tabela 3.3). Po połączeniu poziomów o liczbie obserwacji $\leqslant 15$, utworzyłyśmy następujące zmienne binarne:

$$liczba\ niezgodności\ AB2_i = \begin{cases} 1 & \text{jeżeli } liczba\ niezgodności\ AB_i = 2, \\ 0 & \text{w przeciwnym przypadku.} \end{cases}$$

$$liczba\ niezgodności\ AB3_i = \begin{cases} 1 & \text{jeżeli } liczba\ niezgodnosi\ AB_i \geqslant 3, \\ 0 & \text{w przeciwnym przypadku.} \end{cases}$$

$$liczba\ niezgodności\ DR_i = \begin{cases} 1 & \text{jeżeli } liczba\ niezgodności\ DR_i \geqslant 1, \\ 0 & \text{w przeciwnym przypadku.} \end{cases}$$

$$leki2_i = \begin{cases} 1 & \text{jeżeli } liczba\ leków\ na\ ciśnienie_i = 2, \\ 0 & \text{w przeciwnym przypadku.} \end{cases}$$

$$leki3_i = \begin{cases} 1 & \text{jeżeli } liczba\ leków\ na\ ciśnienie_i = 3, \\ 0 & \text{w przeciwnym przypadku.} \end{cases}$$

$$leki4_i = \begin{cases} 1 & \text{jeżeli } liczba\ leków\ na\ ciśnienie_i \geqslant 4, \\ 0 & \text{w przeciwnym przypadku.} \end{cases}$$

[1]Zaprezentowano najprostszy sposób przekodowania zmiennej jakościowej, stosowany automatycznie przez R. Istnieją altertywne podejścia, np. kontrasty w odchyleniach lub efekty progowe, narzucające inną interpretację współczynników [5].

Poziomami bazowymi są:

- *liczba niezgodności* $AB \leqslant 1$,
- *liczba niezgodności* $DR = 0$,
- *liczba leków na ciśnienie*$_i \leqslant 1$.

3.4. Wybór modelu

W tym rozdziale spróbujemy dobrać jak najlepszy model regresji logistycznej do naszych danych. Jakość modelu oceniamy na podstawie dwóch kryteriów: zgodności z założeniami z punktu (1.3.2) oraz łatwości w interpretacji medycznej, co umożliwia jego praktyczne zastosowanie. Na początek skoncentrujemy się na tym, które zmienne objaśniające uwzględnić i czy nie należy zmienić ich formy funkcyjnej. Opracowano różne metody wyboru zmiennych, przedstawimy krótko te z nich, które będziemy stosować. Pierwszą metodą jest dobór zmiennych objaśniających w oparciu o wiedzę medyczną oraz intuicję. Wadą ograniczenia się do takiego postępowania jest ryzyko niepoprawnego odwzorowania zależności oraz zbytniego rozbudowania modelu. Konsekwencjami użycia tej metody może być również obciążenie oszacowania współczynników lub błędów standardowych. Kolejną metodą jest testowanie istotności statystycznej, przeprowadzając testy oparte między innymi na statystyce Walda z lub statystyce dewiancji D. Odrzucamy wówczas te zmienne objaśniające, dla których p-wartość testu jest większa od ustalonego poziomu istotności. Ten poziom jest wyższy niż stosowane zazwyczaj w testowaniu hipotez 5%. Hosmer i Lemeshow w [3] doradzają wybór $\alpha = 25\%$. Te dwie metody uzupełnimy analizą wykresów wartości zmiennych objaśniających od logarytmu szans modelu. Weźmiemy też pod uwagę model wytypowany przez R metodą krokową (funkcja `step`).

W pierwszym modelu uwzględniłyśmy wszystkie zmienne. Z medycznego punktu widzenia, każda ze zmiennych objaśniających może mieć wpływ na powodzenie przeszczepu. Wyniki regresji logistycznej są następujące:

```
glm(formula = MDRD24 ~ wiek.biorcy + wiek.dawcy + czas.zimnego.niedokrwienia +
    AB + DR + schemat.terapi + czy.cukrzyca + leki, family = binomial())

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9235  -0.7568  -0.4757   0.7980   2.3293

Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                1.8138047  1.0128581   1.791  0.07333 .
wiek.biorcy                0.0001513  0.0166921   0.009  0.99277
wiek.dawcy                -0.0610159  0.0148371  -4.112 3.92e-05 ***
czas.zimnego.niedokrwienia 0.0411187  0.0214472   1.917  0.05521 .
AB2                       -0.3903571  0.4697757  -0.831  0.40601
AB3                       -0.5988876  0.4854233  -1.234  0.21730
DR1                        0.5410881  0.3937269   1.374  0.16936
schemat.terapicm          -0.3140939  0.4273650  -0.735  0.46237
schemat.terapitc          -0.7864773  0.4193628  -1.875  0.06074 .
czy.cukrzyca               0.2024203  0.4193569   0.483  0.62931
```

```
leki2                       -0.5913231  0.3992290  -1.481  0.13856
leki3                       -1.4170921  0.4629947  -3.061  0.00221 **
leki4                       -2.1039987  0.6946424  -3.029  0.00245 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 282.09  on 224  degrees of freedom
Residual deviance: 222.38  on 212  degrees of freedom
AIC: 248.38

Number of Fisher Scoring iterations: 5
```

Test Walda odrzucił hipotezę zerową o nieistotności jedynie dla zmiennych *wiek dawcy*, *leki2* i *leki3* (na poziomie ufności 95%). Te zmienne uzględnimy w naszym końcowym modelu. Zanim postanowimy o tym, które z pozostałych zmiennych usunąć, zrobimy oddzielne regresje logistyczne dla każdego czynnika objaśniającego i porównamy wartości oszacowań współczynników i p-value z dwóch typów modeli. Wyniki przedstawiamy w tabeli 3.5. Ponadto zrobimy wykresy wartości każdej zmiennej od logarytmu szans modelu uwzględniającego wszystkie zmienne, a dla zmiennych dyskretnych także od logarytmu szans modelu zawierającego jedynie tą zmienną i stałą (rysunki 3.1-3.6). Analizując wartości oszacowań współczynników i p-value oraz wykresy, podejmiemy decyzje o usunięciu niektórych zmiennych bądź ich modyfikacji.

