Uniwersytet Warszawski
Wydział Matematyki, Informatyki i Mechaniki

Barbara Rubikowska, Natalia Włodarczyk
Nr albumu: 292151, 292591

Regresja logistyczna – algorytm estymacji współczynników i przykład zastosowania w pakiecie statystycznym *R*

Praca licencjacka
na kierunku MATEMATYKA

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

Lipiec 2012

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 niniejszej pracy przedstawione zostało wprowadzenie do regresji logistycznej. Rozdział pierwszy zawiera wstęp teoretyczny do zagadnienia ze szczególnym uwzględnieniem iteracyjnej metody ważonych najmniejszych kwadratów służącej do wyliczania współczynników regresji. Rozdział drugi zawiera opis funkcji pakietu R przydatnych do budowania modelu regresji logistycznej. Rozdział trzeci zawiera przykład zastosowania pakietu R do tworzenia modeli regresji logistycznej objaśniających wpływ takich czynników jak wiek, płeć, wykształcenie na zainteresowanie kinem i książkami jako nośnikami kultury. Potrzebne dane statystyczne zostały zaczerpnięte z badania „Obiegi kultury".

Słowa kluczowe

statystyka, regresja logistyczna, uogólnione modele liniowe, iteracyjna metoda ważonych najmniejszych kwadratów, szansa, pakiet R

Dziedzina pracy (kody wg programu Socrates-Erasmus)

11.2 Statystyka

Klasyfikacja tematyczna

62J12, 62P25

Tytuł pracy w języku angielskim

Logistic regression – coefficients estimation algorithm with an example of application in statistical package *R*

Spis treści

Spis rysunków

Spis tablic

Udział w przygotowaniu pracy

Praca została napisana przez nas obie – Barbarę Rubikowską i Natalię Włodarczyk. Wszystkie badania przeprowadzałyśmy wspólnie i razem dyskutowałyśmy ich wyniki; nawzajem recenzowałyśmy swoje rozdziały i poprawiałyśmy błędy. Wkład w tworzenie pracy był równomierny. Niełatwo jest wskazać, która z nas była pomysłodawczynią poszczególnych inicjatyw i rozwiązań, jednakże możliwe jest wskazanie autorstwa każdej z sekcji.

Barbara Rubikowska napisała rozdziały:

- Wprowadzenie,
- 1.2 Iteracyjna metoda ważonych kwadratów,
- 2.1 Funkcja `glm()`,
- 3.1 Opis danych,
- 3.2 Transformacje danych,
- 3.3.1 Model badający zainteresowanie kinem,
- dodatek A.

Natalia Włodarczyk napisała rozdziały:

- 1.1 Regresja logistyczna,
- 2.2 Funkcja `step()`,
- 2.3 Inne funkcje przydatne w analizie danych rzeczywistych,
- 3.3.2 Model badający zainteresowanie książkami – zmienna $czy.ksiazki0$,
- 3.3.3 Model badający zainteresowanie książkami – zmienna $czy.ksiazki10$,
- 3.4 Diagnostyka modeli,
- Podsumowanie.

Następujące rozdziały zostały napisane wspólnie:

- 3.5 Interpretacja modeli,
- 3.6 Porównanie modeli i wnioski.

Wprowadzenie

Regresja jest to metoda statystyczna służąca do badania związku pomiędzy zmiennymi (dokładniej: wpływu wybranych zmiennych objaśniających na zmienną objaśnianą) oraz do predykcji nieznanych wartości zmiennej objaśnianej w zależności od znanych wartości zmiennych objaśniających ([8]).

Regresja logistyczna to model regresji używany w przypadku, gdy zmienna objaśniana jest binarna, czyli przyjmuje dokładnie dwie różne wartości (zwyczajowo jedna z tych wartości nazywana jest „sukcesem", a druga „porażką"). Jest to szeroko stosowana metoda statystyczna, zwłaszcza w medycynie, naukach przyrodniczych i społecznych. W naszej pracy postanowiłyśmy zająć się zagadnieniem społeczno-kulturowym i użyć modelu regresji logistycznej do badania czynników różnicujących zainteresowanie dwoma nośnikami kultury: kinem i książkami. Dane zaczerpnęłyśmy z sondażu „Obiegi kultury" zrealizowanego w 2011 roku w ramach programu „Obserwatorium Kultury".

Praca składa się z trzech rozdziałów. Pierwszy z nich stanowi wstęp teoretyczny do zagadnienia regresji logistycznej z naciskiem na pojęcia takie jak funkcja logistyczna, szansa i funkcja wiarogodności, na metody sprawdzania poprawności modelu oraz na iteracyjną metodę ważonych najmniejszych kwadratów zaimplementowaną przez pakiet statystyczny R do estymacji współczynników regresji. Drugi rozdział przedstawia funkcje pakietu R przydatne w budowaniu modelu regresji ze szczególnym uwzględnieniem funkcji `glm()` i `step()`. Trzeci rozdział zawiera analizę danych rzeczywistych – budowę, diagnostykę i interpretację modeli opisujących wpływ zmiennych takich jak płeć, wiek czy stan cywilny na to, czy dana osoba chodzi do kina i czy czyta książki. Do pracy dołączony jest dodatek A, w którym zostały zamieszczone wraz z komentarzami wszystkie kody pakietu R zastosowane przy tworzeniu pracy.

Rozdział 1

Wstęp teoretyczny

1.1. Regresja logistyczna

Regresja opisuje zależność zmiennej objaśnianej (zależnej) Y od wektora zmiennych objaśniających (niezależnych) X. Służy do określania związków pomiędzy poszczególnymi wartościami zmiennych. Zazwyczaj ogólny model przyjmuje postać:

$$Y = \phi(X, \beta) + \epsilon. \tag{1.1}$$

Aby taki model stworzyć, musimy wyestymować współczynniki β stojące przy zmiennych z wektora X. Więcej na temat regresji w postaci ogólnej można przeczytać w [5] s. 143-144.

Trochę inaczej wygląda model dla uogólnionych modeli liniowych (ang. generalized linear models, GLM). Dla tego rodzaju regresji definiujemy funkcję wiążącą, zależną od wartości oczekiwanej zmiennej Y ($\mathbb{E}Y$) i równą iloczynowi $X\beta$.

Odtąd będziemy się zajmować szczególnym modelem regresji – regresją logistyczną. Dla ułatwienia zapisu przyjmiemy, że zmienna X jest jednowymiarowa.

1.1.1. Funkcja logistyczna

Regresja logistyczna jest modelem regresji, w którym zmienna objaśniana ma rozkład dwumianowy:

$$Y \sim Bin(m, \theta). \tag{1.2}$$

Każda z m prób osiąga „sukces" z prawdopodobieństwem θ i „porażkę" z prawdopodobieństwem $1-\theta$, czyli Y jest sumą m niezależnych zmiennych binarnych. Dzięki wyestymowanym współczynnikom β funkcja logistyczna pozwala obliczyć prawdopodobieństwo sukcesu w zależności od wartości zmiennej objaśniającej X. Określa się ją wzorem:

$$\theta(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}. \tag{1.3}$$

1.1.2. Szansa

Szansa (ang. odds) jest funkcją prawdopodobieństwa daną wzorem:

$$Odds(\theta(x)) = \frac{\theta(x)}{1 - \theta(x)}. \tag{1.4}$$

Jest to iloraz prawdopodobieństwa sukcesu i prawdopodobieństwa porażki. Szansa przyjmuje wartości ze zbioru $(0, \infty)$. Łatwo można ten zbiór poszerzyć do $(-\infty, \infty)$ biorąc logarytm

szansy, zwany logitem. Logit jest funkcją wiążącą dla regresji logistycznej i przyjmuje on postać:

$$Logit = \ln\left(\frac{\theta(x)}{1-\theta(x)}\right) = \beta_0 + \beta_1 x = X\beta. \tag{1.5}$$

Widzimy więc, że logit jest liniowo zależny od zmiennej X, co pozwala nam na łatwą interpretację współczynnika β_1: jeśli x rośnie o k jednostek, to szansa wzrasta $\exp(k * \beta_1)$ razy.

1.1.3. Wiarogodność

Dla m niezależnych obserwacji definiujemy funkcję wiarogodności (ang. likelihood) wzorem:

$$L(x) = \prod_{i=1}^{n} P(Y_i = y_i|x_i) = \prod_{i=1}^{n} \binom{m_i}{y_i} \theta(x_i)^{y_i}(1-\theta(x_i))^{m_i - y_i}, \tag{1.6}$$

gdzie m_i to liczba prób w obrębie i-tej obserwacji, a y_i jest liczbą sukcesów w m_i próbach. Ponieważ wykorzystuje się ją do znajdowania estymatorów największej wiarogodności poprzez maksymalizację, to częściej korzystamy z jej logarytmu, co zazwyczaj znacznie upraszcza obliczenia:

$$l(x) = \ln(L(x)) = \sum_{i=1}^{n} \left[y_i(\beta_0 + \beta_1 x_i) - m_i \ln(1 + \exp(\beta_0 + \beta_1 x_i)) + \ln\binom{m_i}{y_i} \right]. \tag{1.7}$$

Maksymalizację log-wiarogodności przeprowadza się przy użyciu metod iteracyjnych, które opiszemy w dalszych rozdziałach np. algorytm Newtona-Raphsona, iteracyjna metoda ważonych najmniejszych kwadratów (patrz 1.2).

1.1.4. Poprawność modelu

Gdy chcemy użyć jakiegoś modelu, to zawsze sprawdzamy jego poprawność, to znaczy czy jest on odpowiednio dobrany do naszych danych. Można to robić na wiele sposobów. Przedstawimy teraz kilka z nich.

Podstawowym pojęciem, które zdefiniujemy jest *dewiancja* (ang. deviance). Jest ona odchyleniem przyjętego modelu od modelu wysyconego (ang. saturated model), czyli takiego, w którym liczba funkcyjnie niezależnych parametrów β jest równa liczbie obserwacji. *Dewiancja* jest podwojoną różnicą log-wiarogodności tych dwóch modeli. Obliczamy ją za pomocą statystyki G^2 zadanej wzorem:

$$\begin{aligned} G^2 &= 2[\ln(L_S) - \ln(L_M)] = \\ &= 2\sum_{i=1}^{n}\left[y_i \ln\left(\frac{y_i}{m_i}\right) + (m_i - y_i)\ln\left(1 - \frac{y_i}{m_i}\right)\right] - 2\sum_{i=1}^{n}\left[y_i \ln\left(\frac{\hat{y}_i}{m_i}\right) + (m_i - y_i)\ln\left(1 - \frac{\hat{y}_i}{m_i}\right)\right] = \\ &= 2\sum_{i=1}^{n}\left[y_i \ln\left(\frac{y_i}{\hat{y}_i}\right) + (m_i - y_i)\ln\left(\frac{m_i - y_i}{m_i - \hat{y}_i}\right)\right], \end{aligned} \tag{1.8}$$

gdzie $\hat{y}_i = m_i\hat{\theta}(x_i)$, a L_S i L_M to log-wiarogodności odpowiednio modelu wysyconego i modelu przyjętego.
H_0: $\beta_i = \hat{\beta}_i$ dla $i = 1, ..., n$, gdzie n jest liczbą współczynników wyestymowanych w przyjętym modelu,
przeciwko alternatywie:

H_1: $\beta_i = \hat{\beta}_i$ dla $i = 1, ..., n$, gdzie n jest liczbą współczynników wyestymowanych w modelu wysyconym.
Dla dostatecznie dużych m_i, przy założeniu, że H_0 jest prawdziwa, G^2 ma rozkład asymptotycznie zbieżny do χ^2_d, gdzie liczba stopni swobody $d = n - p - 1$, gdzie $p + 1$ jest liczbą estymowanych współczynników β ([6] s. 274). W pakiecie R taka dewiancja jest nazywana *dewiancją resztową* (ang. residual deviance).