P-value zmiennej *wiek biorcy* wynosi 0.38 dla regresji tylko tej zmiennej i prawie 1 w modelu ze wszystkimi zmiennymi objaśniającymi. Możliwym wyjaśnieniem tej różnicy jest fakt, że wiek biorcy jest skorelowany z występowaniem chorób, np. nadciśnienia tętniczego. Stąd w modelu uwzględniającym zmienne wskazujące na występowanie tej choroby p-value jest znacznie wyższe, a współczynnik jest przeciwnego znaku niż w modelu zawierającym tylko stałą i wiek biorcy. W obu przypadkach p-value jest większe od poziomu istotności testu Walda (5%) i progu wskazanego przez Hosmera i Lemeshowa, a współczynnik jest bliski 0. Na nieistotność tej zmiennej wskazuje także wykres 3.1 – punkty są chaotycznie rozmieszczone. Nie uwzględnimy więc tej zmiennej w ostatecznym modelu.

Zmienna *wiek dawcy* w teście Walda odrzuca hipotezę zerową o nieistotności na ustalonym poziomie istotności 5% w obu przypadkach – p-wartości nawet mniejsze aż od 0.001. Wykres 3.1 pokazuje wyraźną zależność logarytmu szans od wieku dawcy. Niewątpliwie jest to bardzo wpływowa zmienna, która pojawi się w modelu końcowym.

Dla zmiennej *czas zimnego niedokrwienia* p-wartość testu Walda w modelu z tylko tą zmienną wynosi 0.47, nie odrzuca więc hipotezy zerowej o nieistotności. Jednak dla modelu ze wszystkimi zmiennymi objaśniającymi p-value maleje do 0.055. Formalnie, test na poziomie istotności 5% wciąż nie odrzuca hipotezy zerowej, lecz p-value jest na tyle bliska granicznej wartości, że podejrzewamy, że po uwzględnieniu innych czynników czas zimnego niedokrwienia może mieć istotny wpływ. Z wykresu 3.1 widać, że zależność nie jest tak silna jak w przypadku zmiennej *wiek dawcy*. Punkty rozmieszczone są dosyć chaotycznie, a krzywa lowess (skrót od ang. *locally weighted scatterplot smoothing*) ma nieliniowy kształt. Mimo to, z uwagi na niskie p-value w rozbudowanym modelu (< 0.25), zachowamy tą zmienną.

Według testu Walda, nie ma podstaw do odrzucenia hipotezy o nieistotności zmiennych dotyczących liczby niezgodności AB. W każdym przypadku p-value jest wyższa od 5% poziomu istotności. Jednak dla poziomu odpowiadającego trzem lub więcej niezgodnościom AB

Model ze wszystkimi zmiennymi			Model z jedną zmienną objaśniającą		
	współczynniki	p-wartość		współczynniki	p-wartość
stała	1.8138047	0.07333 .	stała	-0.28023	0.613
wiek.biorcy	0.0001513	0.99277	wiek.biorcy	-0.01134	0.380
			stała	1.95395	0.000221 ***
wiek.dawcy	-0.0610159	3.92e-05 ***	wiek.dawcy	-0.06562	2.11e-07 ***
			stała	-1.06611	0.0192 *
czas.zimnego. niedokrwienia	0.0411187	0.05521 .	czas.zimnego. niedokrwienia	0.01342	0.4671
			stała	-0.4700	0.153
AB2	-0.3903571	0.40601	AB2	-0.1780	0.649
AB3	-0.5988876	0.21730	AB3	-0.5534	0.176
			stała	-1.1632	8.42e-05 ***
DR1	0.5410881	0.16936	DR1	0.5522	0.103
			stała	-0.2992	0.13701
schemat. terapi.cm	-0.3140939	0.46237	schemat. terapi.cm	-0.5762	0.11673
schemat. terapi.tc	-0.7864773	0.06074 .	schemat. terapi.tc	-1.1392	0.00151 **
			stała	-0.7282	7.63e-06 ***
czy.cukrzyca	0.2024203	0.62931	czy.cukrzyca	-0.1101	0.747
			stała	0.09844	0.701011
leki2	-0.5913231	0.13856	leki2	-0.73216	0.038063 *
leki3	-1.4170921	0.00221 **	leki3	-1.55069	0.000237 ***
leki4	-2.1039987	0.00245 **	leki4	-2.33203	0.000405 ***