Możemy też sprawdzać, czy nasz model odbiega od modelu zerowego, czyli takiego, w którym współczynnik β_1 jest równy 0. W pakiecie R takie odchylenie jest wyznaczone przez *dewiancję zerową* (ang. null deviance).

Jeszcze innym sposobem jest obliczenie statystyki, będącej różnicą dewiancji resztowej i zerowej, która testuje:
H_0: $\beta_1 = 0$ (czyli $\theta(x) = \frac{1}{1+\exp(-\beta_0)}$),
przeciwko:
H_1: $\beta_1 \neq 0$ (czyli $\theta(x) = \frac{1}{1+\exp(-[\beta_0+\beta_1 x])}$).
Przy założeniu prawdziwości H_0 taka różnica ma rozkład asymptotyczny χ^2_p, gdzie p jest liczbą estymowanych β_i dla $i \neq 0$.

R^2, zwany współczynnikiem determinacji, określa stopień dopasowania modelu do danych (ang. goodness-of-fit). Istnieje kilka definicji R^2, my użyjemy dwóch do diagnozy modeli w rozdziale 3.4. Jedną z nich jest współczynnik dla regresji logistycznej, określany wzorem:

$$R^2_P = 1 - \frac{G^2_{H_1}}{G^2_{H_0}}. \tag{1.9}$$

Drugą jest R^2_N Nagerkelkego, który można łatwo obliczyć w pakiecie R, o czym będzie mowa w rozdziale 2.2. R^2 przyjmuje wartości od 0 do 1. Im większy R^2, tym lepsze dopasowanie modelu do danych.

Statystyka X^2 Pearsona również mierząca dopasowanie modelu jest zdefiniowana następująco:

$$X^2 = \sum_{i=1}^{n} \frac{\left(\frac{y_i}{m_i} - \hat{\theta}(x_i)\right)^2}{\hat{V}ar\left(\frac{y_i}{m_i}\right)}. \tag{1.10}$$

Ta statystyka ma rozkład asymptotyczny χ^2_d, gdzie $d = n - p - 1$ dla dostatecznie dużych m_i ([6] s. 274). Widzimy więc, że jest to taki sam rozkład jak dla dewiancji.

Ostatnim sposobem jest obliczenie *reszt* (ang. residuals). Przedstawimy 3 rodzaje:

- *Reszty objaśniane* (ang. response residuals):
  $r_i = \frac{y_i}{m_i} - \hat{\theta}(x_i)$,
  gdzie $\hat{\theta}(x_i)$ jest i-tą dopasowaną z modelu wartością.
- *Reszty Pearsona*:
  $r_{Pearson,i} = \frac{\frac{y_i}{m_i} - \hat{\theta}(x_i)}{\sqrt{\hat{V}ar\left(\frac{y_i}{m_i}\right)}}$.

Warto zauważyć, że $\sum_{i=1}^{n} r_{Pearson,i}^2 = X^2$, gdzie X^2 jest statystyką Pearsona (stąd nazwa reszt).

- *Reszty dewiancyjne*:
  $r_{Deviance,i} = sign\left(\frac{y_i}{m_i} - \hat{\theta}(x_i)\right) g_i$,
  gdzie $\sum_{i=1}^{n} g_i^2 = G^2$.

Trochę więcej o resztach można przeczytać w [6] s. 274-277.

1.1.5. Regresja logistyczna dla danych binarnych

Jest to specyficzna odmiana regresji logistycznej, w której liczba poszczególnych obserwacji jest równa 1 ($m_i = 1$). Czyli zmienne Y_i, są zmiennymi o rozkładzie dwupunktowym (binarnym). Tak przedstawione dane nazywamy binarnymi. Możemy teraz przedefiniować wielkości dostosowane do tego typu danych.

- Log-wiarogodność dla danych binarnych:
$$\ln(L) = \sum_{i=1}^{n} \left[y_i \ln(\theta(x_i)) + (1 - y_i) \ln(1 - \theta(x_i)) + \ln \binom{1}{y_i} \right]. \tag{1.11}$$

- Dewiancja dla danych binarnych wraz z wyprowadzeniem (L_M - model otrzymany przez nas, L_S - model wysycony):
$$\begin{aligned} G^2 &= 2[\ln(L_S) - \ln(L_M)] = \\ &= 2\sum_{i=1}^{n} [y_i \ln(y_i) + (1 - y_i) \ln(1 - y_i)] - 2\sum_{i=1}^{n} [y_i \ln(\hat{y}_i) + (1 - y_i) \ln(1 - \hat{y}_i)] = \\ &= -2\sum_{i=1}^{n} [y_i \ln(\hat{y}_i) + (1 - y_i) \ln(1 - \hat{y}_i)]. \end{aligned} \tag{1.12}$$

  Jest ona zależna tylko od logarytmu modelu przyjętego (dla modelu wysyconego wynosi on 0). Ponieważ wszystkie m_i są równe 1, to nie możemy mówić o zbieżności asymptotycznej do rozkładu χ^2, jak to miało miejsce dla zwykłej dewiancji. Jednak już dla różnicy dewiancji, ta asymptotyczność zachodzi ([6] s. 280-281).

Więcej na temat tego rodzaju danych można przeczytać w [6] s. 277-294.

1.2. Iteracyjna metoda ważonych kwadratów

W tym rozdziale opiszemy iteracyjną metodę ważonych kwadratów – technikę stosowaną do estymowania metodą największej wiarogodności współczynników uogólnionych modeli liniowych, w tym regresji logistycznej.

Wyznaczenie estymatora największej wiarogodności (ang. maximum likelihood estimate – MLE) sprowadza się do odnalezienia tych wartości parametrów, dla których funkcja wiarogodności osiąga maksimum, czyli miejsc zerowych pierwszej pochodnej funkcji wiarogodności. Prawdopodobnie najszerzej stosowaną numeryczną metodą służącą do rozwiązania tego problemu jest metoda Newtona-Raphsona, oparta na metodzie Newtona odnajdującej miejsca zerowe funkcji.

1.2.1. Metoda Newtona

Metoda Newtona jest to iteracyjny algorytm wyznaczania przybliżonej wartości pierwiastka funkcji. Przedstawimy jej teoretyczne założenia.

Rozwinięcie Taylora funkcji f wokół punktu x_0 dane jest wzorem:

$$f(x_1) = f(x_0) + (x_1 - x_0)f'(x_0) + \frac{1}{2!}(x_1 - x_0)^2 f''(x_0) + \frac{1}{3!}(x_1 - x_0)^3 f^{(3)}(x_0) + \cdots . \quad (1.13)$$

Załóżmy, że szukamy punktu x_1 takiego, że $f(x_1) = 0$. Wówczas spełnione jest równanie:

$$0 = f(x_0) + (x_1 - x_0)f'(x_0) + \frac{1}{2!}(x_1 - x_0)^2 f''(x_0) + \frac{1}{3!}(x_1 - x_0)^3 f^{(3)}(x_0) + \cdots . \quad (1.14)$$

W większości wykorzystywanych w statystyce metod iteracyjnych tylko dwa pierwsze składniki powyższej sumy są brane pod uwagę. Pomijając pozostałe składniki sumy otrzymujemy przybliżenie szukanego punktu:

$$0 \approx f(x_0) + (x_1 - x_0)f'(x_0). \quad (1.15)$$

Rozwikłując to wyrażenie ze względu na x_1 otrzymujemy:

$$x_1 \approx x_0 - \frac{f(x_0)}{f'(x_0)}. \quad (1.16)$$

Jeżeli za x_0 wstawimy j-te przybliżenie szukanego punktu, a za x_1 wstawimy kolejne $(j+1)$-sze przybliżenie, to otrzymamy wzór na $(j+1)$-szy krok iteracyjnej metody Newtona:

$$x^{(j+1)} = x^{(j)} - \frac{f(x^{(j)})}{f'(x^{(j)})}. \quad (1.17)$$

Warunkiem kończącym iteracje jest:

$$|f(x^{(j)})| < \epsilon. \quad (1.18)$$

Metoda Newtona jest zbieżna kwadratowo – błąd maleje kwadratowo wraz z liczbą iteracji ([2], s. 41).

1.2.2. Algorytm Newtona-Raphsona

Zdefiniujmy funkcję wynikową (ang. score function) – jest to pierwsza pochodna log-wiarogodności:

$$\dot{l}(\theta|Y) = \frac{\partial}{\partial\theta} l(\theta|Y). \quad (1.19)$$

Algorytm Newtona-Raphsona jest to metoda Newtona zastosowana do funkcji wynikowej. Z powyższej definicji wynika, że owy algorytm odnajduje punkty krytyczne log-wiarogodności (punkty, w których zeruje się pierwsza pochodna log-wiarogodności) będące jednocześnie punktami krytycznymi funkcji wiarogodności. Kolejne oszacowania estymatora są wyznaczane w następujących iteracjach:

$$\theta^{(j+1)} = \theta^{(j)} - \frac{\dot{l}(\theta^{(j)}|\mathbf{y})}{\frac{\partial}{\partial\theta}\dot{l}(\theta^{(j)}|\mathbf{y})} = \theta^{(j)} - \frac{\frac{\partial}{\partial\theta} l(\theta^{(j)}|\mathbf{y})}{\frac{\partial^2}{\partial\theta^2} l(\theta^{(j)}|\mathbf{y})}. \quad (1.20)$$

Uogólnijmy powyższy wzór na przypadek wielowymiarowego parametru $\boldsymbol{\theta}$:

$$\boldsymbol{\theta}^{(j+1)} = \boldsymbol{\theta}^{(j)} - \frac{\partial}{\partial \boldsymbol{\theta}} l(\boldsymbol{\theta}^{(j)}|\mathbf{y}) \left(\frac{\partial^2}{\partial \boldsymbol{\theta}^2} l(\boldsymbol{\theta}^{(j)}|\mathbf{y}) \right)^{-1}. \tag{1.21}$$

Czasami macierz Hessego, $\mathbf{H} = \frac{\partial^2}{\partial \boldsymbol{\theta}^2} l(\boldsymbol{\theta}^{(j)}|\mathbf{y})$, zastępuje się łatwiejszą do policzenia jej wartością oczekiwaną: $\mathbf{A} = \mathbb{E}_\theta \left(\frac{\partial^2}{\partial \boldsymbol{\theta}^2} l(\boldsymbol{\theta}^{(j)}|\mathbf{y}) \right)$. Algorytm po tej modyfikacji nazywa się algorytmem oceniania Fishera (ang. Fisher scoring algorithm).