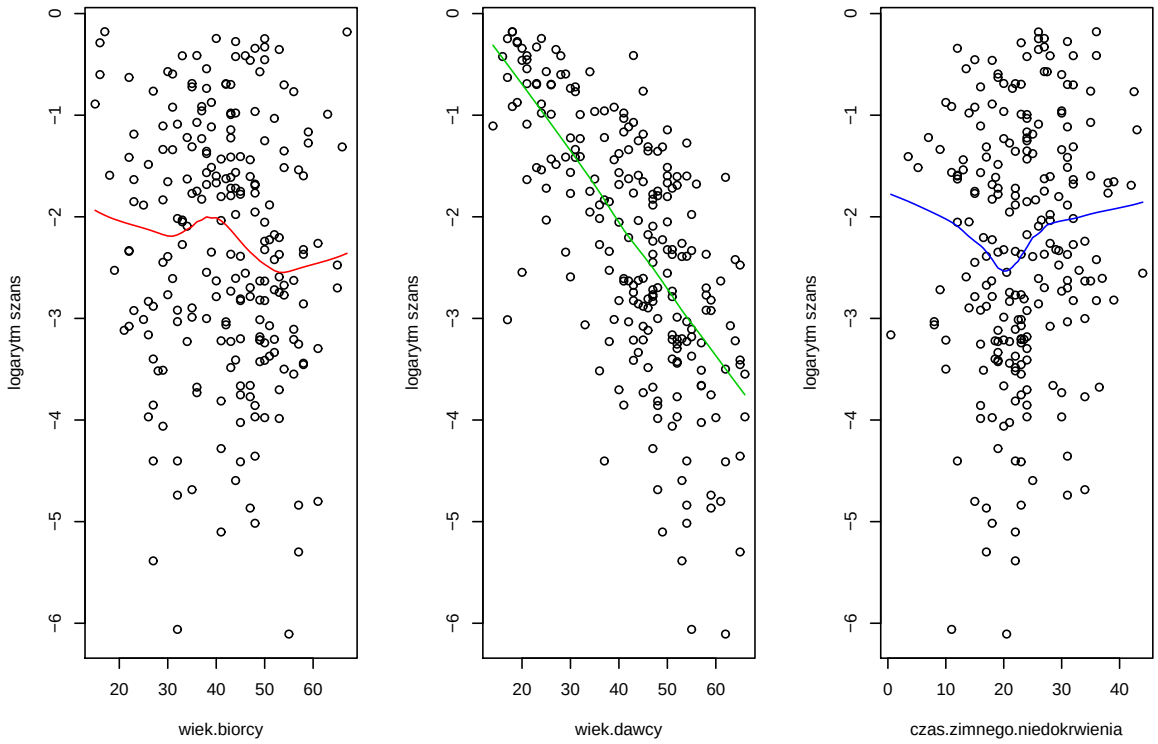
Tabela 3.5: Współczynniki i p-wartości modeli regresji logistycznej dla wszystkich zmiennych i dla każdej zmiennej osobno.

w obu modelach p-value jest mniejsze od progu 0.25, dlatego rozważamy pozostawienie tych zmiennych. Z wykresu 3.2 odczytujemy, że poziomy dla jednej i mniej niezgodności AB oraz dwóch niezgodności mają zbliżony wpływ na logarytm szans. Spróbowałyśmy połączyć te poziomy i wykonać analogiczne regresje. P-value w modelu ze wszystkimi zmiennymi wzrosło do 0.36605, co przesądziło o nieobecności tej zmiennej w ostatecznym modelu.

Inną decyzję podjęłyśmy w przypadku zmiennej *liczba niezgodności DR* (wykres 3.3). P-wartości, choć przekraczają 5% (wynoszą 0,17 w modelu rozbudowanym, 0,103 w modelu tylko ze zmienną DR i stałą), nie są mniejsze od 0.25. Uwzględnimy ją, gdyż pominięcie zmiennej, która wpływa na powodzenie przeszczepu, ma poważniejsze skutki, niż uwzględnienie zmiennej nieistotnej.

W przypadku zmiennej *czy cukrzyca* p-value znacznie przekraczają poziomy istotności 5% i 25%. Również na wykresach (rysunek 3.5) widać, że poziomy tej zmiennej mają podobny wpływ na powodzenie przeszczepu. Nie uwzględnimy jej w końcowym modelu.

Końcowy model zawierał będzie także zmienne dotyczące schematu terapii i liczby leków na nadciśnienie tętnicze. Test Walda wskazuje na odrzucenie hipotezy zerowej o nieistotności schematu *tc* w modelu tylko ze schematem terapii (p-value= $0,0015 << 0.05$), natomiast w modelu rozbudowanym p-value= $0,06$ tylko nieznacznie przekracza poziom istotności 5%

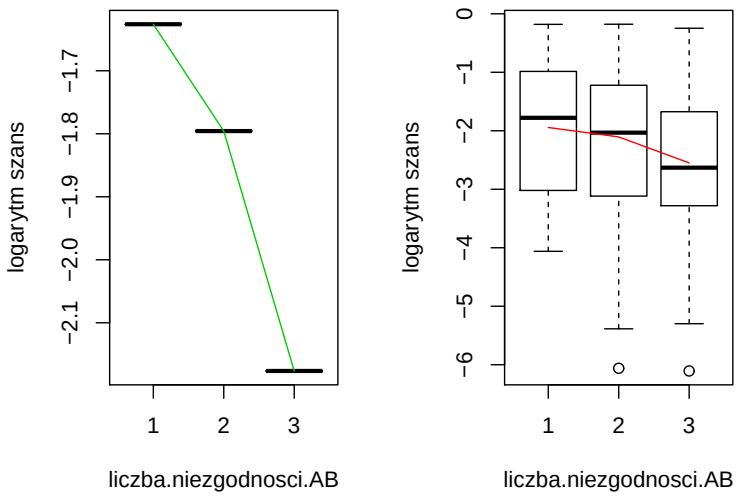


Rysunek 3.1: Wykres zależności zmiennych: *wiek biorcy*, *wiek dawcy* i *czas zimnego niedokrwienia* dla modelu uwzględniającego wszystkie zmienne. Kolorem oznaczono krzywe lokalnie ważonej regresji.