W każdym kroku algorytmu Newtona-Raphsona rozwiązywane jest następujące równanie:

$$(\boldsymbol{\theta}^{(j+1)} - \boldsymbol{\theta}^{(j)})\mathbf{A} = -\frac{\partial}{\partial \boldsymbol{\theta}^{(j)}} l(\boldsymbol{\theta}^{(j)}|\mathbf{y}). \tag{1.22}$$

1.2.3. Metoda najmniejszych kwadratów

Na początek przyjrzyjmy się bliżej modelowi regresji liniowej z r zmiennymi objaśniającymi zapisanemu w postaci macierzowej:

$$Y = \mathbf{X}\beta + \varepsilon. \tag{1.23}$$

Wyjaśnienie znaczenia poszczególnych zmiennych w równaniu modelu:

- ε jest wektorem błędów losowych ϵ_i (zmiennych losowych); dla każdego i zachodzi $\mathbb{E}\epsilon_i = 0$; $Var\varepsilon \sim \sigma^2 Id$, czyli dla każdego i zachodzi $Var\epsilon_i = \sigma$,
- Y jest wektorem wartości zmiennej objaśnianej odnotowanych w kolejnych obserwacjach; $\mathbb{E}Y = \mathbf{X}\beta$, $VarY \sim \sigma^2 Id$,
- macierz $\mathbf{X}$ jest nazywana macierzą planu, x_{ij} jest wartością j-tej zmiennej objaśniającej w i-tej obserwacji; pierwsza kolumna macierzy $\mathbf{X}$ jest złożona z jedynek, bo rozpatrujemy model regresji z wyrazem wolnym,
- β jest wektorem nieznanych parametrów, $\beta^T = (\beta_0, \beta_1, \cdots, \beta_r)$; rozwiązanie zagadnienia regresji polega na wyestymowaniu współczynników β_j na podstawie wartości zmiennych objaśniających i zmiennej objaśnianej w kolejnych próbach.

W teoretycznych rozważaniach dotyczących wyliczania wektora parametrów zakłada się, że macierz $\mathbf{X}$ jest pełnego rzędu ([5], s. 151). Przy tym założeniu możemy wyliczyć w jawnej macierzowej postaci estymator najmniejszych kwadratów (ang. least squares estimate – LSE) parametru β, minimalizując sumę kwadratów błędów (ang. sum of squared errors – SSE):

$$SSE(\beta) = \sum_{i=1}^{n} (Y_i - x_i^T \beta)^2 = (\mathbf{X}\beta - \mathbf{Y})^T (\mathbf{X}\beta - \mathbf{Y}). \tag{1.24}$$

Obliczając gradient względem β, otrzymujemy $\mathbf{X}^T(\mathbf{X}\beta - Y) = 0$, czyli $\mathbf{X}^T\mathbf{X}\beta = \mathbf{X}^T\mathbf{Y}$. Wcześniejsze założenie gwarantuje, że macierz $\mathbf{X}^T\mathbf{X}$ jest odwracalna i otrzymujemy następujący wzór na estymator najmniejszych kwadratów parametru β:

$$LSE(\beta) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}. \tag{1.25}$$

1.2.4. Metoda ważonych najmniejszych kwadratów

Rozważania w poprzednim rozdziale zakładały homoskedastyczność błędu losowego (taką samą wartość wariancji błędu dla każdej z prób). Rozpatrzmy teraz model, w którym błędy poszczególnych obserwacji nie mają stałej wariancji, czyli są heteroskedastyczne:

$$Y = \mathbf{X}\beta + \varepsilon. \tag{1.26}$$

Różni się on od poprzedniego modelu tym, że w wektorze ε błędy mają różne wariancje: $Var\epsilon_i = v_i$. Aby sprowadzić ten model do poprzedniego, uwzględnia się w rozważaniach diagonalną macierz wag $\mathbf{\Omega}$, taką że: $\mathbf{\Omega}_{ii} = \frac{1}{v_i}$, gdzie v_i to wariancja i-tej obserwacji. Wówczas obserwacje z większą wariancją błędu mają mniejszą wagę, więc są mniej brane pod uwagę przy estymowaniu współczynników. Skoro $\mathbf{\Omega}$ jest macierzą odpowiadającą za wariancję, to macierzą odpowiadającą za odchylenie standardowe jest macierz $\mathbf{\Omega}^{\frac{1}{2}}$, również diagonalna, taka że: $\mathbf{\Omega}^{\frac{1}{2}}_{ii} = \sqrt{\frac{1}{v_i}}$. Po przemnożeniu równania naszego modelu przez macierz $\mathbf{\Omega}^{\frac{1}{2}}$, otrzymujemy następujący nowy model:

$$\mathbf{\Omega}^{\frac{1}{2}}\mathbf{Y} = \mathbf{\Omega}^{\frac{1}{2}}\mathbf{X}\beta + \mathbf{\Omega}^{\frac{1}{2}}\varepsilon. \tag{1.27}$$

Łatwo zauważyć, że w tak zdefiniowanym modelu składnik $\mathbf{\Omega}^{\frac{1}{2}}\varepsilon$ odpowiadający za błąd losowy jest homoskedastyczny, czyli udało nam się sprowadzić nowy model do poprzedniego. Suma kwadratów błędów jest następująca:

$$SSE(\beta) = (\mathbf{X}\beta - \mathbf{Y})^T\mathbf{\Omega}^{-1}(\mathbf{X}\beta - \mathbf{Y}). \tag{1.28}$$

Obliczając gradient względem β, otrzymujemy $\mathbf{X}^T\mathbf{\Omega}(\mathbf{X}\beta - Y) = 0$, czyli $\mathbf{X}^T\mathbf{\Omega}\mathbf{X}\beta = \mathbf{X}^T\mathbf{\Omega}\mathbf{Y}$.

Rozwiązując takie zagadnienie otrzymujemy następujący wzór na estymator ważonych najmniejszych kwadratów (ang. weighted least squares estimate – WLSE) parametru β:

$$WLSE(\beta) = (\mathbf{X}^T\mathbf{\Omega}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{\Omega}\mathbf{Y}. \tag{1.29}$$

Zauważmy, że jeśli błędy są homoskedastyczne, to model z macierzą wag $\mathbf{\Omega}$ sprowadza się do modelu, w którym można tę macierz pominąć.

1.2.5. Iteracyjna metoda ważonych najmniejszych kwadratów

Przypuśćmy, że wartości wariancji błędów losowych potrzebne do wyznaczenia wartości na diagonali macierzy $\mathbf{\Omega}$ nie są znane i nie jesteśmy w stanie ich łatwo wyestymować. W iteracyjnej metodzie ważonych najmniejszych kwadratów (ang. iteratively weighted least squares – IWLS) zakłada się, że poszczególne wariancje są funkcjami wartości oczekiwanych zmiennej objaśnianej:

$$v_i = f(\mathbb{E}\mathbf{Y}_i) = f(\mu_i). \tag{1.30}$$

Postać funkcji f można wywnioskować z postaci modelu – zajmujemy się uogólnionymi modelami liniowymi, czyli takimi, w których ustalona jest postać funkcji g (funkcji wiążącej, ang. link function) takiej że:

$$\mathbb{E}\mathbf{Y}_i = \mu_i = g^{-1}(\mathbf{X}\beta). \tag{1.31}$$

Jeśli znana jest wartość oczekiwana μ zmiennej objaśnianej i postać funkcji f, to istnieje bezpośrednia iteracyjna procedura estymująca współczynniki. Polega ona na iteracyjnym estymowaniu wag przy użyciu funkcji wartości oczekiwanej zmiennej objaśniającej μ_i. Znając $\mu_i^{(j)}$ można wyznaczyć wagi w kroku $(j+1)$-ym, a znając wagi w kroku $(j+1)$-ym można wyliczyć $\mu_i^{(j+1)}$.

Oto schemat procedury:

1. Ustalenie wag początkowych: $v_i^{(1)} = 1$ (model regresji nieważonej) i konstrukcja macierzy diagonalnej $\mathbf{\Omega}^{(1)} = Id$,
2. Estymacja parametru $\beta^{(j)}$ przy użyciu algorytmu ważonych kwadratów dla macierzy wag $\mathbf{\Omega}^{(j)}$: $\beta^{(j)} = (\mathbf{X}^T\mathbf{\Omega}^{(j)}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{\Omega}^{(j)}\mathbf{Y}$,
3. Aktualizacja wag wg wzoru: $\frac{1}{v_i^{(j+1)}} = Var\mu_i^{(j)}$, aktualizacja macierzy wag: $\mathbf{\Omega}_{ii}^{(j+1)} = \frac{1}{v_i^{(j+1)}}$,
4. Krok 2 i 3 powtarzany jest aż do momentu gdy $\|\mathbf{X}\beta^{(j+1)} - \mathbf{X}\beta^{(j)}\| < \epsilon$.

Iteracyjna metoda ważonych najmniejszych kwadratów przybliża MLE dla nieznanego wektora parametrów β ([2], s. 44).

Znając postać funkcji wiążącej dla naszego modelu możemy rozszerzyć zagadnienie 1.22 z rozdziału o algorytmie Newtona-Raphsona na uogólnione modele liniowe:

$$(\boldsymbol{\theta}^{(j+1)} - \boldsymbol{\theta}^{(j)})\mathbf{A} = -\frac{\partial l(\boldsymbol{\theta}^{(j)}|\mathbf{y})}{\partial g^{-1}(\boldsymbol{\theta}^{(j)})}\frac{\partial g^{-1}(\boldsymbol{\theta}^{(j)})}{\partial \boldsymbol{\theta}^{(j)}}. \tag{1.32}$$

Łatwo zauważyć, że w przypadku modelu liniowego, gdy funkcja wiążąca jest identycznością, powyższe równanie sprowadza się do równania 1.22.

Procedura IWLS sprowadza się do metody Newtona-Raphsona z ocenianiem Fishera (Fisher scoring) zastosowanej iteracyjnie do powyższego równania ([2], s. 45).

Rozdział 2

Opis funkcji w pakiecie R

2.1. Funkcja `glm()`

Funkcja `glm()` służy do budowy uogólnionych modeli liniowych.

Jeśli podamy argumenty `family="binomial"` (rodzina rozkładów dwumianowych) oraz `link=logit` (deklarujemy funkcję logit jako funkcję wiążącą – przy zadeklarowaniu rozkładu `"binomial"` funkcja wiążąca logit jest wybierana domyślnie), to zbudujemy model regresji logistycznej.