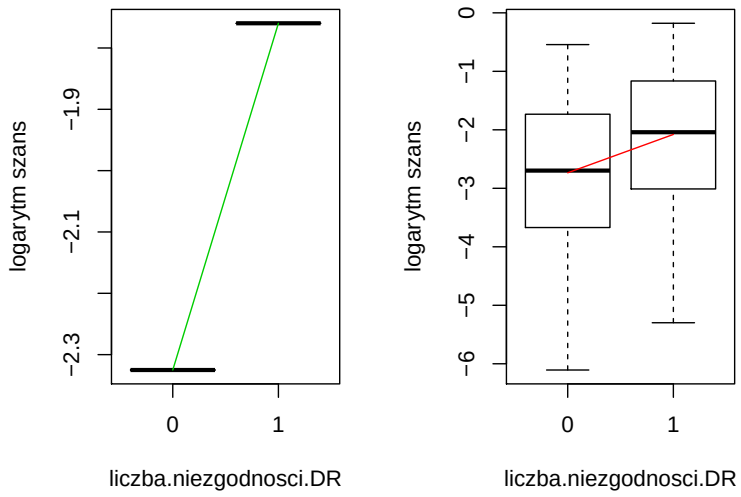
i jest mniejszy od 25%. Test Walda nie rozstrzyga jednak kwestii istotności poziomu *cm*. Pomysł włączenia tego poziomu do poziomu bazowego (*ca*) zarzucamy po analizie wykresu 3.4. Widoczne jest, że poszczególne poziomy mają różny wpływ na powodzenie przeszczepu. Z kolei liczba leków na nadciśnienie jest zmienną bardzo istotną. Wskazują na to zarówno p-wartości w modelu ze wszystkimi zmiennymi, jak i w modelu tylko z jedną zmienną oraz wykres 3.6. W przypadku poziomów charakteryzujących pacjentów przyjmujących 3 leki (*leki*3) oraz 4 i więcej (*leki*4) są one dużo mniejsze od ustalonego poziomu istotności równego 5%. Choć p-wartość dla poziomu *leki*1 przekracza 5% w modelu rozbudowanym, wynosząc 0.13856, jest mniejsza od progu 0.25. Nie rozważamy zatem modyfikacji poziomów.

Podsumowując, w naszym końcowym modelu postanowiłyśmy uwzględnić następujące zmienne objaśniające: *wiek dawcy*, *czas zimnego niedokrwienia*, *liczba niezgodności DR*, *schemat terapii*, *liczba leków na ciśnienie*. Wydruk z programu R z wynikami regresji dla naszego modelu przedstawiamy poniżej:

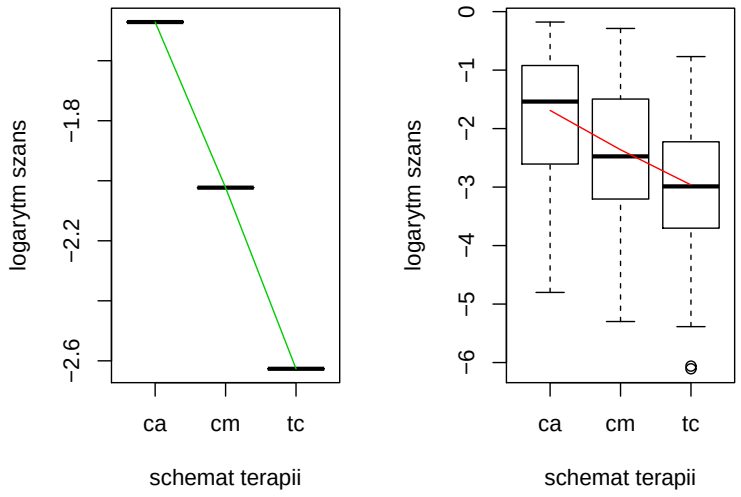
```
glm(formula = MDRD24 ~ wiek.dawcy + czas.zimnego.niedokrwienia +
    DR + schemat.terapi + leki, family = binomial())

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9221  -0.7741  -0.4814   0.7855   2.3726
```



Rysunek 3.2: Wykresy pudełkowe dla logarytmu szans od poszczególnych poziomów zmiennej *liczba niezgodności AB* dla modelu z tylko tą zmienną (z lewej) oraz modelu uwzględniającego wszystkie zmienne (z prawej). W modelu uwzględniającym wszystkie zmienne, logarytm szans przyjmuje różne wartości dla poszczególnych obserwacji, ponieważ zależy od wartości wszystkich zmiennych objaśniających. Na wykresie pudełkowym, dolna podstawa prostokąta wyznaczona jest przez pierwszy kwartyl, górna natomiast przez trzeci kwartyl. Wysokość prostokąta odpowiada zatem rozstępowi ćwiartkowemu. Pozioma linia wewnątrz prostokąta oznacza wartość mediany. Odcinki łączą prostokąt z najmniejszą i największą wartością. Ponadto, na powyższych wykresach połączono odcinkami średnie wartości logarytmu szans dla poszczególnych poziomów zmiennej *liczba niezgodności AB*.



Rysunek 3.3: Wykresy pudełkowe dla logarytmu szans od poszczególnych poziomów zmiennej *liczba niezgodności DR* dla modelu z tylko tą zmienną (z lewej) oraz modelu uwzględniającego wszystkie zmienne (z prawej). Średnie wartości logarytmu szans dla poszczególnych poziomów połączono odcinkami.



Rysunek 3.4: Wykresy pudełkowe dla logarytmu szans od poszczególnych poziomów zmiennej *schemat terapii* dla modelu z tylko tą zmienną (z lewej) oraz modelu uwzględniającego wszystkie zmienne (z prawej). Średnie wartości logarytmu szans dla poszczególnych poziomów połączono odcinkami.

```
Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)                 1.54596    0.79680   1.940  0.05235 .
wiek.dawcy                 -0.06037    0.01403  -4.305 1.67e-05 ***
czas.zimnego.niedokrwienia  0.03696    0.02084   1.774  0.07610 .
DR1                         0.52038    0.38865   1.339  0.18060
schemat.terapicm           -0.31341    0.41870  -0.749  0.45414
schemat.terapitc           -0.75645    0.40779  -1.855  0.06360 .
leki2                      -0.59828    0.39248  -1.524  0.12741
leki3                      -1.39755    0.45992  -3.039  0.00238 **
leki4                      -2.16208    0.68989  -3.134  0.00172 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 282.09  on 224  degrees of freedom
Residual deviance: 224.23  on 216  degrees of freedom
AIC: 242.23
Number of Fisher Scoring iterations: 5
```