Oto przykład wywołania funkcji `glm()` w celu budowy modelu regresji logistycznej, w którym zmienną objaśnianą jest zmienna *czy.ksiazki*0 opisana w rozdziale 3.2.2, a zmiennymi objaśniającymi są zmienne *fac.plec*, *fac.wiek* i *fac.wyksztalcenie* opisane w rozdziale 3.2.3:

```
> regresja<-glm(formula=czy.ksiazki0~fac.plec+fac.wiek+fac.wyksztalcenie,
  family=binomial())
> summary(regresja)
```

Funkcja `summary()` służy do generowania podsumowań wyników różnych funkcji dopasowujących model.

W tym przypadku rezultat powyższych operacji jest następujący:

```
Call:
glm(formula = czy.ksiazki0 ~ fac.plec + fac.wiek + fac.wyksztalcenie,
    family = binomial())

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7514   0.2185   0.3388   0.5262   0.9637

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)         2.21314    0.52956   4.179 2.93e-05 ***
fac.plec1           0.93329    0.22595   4.131 3.62e-05 ***
fac.wiek2          -0.51039    0.24064  -2.121   0.0339 *
fac.wiek3          -0.03878    0.31324  -0.124   0.9015
fac.wyksztalcenie2 -1.17676    0.59088  -1.992   0.0464 *
fac.wyksztalcenie3 -0.30578    0.55517  -0.551   0.5818
fac.wyksztalcenie4  0.61582    0.58332   1.056   0.2911
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

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

```
    Null deviance: 841.82  on 1283  degrees of freedom
Residual deviance: 753.86  on 1277  degrees of freedom
AIC: 767.86

Number of Fisher Scoring iterations: 6
```

Na początku przedstawione są reszty dewiancyjne – minimum, maksimum oraz kwartyle.

Następnie przytoczone są wszystkie oceny współczynników dla kolejnych zmiennych objaśniających, oceny odchyleń standardowych tych ocen, wartości statystyki testowej dla testu, w którym hipotezą zerową jest nieistotność danej zmiennej wraz z p-wartością dla tego testu. Im niższa p-wartość, tym zmienna istotniejsza, co graficznie przedstawiają gwiazdki po prawej stronie.

Ponadto przytoczone są wartości dewiancji zerowej oraz resztowej i kryterium informacyjnego Akaike'a (`AIC`) oraz liczba wykonanych iteracji algorytmu estymującego współczynniki (iteracyjna metoda ważonych najmniejszych kwadratów – algorytm oceniania Fishera).

W wyniku zastosowania funkcji `glm()` powstaje obiekt klasy `glm`. Oto niektóre z właściwości tego obiektu wraz z przykładami ich wywołania dla zbudowanego wyżej modelu:

- `$coefficients` – wyestymowane współczynniki:

```
> regresja$coefficients
       (Intercept)          fac.plec1           fac.wiek2
         2.2131390          0.9332928         -0.5103879
         fac.wiek3 fac.wyksztalcenie2 fac.wyksztalcenie3
        -0.0387839         -1.1767590         -0.3057840
fac.wyksztalcenie4
         0.6158169
```

- `$fitted.values` – dopasowane przez model prawdopodobieństwa „sukcesu" dla każdej z prób – wartości dopasowane przez model przekształcone przez funkcję wiążącą:

```
> regresja$fitted.values
        1         2         3         4         5         6
0.8775760 0.9427765 0.8017022 0.6285479 0.8707217 0.6285479
        7         8         9        10        11        12
0.6285479 0.6285479 0.9448332 0.9772960 0.9587679 0.9587679
(...)
```

- `$linear.predictors` – dla każdej z prób wartości dopasowane przez model przed przekształceniem przez funkcję wiążącą:

```
> regresja$linear.predictors
        1         2         3         4         5         6
1.9696729 2.8018639 1.3969671 0.5259921 1.9073550 0.5259921
        7         8         9        10        11        12
0.5259921 0.5259921 2.8406478 3.7622488 3.1464319 3.1464319
(...)
```

- `$residuals` – reszty dla każdej z prób:

```
> regresja$residuals
         1          2          3          4          5          6
  1.139502   1.060697   1.247346  -2.692137   1.148473   1.590969
         7          8          9         10         11         12
  1.590969   1.590969   1.058388   1.023231   1.043005   1.043005
(...)
```

- `$family` – rodzina rozkładów zmiennej objaśnianej oraz użyta funkcja wiążąca:

```
> regresja$family
Family: binomial
Link function: logit
```

- `$aic` – wartość kryterium informacyjnego Akaike'a dla modelu:

```
> regresja$aic
[1] 767.8587
```

- `$iter` – liczba wykonanych iteracji algorytmu IWLS:

```
> regresja$iter
[1] 6
```

2.2. Funkcja `step()`

W tym podrozdziale przedstawimy funkcję `step()`, drugą ważną funkcję przydatną w tworzeniu modelu regresji logistycznej. Tak wygląda jej deklaracja:

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

Funkcja `step()` stosując metodę krokową kolejno usuwa lub dodaje do modelu zmienne objaśniające. Na podstawie pewnego kryterium wybiera albo zmienną najmniej istotną i ją odrzuca albo najbardziej istotną i ją do tego modelu dodaje i robi to tak długo, aż w modelu będą same zmienne istotne. Domyślnym kryterium jest AIC (Akaike'a), ale możemy je zmienić dzięki zmiennej `k` (np. dla `k=log(n)`, gdzie n to liczba obserwacji, stosowane będzie BIC, czyli kryterium Schwarza). Kolejnym argumentem funkcji jest `step`. Określa on liczbę kroków dopuszczalnych w jednej operacji, domyślnie ustawiony jest na 1000. Metoda, którą będzie stosować funkcja zależy od argumentu `direction`. Może on przyjmować wartości:

- `backward` - funkcja zaczyna od pełnego modelu i stopniowo odrzuca zmienne, które są nieistotne,
- `forward` - funkcja zaczyna od modelu złożonego tylko z wyrazu wolnego i stopniowo dokłada do niego zmienne, które mogą być istotne,
- `both` - funkcja zaczyna od modelu złożonego tylko z wyrazu wolnego, dodaje do niego zmienną z najmniejszą p-wartością (jeśli kryterium AIC jest spełnione), a następnie usuwa zmienną z największą p-wartością (jeśli kryterium AIC nie jest spełnione).

Wszystkie te czynności są wykonywane dopóki nie będzie już czego dodać, ani czego usunąć z modelu. Przykładowe użycie funkcji `step()`:

```
Start:  AIC=778.12
czy.ksiazki0 ~ fac.plec + fac.wiek + fac.wyksztalcenie + fac.miejscowosc +
    fac.liczbaosob + fac.stan

                    Df Deviance    AIC
- fac.liczbaosob     2   750.67 774.67
- fac.miejscowosc    2   750.68 774.68
- fac.stan           3   753.01 775.01
<none>                   750.12 778.12
- fac.wiek           2   754.14 778.14
- fac.plec           1   768.34 794.34
```

```
- fac.wyksztalcenie  3   784.21 806.21

Step:  AIC=774.67
czy.ksiazki0 ~ fac.plec + fac.wiek + fac.wyksztalcenie + fac.miejscowosc +
    fac.stan

                    Df Deviance    AIC
- fac.miejscowosc    2   751.22 771.22
- fac.stan           3   753.25 771.25
<none>                   750.67 774.67
- fac.wiek           2   754.93 774.93
- fac.plec           1   769.37 791.37
- fac.wyksztalcenie  3   784.30 802.30

Step:  AIC=771.22
czy.ksiazki0 ~ fac.plec + fac.wiek + fac.wyksztalcenie + fac.stan

                    Df Deviance    AIC
- fac.stan           3   753.86 767.86
<none>                   751.22 771.22
- fac.wiek           2   755.73 771.73
- fac.plec           1   771.27 789.27
- fac.wyksztalcenie  3   786.68 800.68

Step:  AIC=767.86
czy.ksiazki0 ~ fac.plec + fac.wiek + fac.wyksztalcenie

                    Df Deviance    AIC
<none>                   753.86 767.86
- fac.wiek           2   760.18 770.18
- fac.plec           1   772.43 784.43
- fac.wyksztalcenie  3   790.13 798.13

Call:  glm(formula = czy.ksiazki0 ~ fac.plec + fac.wiek + fac.wyksztalcenie,
    family = binomial())

Coefficients:
                 (Intercept)               fac.pleckobieta
                     2.21314                       0.93329
               fac.wiek26-40                 fac.wiek41-50
                    -0.51039                      -0.03878
   fac.wyksztalceniezawodowe      fac.wyksztalceniesrednie
                    -1.17676                      -0.30578
fac.wyksztalceniepomaturalne
                     0.61582

Degrees of Freedom: 1283 Total (i.e. Null);  1277 Residual
Null Deviance:      841.8
Residual Deviance: 753.9        AIC: 767.9
```

2.3. Inne funkcje przydatne w analizie danych rzeczywistych

Przedstawimy teraz resztę funkcji, których będziemy używać w rozdziale 3 do analizy danych rzeczywistych.

2.3.1. Funkcja `factor()`

Jedną z przydatnych funkcji jest `factor()`. Możemy nią faktoryzować wektory, czyli ze zmiennej ilościowej tworzyć zmienną jakościową. Składnia tej funkcji wygląda tak:

```
factor(nazwa_zmiennej, labels=levels)
```

gdzie `levels` to wektor składający się z napisów, które mają być nazwami kolejnych poziomów np. c(„malo", „duzo") lub c(„1", „1", „2"). Zwróćmy uwagę, że dzięki tej funkcji możemy łączyć niektóre poziomy, w rezultacie możemy ze zmiennej ilościowej stworzyć nawet zmienną binarną.

2.3.2. Funkcja `plot()`

Funkcją służącą do rysowania wykresów pudełkowych jest `plot()`. Składnia tej funkcji wygląda tak:

```
plot(x, y, xlab="etykieta␣osi␣x", ylab="etykieta␣osi␣y")
```

W naszym przypadku wykres będzie obrazował zależność pomiędzy x – pojedynczą zmienną objaśniającą a y – szansą modelu regresji logistycznej, wyznaczonego przez funkcję `glm()`. Przykładowe użycie tej funkcji można zobaczyć na rysunkach 3.2, 3.3, 3.4.

2.3.3. Obliczanie X^2 i R^2_N

Przy diagnozie modelu będziemy obliczać statystykę Pearsona X^2. W R robi się to za pomocą funkcji:

```
sum(residuals(nazwa_modelu, type="pearson")^2)
```

wiedząc, że X^2 to suma kwadratów reszt Pearsona. Będziemy także znajdować R^2_N Nagelkerkego. Do tego skorzystamy z pakietu `rms` i wbudowanej w niego funkcji `lrm()`:

```
library(rms)
nazwa<-lrm(nazwa_modelu)
nazwa$stats
```

Dla zmiennej *czy.ksiazki*0 dostajemy:

```
     Obs     Max Deriv   Model L.R.          d.f.             P             C
1.284000e+03 4.632794e-11 8.796338e+01 6.000000e+00 1.110223e-16 7.364051e-01
         Dxy         Gamma         Tau-a            R2         Brier             g
4.728103e-01 5.160983e-01 8.611413e-02 1.376914e-01 8.389097e-02 1.020065e+00
          gr            gp
2.773374e+00 8.614185e-02
```

R^2 które widnieje w wyniku to właśnie R^2 Nagelkerkego.

2.3.4. Funkcja `read.table()`

Ostatnia funkcja służy do wczytywania danych z pliku. Jej składnia wygląda następująco:

```
read.table("sciezka_dostepu", sep=";", header=T)
```

Za pomocą argumentu `sep` określamy znak separatora kolejnych kolumn w każdym wierszu. Argument `header` służy do wyszczególnienia, czy w tabeli są nagłówki (`T`), czy ich nie ma (`F`).

Rozdział 3

Analiza danych rzeczywistych

3.1. Opis danych

Dane, z których skorzystałyśmy, pochodzą z badania „Obiegi kultury" zrealizowanego w 2011 roku ze środków Narodowego Centrum Kultury w ramach programu „Obserwatorium Kultury". Badanie skupiało się na zagadnieniu nieformalnego obiegu treści kultury. Autorzy raportu z badania odchodzą od standardowego podziału na legalne i nielegalne obiegi kultury i starają się wykazać, że osoby, które nazywane są internetowymi „piratami", są jednocześnie najbardziej aktywnymi użytkownikami treści szeroko pojętej kultury: książek, filmów, muzyki. Raport „Obiegi kultury" ukazuje „piratów" w nowym świetle – jako tych, którzy zamiast niszczyć tworzą świat kultury.

W ramach projektu przeprowadzono dwa badania sondażowe:

- pilotaż – badanie na reprezentatywnej dla populacji Polski grupie 1004 osób w wieku powyżej 15 lat przeprowadzone metodą osobistego wywiadu wspomaganego komputerowo, które wykazało, że w pozarynkowej wymianie treści kultury udział biorą niemal wyłącznie aktywni internauci,

- sondaż – badanie na reprezentatywnej dla populacji polskich internautów grupie 1284 osób w wieku 16-50 lat przeprowadzone w całości za pośrednictwem Internetu.

Dane zebrane za pomocą pilotażu były bardzo okrojone i nie nadawały się do szerszych analiz, dlatego postanowiłyśmy skorzystać z licznych, szczegółowych danych zgromadzonych za pomocą sondażu. Takie podejście ma słabą stronę – wszystkie uzyskane przez nas wyniki będą prawdziwe dla populacji polskich aktywnych internautów, a nie dla populacji wszystkich Polaków. O tym, jak mogło to zniekształcić wyniki naszych analiz, napiszemy w rozdziale z porównaniem modeli.

W naszych analizach postanowiłyśmy zbadać nie to, co stanowi główne zagadnienie raportu z badań, czyli nieformalne obiegi treści kulturowych. Skupiłyśmy się na modelowaniu poziomu zainteresowania nośnikami kultury takimi jak książki i kino w zależności od pewnych czynników.

Zmiennymi objaśnianymi są:

- prawdopodobieństwo, że dana osoba chodzi do kina,

- prawdopodobieństwo, że dana osoba czyta książki.

Zaproponowane przez nas zmienne objaśniające to:

- płeć,
- wiek,
- wykształcenie,
- miejsce zamieszkania (wieś, małe miasto, duże miasto),
- liczba osób w gospodarstwie domowym,
- stan cywilny.

Rozważałyśmy dołączenie do grupy zmiennych objaśniających przeciętnego dochodu uzyskiwanego przez daną osobę. Niestety znaczna grupa badanych odmówiła odpowiedzi na takie pytanie. Gdybyśmy postanowiły uwzględnić tę zmienną, byłybyśmy zmuszone zawęzić liczebność badanej próby z 1284 osób do około 1000.

Postanowiłyśmy zbadać, które z wyżej wymienionych są zmiennymi istotnie różnicującymi zainteresowanie wybranymi nośnikami kultury oraz jak na to zainteresowanie wpływają. Wszystkie dane potrzebne do analiz pobrałyśmy ze strony internetowej badania „Obiegi kultury", z ogólnodostępnego pakietu z surowymi danymi uzyskanymi z sondażu oraz ze wzoru kwestionariusza.

3.2. Transformacje danych

3.2.1. Pytania z sondażu

Każda wybrana przez nas zmienna jest odpowiedzią na pytanie z sondażu. Tablice 3.1 i 3.2 przedstawiają interesujące nas pytania i możliwe odpowiedzi.

3.2.2. Transformacje zmiennych objaśnianych

Z postaci pytań i możliwych odpowiedzi widać, że dane wymagają pewnych transformacji. W szczególności, na potrzeby budowania modeli regresji logistycznej, należy utworzyć binarne zmienne objaśniane, odpowiadające na pytania: czy dana osoba chodzi do kina oraz czy dana osoba czyta książki.

Zacznijmy od zmiennej opisującej zainteresowanie kinem. Zmienną *kino*, zawierającą odpowiedź na pytanie o częstotliwość wizyt w kinie, postanowiłyśmy zbinaryzować w następujący sposób:

$$czy.kino[i] = \begin{cases} 0 & \text{jeżeli } kino[i] \geq 5 \\ 1 & \text{jeżeli } kino[i] < 5. \end{cases}$$

Tak więc ustaliłyśmy, że jeśli ktoś chodzi do kina rzadziej niż raz w roku bądź nigdy, to „nie chodzi do kina", natomiast jeśli chodzi raz w roku bądź częściej, to „chodzi do kina".

Nazwy zmiennych objaśnianych zaczynają się od słówka „czy".

Odpowiedzi na pytanie o liczbę przeczytanych w ciągu ostatniego roku książek były bardzo zróżnicowane, wahały się od 0 do 300. Pojawiła się też jedna wartość odstająca od pozostałych wyników: 3000. Gdybyśmy w naszych analizach traktowały liczbę książek jako zmienną ilościową, z pewnością odrzuciłybyśmy tę obserwację (nie sądzimy, aby naprawdę ktoś był w stanie przeczytać 3000 książek w ciągu roku), jednak przy binaryzacji zmiennej przestaje ona być wartością odstającą – po prostu stwierdzamy, że ta osoba czyta książki i nie interesuje

Pytanie	**Możliwe odpowiedzi**
Jak często zdarza się Panu/i oglądać filmy w kinie?	1. Przynajmniej raz w miesiącu. 2. Raz na 2-3 miesiące. 3. 2-3 razy w roku. 4. Mniej więcej raz do roku. 5. Rzadziej niż raz do roku. 6. Nigdy.
Proszę wpisać, ile książek przeczytał/a Pan/i w ciągu ostatniego roku?	dowolna liczba całkowita nieujemna

Tablica 3.1: Pytania i odpowiedzi z sondażu dotyczące zmiennych objaśnianych

Pytanie	**Możliwe odpowiedzi**
Płeć:	1. Kobieta. 2. Mężczyzna.
Rok urodzenia:	1961-1995
Jakie ma Pan(i) wykształcenie?	1. Podstawowe (niepełne podstawowe, podstawowe). 2. Zasadnicze zawodowe. 3. Średnie (niepełne średnie, średnie zawodowe lub ogólnokształcące). 4. Wyższe (pomaturalne, niepełne wyższe, licencjackie, wyższe, doktorat, podyplomowe).
Jaka jest wielkość miejscowości, w której Pan(i) mieszka?	1. Miejscowość w gminie wiejskiej. 2. Miasto do 10 tys. mieszkańców. 3. Miasto 10.000-19.999 mieszkańców. 4. Miasto 20.000-49.999 mieszkańców. 5. Miasto 50.000-99.999 mieszkańców. 6. Miasto 100.000-199.999 mieszkańców. 7. Miasto 200.000-499.999 mieszkańców. 8. Miasto 500.000-999.999 mieszkańców. 9. Miasto powyżej 1.000.000 mieszkańców.
Ile osób, łącznie z Panem/Panią liczy Pana/i gospodarstwo domowe?	1. Mieszkam samodzielnie. 2. Dwie osoby. 3. Trzy osoby. 4. Cztery osoby. 5. Pięć osób. 6. Sześć osób. 7. Siedem osób. 8. Osiem i więcej osób.
Jaki jest Pana/i stan cywilny?	1. Stan wolny. 2. Żyję z partnerką / partnerem. 3. Żonaty / mężatka. 4. Rozwiedziony/a / w separacji / wdowiec / wdowa.

Tablica 3.2: Pytania i odpowiedzi z sondażu dotyczące zmiennych objaśniających

nas liczba tych książek. W toku analiz, przeprowadzając symulacje na różnych wariantach binaryzacji tej zmiennej, postanowiłyśmy przedstawić w pracy takie dwa warianty:

$$czy.ksiazki0[i] = \begin{cases} 0 & \text{jeżeli } liczbaksiazek[i] = 0 \\ 1 & \text{jeżeli } liczbaksiazek[i] > 0 \end{cases}$$

$$czy.ksiazki10[i] = \begin{cases} 0 & \text{jeżeli } liczbaksiazek[i] \leq 10 \\ 1 & \text{jeżeli } liczbaksiazek[i] > 10. \end{cases}$$

3.2.3. Transformacje zmiennych objaśniających

Płeć zdefiniowałyśmy następująco:

- 0: mężczyzna,
- 1: kobieta.

Oczywiście tak utworzonej zmiennej nie traktujemy jako ilościową, tylko jakościową, co uzyskałyśmy korzystając z funkcji `factor()`, opisanej w rozdziale 2. Zmienna ostatecznie nazywa się $fac.plec$ (nazwy zmiennych poddanych faktoryzacji zaczynają się od słówka „fac").

Kolejną zmienną jest wiek. Oczywiście obliczyłyśmy go z wzoru: $wiek[i] = 2011 - rokurodzenia[i]$ (badanie było przeprowadzone w 2011 roku). Zastanawiałyśmy się, czy potraktować wiek jako zmienną ilościową czy też sfaktoryzować ją, czyli zdefiniować przedziały wiekowe i z tych przedziałów utworzyć zmienną jakościową. Po wstępnych analizach zdecydowałyśmy się na to drugie podejście. Ogół respondentów podzieliłyśmy na trzy grupy wiekowe:

- 1: 16-25 lat,
- 2: 26-40 lat,
- 3: 41-50 lat.

Tak utworzoną zmienną jakościową nazwałyśmy $fac.wiek$. Zasięg pierwszej grupy został przez nas ustalony „sztywno" od samego początku – chciałyśmy odróżnić od reszty respondentów osoby w wieku 16-25 lat jako grupę „młodych", prawdopodobnie jeszcze uczących się osób. Po takiej operacji został nam ogół osób w wieku 26-50 lat, których postanowiłyśmy podzielić na dwie grupy, żeby długości przedziałów wiekowych były zbliżone. Ostateczną granicę pomiędzy drugim a trzecim przedziałem uzyskałyśmy rozpatrując różne jej warianty i badając istotność tak utworzonych zmiennych przy budowaniu naszych modeli. Okazało się, że przy badaniu istotności zmiennej $fac.wiek$ w budowanych przez nas modelach, najmniejsze p-wartości dla testu z hipotezą zerową mówiącą o nieistotności tej zmiennej są osiągane wtedy, gdy za wartość graniczną między poziomami 2 i 3 przyjmiemy 40 lat.

Uważamy, że pytanie o wykształcenie zostało zadane w dość niefortunny sposób. Podpowiedzi w nawiasach umieszczone przy kolejnych możliwych odpowiedziach (patrz tablica 3.2) sprawiają, że trudno jest jednoznacznie stwierdzić, jak autorzy pytań definiują wykształcenie średnie i wyższe. Włączenie do kategorii wykształcenia wyższego takich określeń jak „pomaturalne" i „niepełne wyższe" prawdopodobnie sprawiło, że osoby z wykształceniem średnim, w szczególności studenci, niesłusznie zawyżyły swoje wykształcenie. Wskazują na to: zaskakująco duży odsetek ludzi z wyższym wykształceniem oraz wiek osób deklarujących wyższe wykształcenie (często zbyt wczesny na uzyskanie tytułu licencjata). Być może był to celowy zabieg mający na celu włączenie do grupy ludzi z wyższym wykształceniem ogółu studentów.

Zastanawiałyśmy się nad scaleniem grup ludzi ze średnim i wyższym wykształceniem, aby pozbyć się tej niejednoznaczności, ale postanowiłyśmy zrezygnować z tego pomysłu i uznać, że ostatnią grupę stanowią ludzie z wykształceniem wyższym oraz studenci.

Tak więc faktoryzacja zmiennej *wyksztalcenie* wygląda tak:

- 1: podstawowe (odpowiedź nr 1),
- 2: zasadnicze zawodowe (odpowiedź nr 2),
- 3: średnie (odpowiedź nr 3),
- 4: wyższe i studenci (odpowiedź nr 4).

Tak utworzoną zmienną jakościową nazwałyśmy $fac.wyksztalcenie$.

W sondażu jest 9 różnych odpowiedzi na pytanie o wielkość miejscowości. Postanowiłyśmy scalić niektóre odpowiedzi i stworzyć dla tej zmiennej trzy poziomy:

- 1: miejscowość w gminie wiejskiej (odpowiedź nr 1),
- 2: małe miasto – do 99.999 mieszkańców (odpowiedzi nr 2-5),
- 3: duże miasto – co najmniej 100.000 mieszkańców (odpowiedzi nr 6-9).

Tak utworzoną zmienną nazwałyśmy $fac.miejscowosc$.

Podobnie postąpiłyśmy ze zmienną wskazującą na liczbę osób w gospodarstwie domowym – utworzyłyśmy 3 poziomy:

- 1: 1 osoba (odpowiedź nr 1),
- 2: mała rodzina – 2-4 osoby (odpowiedzi nr 2-4),
- 3: duża rodzina – co najmniej 5 osób (odpowiedzi nr 5-8).

Zmienną nazwałyśmy $fac.liczbaosob$.

Dla zmiennej opisującej stan cywilny utworzyłyśmy 4 poziomy tożsame z możliwymi odpowiedziami na pytanie w sondażu:

- 1: stan wolny (odpowiedź nr 1),
- 2: wolny związek (odpowiedź nr 2),
- 3: małżeństwo (odpowiedź nr 3),
- 4: rozwodnik/wdowiec (odpowiedź nr 4).

Zmienna nazywa się $fac.stan$.

Przy estymacji współczynników modelu regresji logistycznej program R traktuje sfaktoryzowaną zmienną w następujący sposób: pierwszy z poziomów ustala jako poziom bazowy, a dla pozostałych poziomów tworzy zmienne binarne. Na przykład dla zmiennej $fac.liczbaosob$ poziomem bazowym jest poziom 1 (1 osoba), a utworzonymi przez R zmiennymi binarnymi są:

$$fac.liczbaosob2[i] = \begin{cases} 1 & \text{jeżeli } fac.liczbaosob[i] = 2 \\ 0 & \text{jeżeli } fac.liczbaosob[i] \neq 2. \end{cases}$$

$$fac.liczbaosob3[i] = \begin{cases} 1 & \text{jeżeli } fac.liczbaosob[i] = 3 \\ 0 & \text{jeżeli } fac.liczbaosob[i] \neq 3. \end{cases}$$

Analogicznie program R ustala poziom bazowy i nowe zmienne binarne dla pozostałych zmiennych objaśniających.

3.3. Budowa modeli

3.3.1. Model badający zainteresowanie kinem

Najpierw zbadamy model uwzględniający wszystkie zmienne objaśniające. Oto oceny współczynników otrzymane za pomocą funkcji `glm()`:

```
Call:
glm(formula = czy.kino ~ fac.plec + fac.wiek + fac.wyksztalcenie +
    fac.miejscowosc + fac.liczbaosob + fac.stan, family = binomial())

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)         1.788026   0.592221   3.019 0.002535 **
fac.plec1          -0.199946   0.174910  -1.143 0.252980
fac.wiek2          -0.892130   0.250486  -3.562 0.000369 ***
fac.wiek3          -0.907820   0.299279  -3.033 0.002418 **
fac.wyksztalcenie2 -0.406413   0.436165  -0.932 0.351446
fac.wyksztalcenie3  0.461003   0.396388   1.163 0.244826
fac.wyksztalcenie4  1.544811   0.419291   3.684 0.000229 ***
fac.miejscowosc2    0.248074   0.188686   1.315 0.188595
fac.miejscowosc3    0.496878   0.206830   2.402 0.016290 *
fac.liczbaosob2    -0.337223   0.456886  -0.738 0.460459
fac.liczbaosob3    -0.489516   0.494227  -0.990 0.321945
fac.stan2          -0.005641   0.273051  -0.021 0.983516
fac.stan3           0.029906   0.240253   0.124 0.900939
fac.stan4          -0.714305   0.453438  -1.575 0.115185
---
```

Zmiennymi objaśniającymi o największym znaczeniu, czyli tymi, dla których została na poziomie istotności 5% odrzucona hipoteza o ich nieistotności, okazały się wiek (zarówno dla poziomu 2, jak i dla poziomu 3 p-wartość jest niska, wynosi odpowiednio 0,0369% i 0,2418%), wykształcenie (dla poziomu 4, czyli dla wykształcenia wyższego p-wartość to 0,0229%) i wielkość miejscowości (dla poziomu 3, czyli dla dużej miejscowości p-wartość wynosi 1,629%). Na tym etapie możemy ustalić, że te trzy zmienne znajdą się w ostatecznym modelu. Inne zmienne okazały się mniej istotne według tego kryterium.

Zobaczmy, jakie zmienne zostaną wykluczone za pomocą funkcji `step()`:

```
Call: glm(formula = czy.kino ~ fac.wiek + fac.wyksztalcenie + fac.miejscowosc,
    family = binomial())

Coefficients:
       (Intercept)           fac.wiek2           fac.wiek3  fac.wyksztalcenie2
            1.3186              -0.8619             -0.8966             -0.3825
fac.wyksztalcenie3  fac.wyksztalcenie4   fac.miejscowosc2    fac.miejscowosc3
            0.4728              1.5066              0.2600              0.4817

Degrees of Freedom: 1283 Total (i.e. Null);  1276 Residual
Null Deviance:      1141
Residual Deviance: 1033         AIC: 1049
```

Przy użyciu funkcji `step()` za optymalny zbiór zmiennych objaśniających według kryterium Akaike'a uznany został zbiór złożony ze zmiennych: *fac.wiek*, *fac.wyksztalcenie* i *fac.miejscowosc*, natomiast zmienne *fac.plec*, *fac.liczbaosob* i *fac.stan* zostały wykluczone.

Dla każdej z trzech wykluczonych zmiennych zobaczmy, jak będą wyglądały wyniki regresji przeprowadzonej z uwzględnieniem tylko tej jednej ze zmiennych objaśniających.

Wyniki regresji z uwzględnieniem zmiennej *fac.plec*:

```
Call:
glm(formula = czy.kino ~ fac.plec, family = binomial())

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.4749     0.1029  14.330   <2e-16 ***
fac.plec1     0.3346     0.1520   2.201   0.0277 *
---
```

Wyniki regresji z uwzględnieniem zmiennej *fac.liczbaosob*:

```
Call:
glm(formula = czy.kino ~ fac.liczbaosob, family = binomial())

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)       2.1316     0.3997   5.333 9.69e-08 ***
fac.liczbaosob2  -0.4894     0.4089  -1.197    0.231
fac.liczbaosob3  -0.6331     0.4363  -1.451    0.147
---
```

Wyniki regresji z uwzględnieniem zmiennej *fac.stan*:

```
Call:
glm(formula = czy.kino ~ fac.stan, family = binomial())

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.9841     0.1438  13.797  < 2e-16 ***
fac.stan2    -0.1924     0.2415  -0.797  0.42568
fac.stan3    -0.5716     0.1780  -3.211  0.00132 **
fac.stan4    -1.0678     0.4009  -2.664  0.00772 **
---
```

W tych modelach regresji, gdzie zbiór zmiennych objaśniających został ograniczony do tylko jednej zmiennej, hipoteza o nieistotności została odrzucona tylko w przypadku zmiennych *fac.stan* (dla poziomu 3 p-wartość wynosi 0,132 %, a dla poziomu 4 p-wartość wynosi 0,772%) oraz *fac.plec* (p-wartość dla poziomu 1 to 2,77 %). Dla zmiennej *fac.liczbaosob* nawet w takim modelu regresji hipoteza o nieistotności nie została odrzucona, dlatego w tym momencie możemy podjąć decyzję o nieuwzględnianiu jej w ostatecznym modelu.

Zbadajmy model nieuwzględniający tej zmiennej:

```
Call:
glm(formula = czy.kino ~ fac.plec + fac.wiek + fac.wyksztalcenie +
    fac.miejscowosc + fac.stan, family = binomial())