Na podstawie testu Walda na poziomie istotności 5% możemy stwierdzić, że zmienne *wiek dawcy*, *leki*3 i *leki*4 mają istotny wpływ na powodzenie przeszczepu nerki. Poniżej zamieszczamy krańce przedziałów ufności dla oszacowań współczynników:

```
library(MASS)
confint(glm.koncowy)
                                  2.5 %      97.5 %
(Intercept)                 0.0001090256  3.14055053
wiek.dawcy                 -0.0889266886 -0.03370806
czas.zimnego.niedokrwienia -0.0035671953  0.07851351
DR1                        -0.2277311660  1.30398141
schemat.terapicm           -1.1487571272  0.50158832
schemat.terapitc           -1.5757628283  0.03164910
leki2                      -1.3761822006  0.16811632
leki3                      -2.3315843894 -0.51798979
leki4                      -3.7109951364 -0.92789707
```

Na koniec przyjrzyjmy się jeszcze modelowi, który proponuje pakiet R. Po użyciu funkcji `step(glm)`otrzymujemy następujący wynik:

```
Step:  AIC=241.41

MDRD24 ~ wiek.dawcy + czas.zimnego.niedokrwienia + leki

                            Df Deviance    AIC
<none>                           229.41 241.41
- czas.zimnego.niedokrwienia  1   231.61 241.61
- leki                        3   249.38 255.38
- wiek.dawcy                  1   256.47 266.47
```

Model ten różni się od naszego końcowego modelu. Brakuje w nim dwóch zmiennych: *liczba niezgodności DR* oraz *schemat terapii*. Różnice te są wynikiem brania innych kryteriów

przy doborze modelu. Funkcja `step` buduje model na podstawie kryterium AIC, odrzucając zmienne posiadające najniższe wartości tego współczynnika. Do budowy naszego modelu posłużył nam natomiast test Walda, wykresy diagnostyczne oraz intuicja oparta na wiedzy medycznej.

3.5. Diagnostyka współliniowości

Sprawdzimy, czy w naszym modelu nie występuje problem współliniowości. Zmienne objaśniające są współliniowe, gdy są mocno skorelowane ze sobą – kolumny macierzy X są wówczas bliskie liniowej zależności. Wtedy model regresji może mieć zawyżone oszacowania współczynników i duże wartości błędów standardowych. Efekt ten wyrażany jest poprzez współczynnik VIF_i (ang. *variance inflation factor*), który pokazuje, o ile wariancje współczynników są zawyżone z powodu zależności liniowych w badanym modelu regresji. Obliczamy go ze wzoru: $VIF_i = \frac{1}{1-R_i^2}$, gdzie R_i^2 jest współczynnikiem wielokrotnej determinacji dla i-tej zmiennej w modelu regresji liniowej. W modelu tym zmienną objaśnianą staje się zmienna x_i, natomiast zmienne x_j, gdzie $j \neq i$ pozostają zmiennymi objaśniającymi. Przyjmuje się, że wartość $VIF_i > 10$ wskazuje na obecność współliniowości w modelu [5]. Dla naszego modelu wartośći współczynników VIF_i są następujące:

```
library(Design)
vif(glm_nasz)
                wiek.dawcy czas.zimnego.niedokrwienia
                  1.092589                   1.074570
                       DR1           schemat.terapicm
                  1.046014                   1.139231
          schemat.terapitc                      leki2
                  1.193395                   1.347153
                     leki3                      leki4
                  1.311647                   1.134510
```

Otrzymujemy wynik wskazujący na brak współliniowości w modelu, gdyż dla każdej ze zmiennych współczynnik VIF_i tylko nieznacznie przekracza 1. Oznacza to, że przedziały ufności dla poszczególnych zmiennych objaśniających nie są poszerzone na skutek korelacji tych zmiennych z pozostałymi. Ogólnie, pierwiastek ze współczynnika VIF mówi o tym, ile razy zwiększony jest błąd standardowy oszacowania w porównaniu do modelu, w którym nie występowałaby korelacja tej zmiennej z pozostałymi zmiennymi objaśniającymi.

3.6. Interpretacja modelu

Wykonamy analizę współczynników końcowego modelu, przedstawiając je w terminach szans. Przypomnijmy, że szansa jest to iloraz prawdopodobieństwa sukcesu do prawdopodobieństwa porażki. W regresji logistycznej szansa zmiennej objaśniającej to $\exp(\beta)$, gdzie β jest współczynnikiem stojącym przy tej zmiennej.

```
exp(coef(glm_koncowy))
          (Intercept)             wiek.dawcy     czas.zimnego.niedokrwienia
            4.6924673              0.9414140              1.0376503

        schemat.terapicm       schemat.terapitc                   DR1
```

```
0.7309514          0.4693316          1.6826625

   leki2              leki3              leki4
0.5497541          0.2472027          0.1150851
```

Szansa dla zmiennej *wiek dawcy* wynosi 0.94, więc czynnik opisany tą zmienną działa ograniczająco na dobre działanie nerki po roku od przeszczepu. Wraz ze wzrostem o rok wieku dawcy szanse na przyjęcie się nerki maleją o ok. 6% (100%-94%).

Zaskakującym wynikiem okazała się szansa zmiennej opisującej *czas zimnego niedokrwienia*. Wynika bowiem z niej, że wraz ze wzrostem o jedną jednostkę czasu zimnego niedokrwienia nerki szansa na jej prawidłowe funkcjonowanie rośnie (o 4%). Podobnie nieoczekiwany wynik uzyskałyśmy dla liczby niezgodności DR – osoby posiadające co najmniej jedną niezgodność DR mają o 68% większą szansę na przyjęcie się nerki niż pacjenci bez niezgodnośći. Są to wnioski nieintuicyjne, przeczy im także medycyna. Mogą być wynikiem źle dopasowanego modelu.

Przeanalizujemy szanse zajścia zdarzenia dla pacjentów przyjmujących trzy rodzaje terapii. Dla pacjenta leczonego według terapii cm szansa przyjęcia się nerki maleje o ok. 27% w stosunku do pacjenta leczonego terapią ca. Natomiast szansa pacjenta, który otrzymuje terapię tc maleje o ok. 53%.

Szanse zmiennych $leki_i$ przedstawimy za pomocą ilorazów szans obliczonych zgodnie ze wzorem (1.11). Prezentujemy je w tabeli 3.5.

$\frac{o(kolumna)}{o(wiersz)}$	leki2	leki3	leki4
leki2	1	0.7389305	0.647479
leki3	1.353307	1	0.876238
leki4	1.544452	1.141243	1

Tabela 3.6: Ilorazy szans dla poszczególnych grup pacjentów biorących różne dawki leków.

Dokładną interpretację przedstawimy na przykładzie ilorazu szans dla pacjentów biorących co najwyżej dwa leki w stosunku do osób zażywających trzy leki. Iloraz szans dla tych grup wyniósł w przybliżeniu 1.35, zatem grupa pierwsza ma 1.35 razy większą szansę na udany przeszczep niż druga. Ponadto możemy powiedzieć, że przyjmowanie mniejszej ilości leków powoduje wzrost szansy na przyjęcie nerki o ok 35%. Z obliczeń zaprezentowanych w tabeli widzimy, że wraz z każdym dodatkowo przyjmowanym lekiem na nadciśnienie szansa na prawidłowe funkcjonowanie nerki po 24 miesiącach maleje.

3.7. Przykład zastosowania modelu do prognozy

Model logistyczny znajduje również swoje zastosowanie do prognozowania szansy przyjęcia się nerki dla pacjentów spoza wykorzystanego zbioru danych. Może być to istotne jako informacja dla pacjenta lub wsparcie w procesie wyboru osoby, która otrzyma nerkę od zmarłego dawcy. Przedstawimy prognozy szans dla pacjentów, których dane zamieszczamy w tabeli 3.7. Za pomocą funkcji `predict` otrzymałyśmy wartości regresji dla poszczególnych pacjentów. Następnie wyliczyłyśmy szanse każdego z nich na powodzenie przeszczepu.

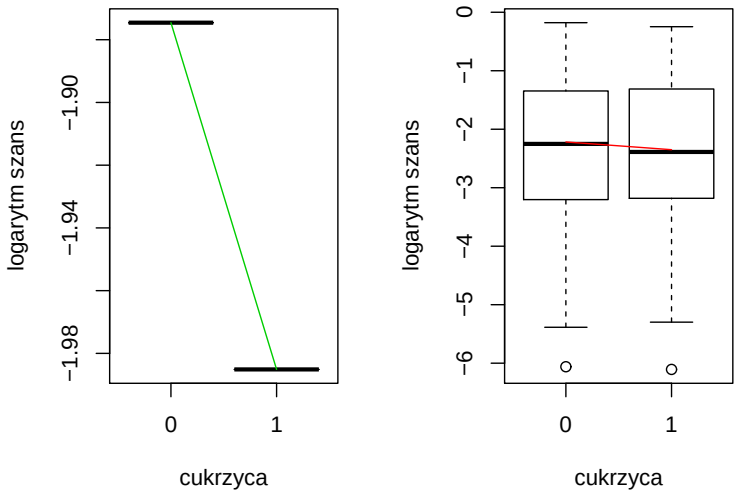
```
data.frame(wiek.dawcy=c(40,67,15), czas.zimnego.niedokrwienia=c(10, 8,5),
           DR=c(1,1,0), leki=c(3,3,1), schemat.terapi=c('tc','tc','cm'))
```

pacjent	wiek dawcy	czas zimnego niedokrwienia	liczba niezgodności DR	liczba leków na ciśnienie	schemat terapii
1	40	10	1	3	tc
2	67	8	1	3	tc
3	15	5	0	1	cm

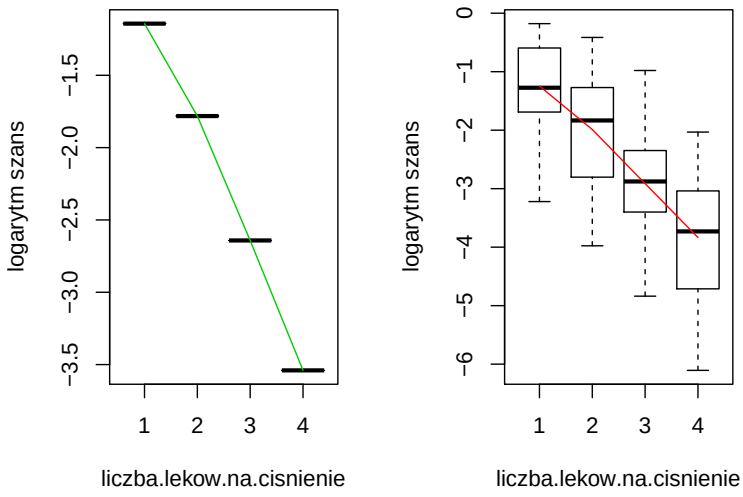
Tabela 3.7: Przykładowe dane pacjentów.

```
K2$DR=as.factor(K2$DR)
K2$leki=as.factor(K2$leki)

predict(glm.koncowy, K2)
        1         2         3
-2.133241 -3.847906 0.5536762

exp(predict(glm.koncowy, K2)
         1          2          3
0.11845276 0.02132434 1.73963650
```

Pierwszy pacjent ma szansę powodzenia przeszczepu ok. 12:100. Oznacza to, że na 112 pacjentów charakteryzujących się takimi wartościami zmiennych jak pacjent pierwszy, przewidujemy, że tylko u dwunastu zaobserwujemy sukces. Dla drugiego pacjenta szanse wynoszą tylko 2:100, natomiast dla trzeciego aż 170:100. Korzystając ze wzoru (1.8), wyrazimy wartości szans pacjentów w terminach prawdopodobieństwa. Prognozowane prawdopodobieństwo przyjęcia nerki pierwszego pacjenta wynosi 0.106, drugiego – 0.021, a trzeciego – 0.635. Możemy zauważyć, że chociaż dane pacjentów oznaczonych numerami 1 i 2 różnią się tylko wiekiem dawcy oraz czasem zimnego niedokrwienia, ich przewidywane prawdopodobieństwa sukcesu różnią już się znacząco. Na podstawie naszego modelu prognozujemy, że największe szanse na prawidłowe funkcjonowanie nerki po przeszczepie ma pacjent numer 3.