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)         1.38597    0.38159   3.632 0.000281 ***
fac.plec1          -0.21389    0.17434  -1.227 0.219882
fac.wiek2          -0.84137    0.24484  -3.436 0.000590 ***
fac.wiek3          -0.85920    0.29419  -2.921 0.003494 **
fac.wyksztalcenie2 -0.37168    0.43374  -0.857 0.391489
fac.wyksztalcenie3  0.48751    0.39459   1.235 0.216650
fac.wyksztalcenie4  1.60278    0.41519   3.860 0.000113 ***
fac.miejscowosc2    0.26538    0.18790   1.412 0.157847
fac.miejscowosc3    0.52349    0.20492   2.555 0.010632 *
fac.stan2          -0.04259    0.26742  -0.159 0.873464
fac.stan3          -0.02130    0.23077  -0.092 0.926450
fac.stan4          -0.66252    0.44897  -1.476 0.140042
---
```

Teza o nieistotności zmiennych *fac.plec* i *fac.stan* nadal nie została odrzucona, mimo że w modelach, w których te zmienne stanowią jedyne zmienne objaśniające teza o ich nieistotności została odrzucona. Być może te zmienne są silnie skorelowane z tymi zmiennymi objaśniającymi, dla których teza o nieistotności została odrzucona. Zbadajmy to sposobem zaproponowanym w [6] s. 290, czyli dodając do zbioru zmiennych objaśniających interakcje tych zmiennych, co do których podejrzewamy, że mogą być skorelowane.

Pary zmiennych, które mogą być ze sobą powiązane to:

- wiek i stan cywilny,
- wiek i wykształcenie,
- płeć i wykształcenie,
- wielkość miejscowości i stan cywilny.

Uwzględnimy interakcje tych zmiennych w kolejnym modelu:

```
Call:
glm(formula = czy.kino ~ fac.plec + fac.wiek + fac.wyksztalcenie +
fac.miejscowosc + fac.stan + fac.stan:fac.wiek + fac.plec:fac.wyksztalcenie +
fac.stan:fac.miejscowosc + fac.wiek:fac.wyksztalcenie, family = binomial())

Coefficients:
                               Estimate Std. Error z value Pr(>|z|)
(Intercept)                    1.988947   0.630812   3.153  0.00162 **
fac.plec1                     -0.885699   0.980100  -0.904  0.36616
fac.wiek2                     -2.918514   1.086817  -2.685  0.00724 **
fac.wiek3                     10.523386 535.412238   0.020  0.98432
fac.wyksztalcenie2            -1.490392   0.925396  -1.611  0.10728
fac.wyksztalcenie3            -0.391345   0.666057  -0.588  0.55683
fac.wyksztalcenie4             0.705632   0.764922   0.922  0.35627
fac.miejscowosc2               0.839476   0.374807   2.240  0.02511 *
fac.miejscowosc3               0.865270   0.394886   2.191  0.02844 *
fac.stan2                      0.911171   0.626192   1.455  0.14564
fac.stan3                     -0.738590   0.535589  -1.379  0.16789
fac.stan4                     11.214024 535.412598   0.021  0.98329
fac.wiek2:fac.stan2           -0.210996   0.639102  -0.330  0.74129
fac.wiek3:fac.stan2            0.142563   0.939016   0.152  0.87933
fac.wiek2:fac.stan3            1.169547   0.582597   2.007  0.04470 *
fac.wiek3:fac.stan3            1.315301   0.808756   1.626  0.10388
fac.wiek2:fac.stan4           -9.602568 535.412981  -0.018  0.98569
fac.wiek3:fac.stan4          -12.253480 535.412056  -0.023  0.98174
fac.plec1:fac.wyksztalcenie2   0.499063   1.066287   0.468  0.63976
fac.plec1:fac.wyksztalcenie3   0.778848   1.009280   0.772  0.44030
fac.plec1:fac.wyksztalcenie4   0.935366   1.035280   0.903  0.36627
fac.miejscowosc2:fac.stan2    -0.732198   0.686669  -1.066  0.28629
fac.miejscowosc3:fac.stan2    -1.595943   0.654496  -2.438  0.01475 *
fac.miejscowosc2:fac.stan3    -0.842332   0.454975  -1.851  0.06411 .
fac.miejscowosc3:fac.stan3    -0.169366   0.495136  -0.342  0.73231
fac.miejscowosc2:fac.stan4    -1.885135   1.454259  -1.296  0.19488
fac.miejscowosc3:fac.stan4    -0.003978   1.266635  -0.003  0.99749
fac.wiek2:fac.wyksztalcenie2   2.204033   1.276771   1.726  0.08430 .
fac.wiek3:fac.wyksztalcenie2 -11.073941 535.412472  -0.021  0.98350
fac.wiek2:fac.wyksztalcenie3   1.998103   1.097778   1.820  0.06874 .
fac.wiek3:fac.wyksztalcenie3 -11.674383 535.411975  -0.022  0.98260
fac.wiek2:fac.wyksztalcenie4   1.925500   1.140906   1.688  0.09147 .
fac.wiek3:fac.wyksztalcenie4 -11.783450 535.412106  -0.022  0.98244
---
```

Hipoteza o nieistotności poszczególnych interakcji została na poziomie istotności 5% odrzucona tylko w przypadku interakcji między zmiennymi *fac.wiek* i *fac.stan* (jest istotna korelacja między byciem w przedziale wiekowym 26-40 lat a życiem w małżeństwie) oraz między zmiennymi *fac.miejscowosc* i *fac.stan* (korelacja między mieszkaniem w dużym mieście a byciem w związku nieformalnym).

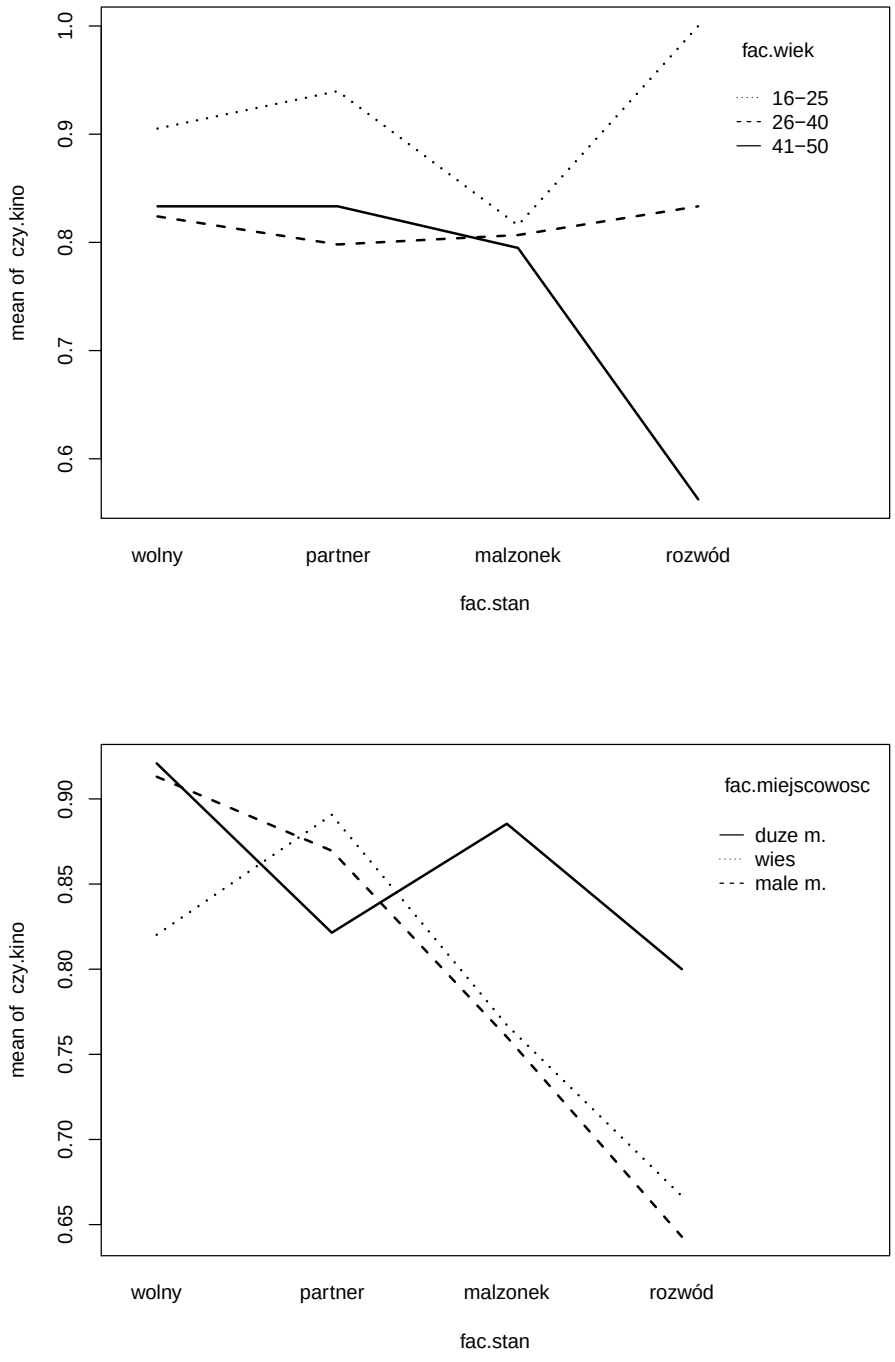
Przy użyciu funkcji `step()` za optymalny zbiór zmiennych objaśniających według kryterium Akaike'a uznany został zbiór złożony ze zmiennych: *fac.wiek*, *fac.wyksztalcenie*, *fac.miejscowosc*, *fac.stan* oraz interakcji: *fac.wiek* i *fac.stan* oraz *fac.miejscowosc* i *fac.stan*, natomiast zmienna *fac.plec* oraz interakcje zmiennych *fac.wiek* i *fac.wyksztalcenie* oraz zmiennych *fac.plec* i *fac.wyksztalcenie* zostały wykluczone. Wykluczmy te dwie interakcje oraz zmienną opisującą płeć z ostatecznego modelu. Zbadajmy tak zmodyfikowany model:

```
Call:
glm(formula = czy.kino ~ fac.wiek + fac.wyksztalcenie + fac.miejscowosc +
    fac.stan + fac.stan:fac.wiek + fac.stan:fac.miejscowosc,
    family = binomial())