Rysunek 3.5: Wykresy pudełkowe dla logarytmu szans od poszczególnych poziomów zmiennej *czy cukrzyca* dla modelu z tylko tą zmienną (z lewej) oraz modelu uwzględniającego wszystkie zmienne (z prawej). Średnie wartości logarytmu szans dla poszczególnych poziomów połączono odcinkami.



Rysunek 3.6: Wykresy pudełkowe dla logarytmu szans od poszczególnych poziomów zmiennej *liczba leków na ciśnienie* dla modelu z tylko tą zmienną (z lewej) oraz modelu uwzględniającego wszystkie zmienne. Średnie wartości logarytmu szans dla poszczególnych poziomów połączono odcinkami.

Podsumowanie

Praca zawiera wprowadzenie do modelu regresji logistycznej. Modele regresji służą przede wszystkim do opisu zależności pomiędzy zmiennymi. Są powszechnie stosowane w badaniach empirycznych z różnych dziedzin. W naszej pracy zastosowałyśmy regresję do danych medycznych.

W rozdziale teoretycznym wprowadziłyśmy podstawowe pojęcia dotyczące uogólnionych modeli liniowych, lecz najwięcej uwagi poświęciłyśmy szczególnemu przypadkowi takich modeli – regresji logistycznej. Charakterystyczna dla regresji logistycznej jest binarna zmienna objaśniana, zwykle wskazująca na wystąpienie lub brak wystąpienia zdarzenia, które chcemy prognozować. Regresja logistyczna umożliwia modelowanie prawdopodobieństwa tego zdarzenia. Współczynniki modelu regresji logistycznej można interpretować jako szanse.

W przypadku danych, które analizowałyśmy, za zmienną objaśnianą wybrałyśmy współczynnik filtracji kłębuszkowej badany po 24 miesiącach od przeszczepu nerki. Zmienne objaśniające, które miałyśmy do dyspozycji to wiek dawcy, wiek biorcy, czas zimnego niedokrwienia, schemat terapii, liczba niezgodności AB, liczba niezgodności DR, występowanie cukrzycy, liczba leków na ciśnienie.

Konstrukcji modelu regresji logistycznej można dokonać na podstawie różnych kryteriów. W pracy skupiłyśmy się na testowaniu modelu na podstawie statystyki Walda. Postanowiłyśmy również oceniać p-value konkretnych zmiennych oraz analizować wykresy diagnostyczne. Upewniłyśmy się, że w preferowanym modelu nie zachodzi problem współliniowości, a następnie zinterpretowałyśmy oszacowania współczynników w terminach szans. Na końcu podałyśmy przykład zastosowania modelu do prognozowania szansy powodzenia przeszczepu nerki.

Zbudowany przez nas model nieco różni się od modelu wyestymowanego przez dostępną w pakiecie R funkcję `step`. Różnice te zapewne wiążą się z odmienną od naszej metodą budowy modelu regresji stosowanej przez tę funkcję.

Dodatek A

Kody pakietu R użyte w pracy

Zamieszczone kody pakietu R są to funkcje, których użyłyśmy w celu doboru odpowiedniego modelu. Najpierw zamieszczamy modyfikację zmiennych, dokładniej opisaną w rozdziale 3.3:

```
leki=factor(factor(liczba.lekow.na.cisnienie, labels=c("1", "1", "2", "3",
"4", "4", "4")))
AB=factor(factor(liczba.niezgodnosci.AB, labels=c("1","1","2","3","3")))
DR=factor(factor(liczba.niezgodnosci.DR, labels=c("0","1","1")))
```

Model ze wszystkimi zmiennymi wywołujemy następującą komendą:

```
glm <- glm(formula = MDRD24 ~ wiek.biorcy + wiek.dawcy + czas.zimnego.
niedokrwienia + DR + AB + schemat.terapi + leki, family = binomial())
summary(glm)
```

Wykresy dla poszczególnych zmiennych posłużyły nam w ocenie istotności danej zmiennej w modelu. Dla poszczególnych zmiennych wykonałyśmy po dwa wykresy. Przedstawiają one zależności zmiennej od logarytmu szans modelu tylko z tą zmienną oraz ze wszystkimi zmiennymi objaśniającymi. Wykresy uzyskujemy funkcją `plot`, a krzywą lokalnie ważonej regresji - `lines(lowess)`.

```
plot(wiek.biorcy, log((glm$fitted.value)/(1-log(glm$fitted.value))), xlab=
"wiek.biorcy", ylab="logarytm szans")
lines(lowess(wiek.biorcy, log((glm$fitted.value)/(1-log(glm$fitted.value))),
col = 2)

plot(wiek.dawcy, log(glm$fitted.value)- log(1-glm$fitted.value), xlab=
"wiek.dawcy", ylab="logarytm szans")
lines(lowess(wiek.dawcy, log((glm$fitted.value)/(1-log(glm$fitted.value)))),
col = 2)

plot(czas.zimnego.niedokrwienia, log(glm$fitted.value)-
log(1-glm$fitted.values), xlab="czas.zimnego.niedokrwienia", ylab= "logarytm
szans")
lines(lowess(czas.zimnego.niedokrwienia, log(glm$fitted.value)-
log(1-glm$fitted.value)), col = 4)
```

```
glm_AB <- glm( formula = MDRD24 ~ AB, family = binomial() )
plot(factor(AB), log((glm_AB$fitted.value)/(1-log(glm_AB$fitted.value))),
xlab="liczba.niezgodności.AB", ylab="logarytm szans")
lines(lowess(AB, log((glm_AB$fitted.value)/(1-log(glm_AB$fitted.value)))),
col = 3)
plot(factor(AB), log((glm$fitted.value)/(1-log(glm$fitted.value))), xlab=
"liczba.niezgodności.AB", ylab="logarytm szans")
lines(lowess(AB, log((glm$fitted.value)/(1-log(glm$fitted.value)))),
col = 2)

glm_DR <- glm( formula = MDRD24 ~ DR, family = binomial() )
plot(factor(DR), log((glm_DR$fitted.value)/(1-log(glm_DR$fitted.value))),
xlab="liczba.niezgodności.DR", ylab="logarytm.szans")
lines(lowess(DR, log((glm_DR$fitted.value)/(1-log(glm_DR$fitted.value)))),
col = 3)
plot(factor(DR), log((glm$fitted.value)/(1-log(glm$fitted.value))), xlab=
"liczba.niezgodności.DR", ylab="logarytm szans")
lines(lowess(DR, log((glm$fitted.value)/(1-log(glm$fitted.value)))),col = 2)

glm_schemat <- glm(formula = MDRD24 ~ schemat.terapi, family = binomial())
plot(schemat.terapi, log(glm_schemat$fitted.value)-
log(1-glm_schemat$fitted.values), xlab="schemat.terapi", ylab="logarytm
szans")
lines(lowess(schemat.terapi, log(glm_schemat$fitted.value)-
log(1-glm_schemat$fitted.value)), col = 3)
plot(schemat.terapi, log(glm$fitted.value)- log(1-glm$fitted.values),
xlab="schemat.terapi", ylab="logarytm szans")
lines(lowess(schemat.terapi, log(glm$fitted.value)-
log(1-glm$fitted.value)), col = 2)

glm_cukrzyca <- glm( formula = MDRD24 ~ factor(czy.cukrzyca), family=
binomial())
plot(factor(czy.cukrzyca),log((glm_cukrzyca$fitted.value)/(1-log(glm_cukrzyca
$fitted.value))), xlab= czy.cukrzyca, ylab= logarytm szans)
lines(lowess(factor(czy.cukrzyca),log((glm$fitted.value)/(1-log(glm$fitted.
value)))), col = 3)
plot(factor(czy.cukrzyca), log((glm$fitted.value)/(1-log(glm$fitted.
value))), xlab= czy.cukrzyca, ylab= logarytm szans)
lines(lowess(factor(czy.cukrzyca),log((glm$fitted.value)/(1-log(glm$fitted.
value)))), col = 2)

glm_leki <- glm( formula = MDRD24 ~ leki, family = binomial())
plot(leki, log((glm_leki$fitted.value)/(1-log(glm_leki$fitted.value))), xlab=
"liczba.lekow.na.cisnienie, ylab="logarytm szans")
lines(lowess(leki, log((glm_leki$fitted.value)/(1-log(glm_leki$fitted.
value)))), col = 3)
plot(leki, log((glm$fitted.value)/(1-log(glm$fitted.value))), xlab="liczba.
lekow.na.cisnienie", ylab="logarytm szans")
lines(lowess(leki, log((glm$fitted.value)/(1-log(glm$fitted.value)))),
```

```
col = 2)
```

Nasz końcowy model uzyskałyśmy komendą:

```
glm_koncowy <- glm(formula = MDRD24 ~ wiek.dawcy + czas.zimnego.niedokrwienia +
               DR + schemat.terapi + leki, family = binomial())

summary(glm_koncowy)
```

Natomiast przedziały ufności dla końcowego modelu uzyskałyśmy następująco:

```
library(MASS)
confint(glm.koncowy)
```

Funkcją `step` uzyskałyśmy model zbudowany na podstawie kryterium Akaike:

```
step(model_koncowy, direction="backward")
```

Problem współliniowości zbadałyśmy funkcją `vif`:

```
library(Design)
vif(glm_koncowy)
```

Szanse dla poszczególnych zmiennych uzyskałyśmy:

```
exp(coef(glm_koncowy))
```

Utworzyłyśmy wektory z przykładowymi danymi trzech pacjentów w celu przewidzenia szans powodzenia przeszczepu nerki na podstawie naszego końcowego modelu. Funkcja `predict` wypisuje wartości regresji dla poszczególnych pacjentów. Obliczyłyśmy dla nich również szanse powodzenia przeszczepu.

```
data.frame(wiek.dawcy=c(40,67,15), czas.zimnego.niedokrwienia=c(10, 8,5),
            DR=c(1,1,0), leki=c(3,3,1), schemat.terapi=c('tc','tc','cm'))

K2$DR=as.factor(K2$DR)
K2$leki=as.factor(K2$leki)

predict(glm.koncowy, K2)

exp(predict(glm.koncowy, K2)
```

Bibliografia

[1] Przemysław Biecek, *Przewodnik po pakiecie R*, 2008

[2] Julian J.Faraway, *Extending the Linear Model with R. Generalized Linear, Mixed Effects and Nonparametric Regression Models*, 2006

[3] David W. Hosmer, Stanley Lemeshow, *Applied Logistic Regression*, 2000

[4] Scott Menard, *Applied Logistic Regression Analysis*, 2001

[5] Jerzy Mycielski, *Ekonometria*, 2010

[6] Daryl Pregibon, *Logistic Regression Diagnostics*, The Annals of Statistics, 1981

[7] Pod redakcją prof. dr. hab. Andrzeja Szczeklika, *Choroby wewnętrzne*, 2006, p. 1263.

[8] *Dokumentacja R*, `http://finzi.psych.upenn.edu/R/library/stats/html/`, dostęp dnia 11.04.2011 r.

[9] *SAS/INSIGHT® 9.1 User's Guide, Volumes 1 and 2*, SAS Institute Inc., 2004