Coefficients:
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                  1.154024   0.399105   2.892 0.003834 **
fac.wiek2                   -1.035749   0.331937  -3.120 0.001807 **
fac.wiek3                   -1.198687   0.633562  -1.892 0.058494 .
fac.wyksztalcenie2          -0.352192   0.442450  -0.796 0.426030
fac.wyksztalcenie3           0.503067   0.402525   1.250 0.211380
fac.wyksztalcenie4           1.629164   0.420307   3.876 0.000106 ***
fac.miejscowosc2             0.854600   0.373395   2.289 0.022095 *
fac.miejscowosc3             0.844374   0.390940   2.160 0.030784 *
fac.stan2                    0.893805   0.622351   1.436 0.150953
fac.stan3                   -0.844628   0.515895  -1.637 0.101587
fac.stan4                   11.069782 535.412575   0.021 0.983505
fac.wiek2:fac.stan2         -0.218642   0.629668  -0.347 0.728416
fac.wiek3:fac.stan2          0.269708   0.940126   0.287 0.774200
fac.wiek2:fac.stan3          1.205899   0.569344   2.118 0.034171 *
fac.wiek3:fac.stan3          1.427331   0.802261   1.779 0.075218 .
fac.wiek2:fac.stan4         -9.445954 535.412954  -0.018 0.985924
fac.wiek3:fac.stan4        -12.180255 535.412052  -0.023 0.981850
fac.miejscowosc2:fac.stan2  -0.810401   0.678815  -1.194 0.232538
fac.miejscowosc3:fac.stan2  -1.640744   0.648123  -2.532 0.011357 *
fac.miejscowosc2:fac.stan3  -0.820321   0.451460  -1.817 0.069211 .
fac.miejscowosc3:fac.stan3  -0.152073   0.490675  -0.310 0.756618
fac.miejscowosc2:fac.stan4  -1.946363   1.444171  -1.348 0.177743
fac.miejscowosc3:fac.stan4  -0.005181   1.264583  -0.004 0.996731
---
    Null deviance: 1140.8  on 1283  degrees of freedom
Residual deviance: 1003.7  on 1261  degrees of freedom
AIC: 1049.7
```

Graficzne porównanie średnich wartości zmiennej *czy.kino* w zależności od kombinacji poziomów zmiennych objaśniających, których interakcje zostały uwzględnione w powyższym modelu, zostało przedstawione na wykresach typu `interaction.plot()` na rysunku 3.1. W ogólności, różne nachylenia poszczególnych krzywych na wykresach oraz przecinanie się tych krzywych świadczy o istnieniu istotnych interakcji ([1], s. 214-216). Przy interakcji zmiennych *fac.wiek* i *fac.stan* krzywe mają podobne nachylenia dla wszystkich wartości zmiennej *fac.stan* oprócz ostatniej – wdowców i rozwodników. Być może wynika to z niskiej liczności tej grupy; w szczególności była tylko jedna osoba młoda (16-25 lat) w tej grupie. Krzywe na tym wykresie przecinają się, ale pod małym kątem. Sytuacja na wykresie interakcji

$fac.miejscowosc$ i $fac.stan$ wygląda ciekawiej – krzywe mają różne nachylenia i przecinają się w licznych miejscach, zwłaszcza w okolicach poziomu „partner" dla zmiennej $fac.stan$ i poziomu „duże miasto" dla zmiennej $fac.miejscowosc$.



Rysunek 3.1: Wykresy `interaction.plot()` ukazujące interakcje między zmiennymi $fac.wiek$ i $fac.stan$ oraz $fac.miejscowosc$ i $fac.stan$ w modelu badającym zmienną $czy.kino$

Powstały w ten sposób model okazuje się optymalny według kryterium Akaike'a (funkcja `step()` nie odrzuca żadnej ze zmiennych objaśniających). Dlatego też nie będziemy już modyfikować tego modelu. Jednak dla naszych analiz, w których będziemy chciały porównać wpływ zmiennych objaśniających na różne zmienne objaśniane, taki model, choć optymalny, okazuje się zbyt skomplikowany. Dlatego też nasz wybór ostatecznego modelu opisującego zainteresowanie kinem jako nośnikiem kultury pada na model prostszy, nieuwzględniający odpowiednich interakcji zmiennych objaśniających ani zmiennej *fac.stan*, czyli na model proponowany przez pierwszą użytą przez nas funkcję `step()`, w którym za optymalny zbiór zmiennych objaśniających według kryterium Akaike'a uznany został zbiór złożony ze zmiennych: *fac.wiek*, *fac.wyksztalcenie* i *fac.miejscowosc*. Jest to w pewnym sensie kompromis pomiędzy dopasowaniem do danych a liczbą uwzględnionych zmiennych objaśniających.

```
Call:
glm(formula = czy.kino ~ fac.wiek + fac.wyksztalcenie + fac.miejscowosc,
    family = binomial())

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)          1.3186     0.3737   3.529 0.000417 ***
fac.wiek2           -0.8619     0.2083  -4.138  3.5e-05 ***
fac.wiek3           -0.8966     0.2507  -3.577 0.000347 ***
fac.wyksztalcenie2  -0.3825     0.4317  -0.886 0.375519
fac.wyksztalcenie3   0.4728     0.3932   1.202 0.229244
fac.wyksztalcenie4   1.5066     0.4087   3.687 0.000227 ***
fac.miejscowosc2     0.2600     0.1868   1.392 0.164051
fac.miejscowosc3     0.4817     0.2031   2.372 0.017705 *
---
    Null deviance: 1140.8  on 1283  degrees of freedom
Residual deviance: 1033.2  on 1276  degrees of freedom
AIC: 1049.2
```

Dewiancja jest większa niż w przypadku poprzedniego modelu. Różnica ta jednak nie jest znacząca, natomiast model jest dużo prostszy – ma o połowę mniej zmiennych objaśniających. Właśnie z tych powodów postanowiłyśmy ostatecznie wybrać model oparty na trzech zmiennych objaśniających.

Rysunek 3.2 przedstawia wykresy ukazujące, jak wybrane ostatecznie zmienne objaśniające wpływają na zmienną objaśnianą. Interpretacja wykresów znajduje się w rozdziale 3.5.

3.3.2. Model badający zainteresowanie książkami – zmienna *czy.ksiazki*0

Analogicznie jak dla kina, zbudujemy model dla książek. Na początku weźmiemy na warsztat zmienną *czy.ksiazki*0, czyli taką, która za osobę czytającą książki uznaje tę, która w ciągu roku przeczytała co najmniej jedną książkę. Zaczniemy od modelu pełnego, czyli zawierającego wszystkie zmienne objaśniające:

```
Call:
glm(formula = czy.ksiazki0 ~ fac.plec + fac.wiek + fac.wyksztalcenie +
   fac.miejscowosc + fac.liczbaosob + fac.stan, family = binomial())

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        1.90789    0.71111   2.683   0.0073 **
fac.plec1          0.94816    0.23090   4.106 4.02e-05 ***
fac.wiek2         -0.29148    0.29264  -0.996   0.3192
fac.wiek3          0.19793    0.36885   0.537   0.5915
fac.wyksztalcenie2 -1.10234    0.59447  -1.854   0.0637 .
```

```
fac.wyksztalcenie3 -0.26011    0.55799  -0.466   0.6411
fac.wyksztalcenie4  0.66786    0.58716   1.137   0.2554
fac.miejscowosc2   -0.04465    0.22789  -0.196   0.8447
fac.miejscowosc3    0.14123    0.26004   0.543   0.5871
fac.liczbaosob2     0.36274    0.48453   0.749   0.4541
fac.liczbaosob3     0.39073    0.53790   0.726   0.4676
fac.stan2          -0.50097    0.31996  -1.566   0.1174
fac.stan3          -0.37788    0.29583  -1.277   0.2015
fac.stan4          -0.54145    0.58731  -0.922   0.3566
---
```

Widzimy, że najmniejszą p-wartość ma płeć i w tym momencie tylko tej zmiennej nie możemy odrzucić.

Sprawdzamy, że funkcja `step()` usuwa z modelu zmienne *fac.miejscowosc*, *fac.liczbaosob* i *fac.stan*:

```
Call:  glm(formula = czy.ksiazki0 ~ fac.plec + fac.wiek + fac.wyksztalcenie,
   family = binomial())

Coefficients:
      (Intercept)           fac.plec1           fac.wiek2           fac.wiek3
          2.21314             0.93329             -0.51039             -0.03878
fac.wyksztalcenie2  fac.wyksztalcenie3  fac.wyksztalcenie4
         -1.17676            -0.30578             0.61582
```

Zobaczmy jednak jak wyglądają modele, w których te zmienne są jedynymi zmiennymi objaśniającymi.

Miejscowość:

```
Call:
glm(formula = czy.ksiazki0 ~ fac.miejscowosc, family = binomial())

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)        2.1401     0.1559  13.729   <2e-16 ***
fac.miejscowosc2  -0.2132     0.2143  -0.995   0.3197
fac.miejscowosc3   0.4149     0.2430   1.707   0.0878 .
---
```

Liczba osób w gospodarstwie:

```
Call:
glm(formula = czy.ksiazki0 ~ fac.liczbaosob, family = binomial())

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)     2.13163    0.39974   5.333 9.69e-08 ***
fac.liczbaosob2  0.04246    0.41317   0.103    0.918
fac.liczbaosob3  0.11212    0.46092   0.243    0.808
---
```

Stan cywilny:

```
Call:
glm(formula = czy.ksiazki0 ~ fac.stan, family = binomial())

Coefficients:
           Estimate Std. Error z value Pr(>|z|)
(Intercept)   2.4849     0.1759  14.124   <2e-16 ***
fac.stan2    -0.3029     0.2855  -1.061   0.2887
fac.stan3    -0.4720     0.2182  -2.163   0.0306 *
fac.stan4    -0.6931     0.5141  -1.348   0.1776
---
```

Możemy odrzucić zmienne *fac.liczbaosob* i *fac.miejscowosc*, gdyż na poziomie istotności 5% hipoteza zerowa o ich nieistotności nie jest odrzucona.

Dalej zgodnie z wcześniejszym postępowaniem dorzucamy do modelu interakcje zmiennych, które podejrzewamy o silne skorelowanie. W tym wypadku mogą to być jedynie wiek i stan cywilny.

```
Call:
glm(formula = czy.ksiazki0 ~ fac.plec + fac.wiek + fac.wyksztalcenie +
   fac.stan + fac.wiek:fac.stan, family = binomial())

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)           2.4182     0.5440   4.446 8.76e-06 ***
fac.plec1             0.9721     0.2304   4.220 2.45e-05 ***
fac.wiek2            -0.8814     0.3835  -2.298   0.0215 *
fac.wiek3             0.3630     1.0779   0.337   0.7363
fac.wyksztalcenie2   -1.0754     0.5978  -1.799   0.0720 .
fac.wyksztalcenie3   -0.2196     0.5615  -0.391   0.6957
fac.wyksztalcenie4    0.7499     0.5904   1.270   0.2040
fac.stan2            -0.9768     0.4613  -2.118   0.0342 *
fac.stan3            -1.1780     0.6123  -1.924   0.0544 .
fac.stan4            10.3954   535.4113   0.019   0.9845
fac.wiek2:fac.stan2   1.2949     0.6503   1.991   0.0465 *
fac.wiek3:fac.stan2  -0.4629     1.2553  -0.369   0.7123
fac.wiek2:fac.stan3   1.2212     0.6918   1.765   0.0775 .
fac.wiek3:fac.stan3   0.5655     1.2430   0.455   0.6491
fac.wiek2:fac.stan4 -10.7886   535.4117  -0.020   0.9839
fac.wiek3:fac.stan4 -11.1452   535.4133  -0.021   0.9834
---
```

W tak skonstruowanym modelu wszystkie zmienne oprócz wykształcenia wydają się znacząco wpływać na nasz model. Jednak funkcja `step()` zostawia nam te same zmienne, co w modelu pierwotnym, czyli *fac.plec*, *fac.wiek*, *fac.wyksztalcenie*, a resztę odrzuca:

```
Call:  glm(formula = czy.ksiazki0 ~ fac.plec + fac.wiek + fac.wyksztalcenie,
   family = binomial())

Coefficients:
      (Intercept)           fac.plec1           fac.wiek2           fac.wiek3
          2.21314             0.93329            -0.51039            -0.03878
fac.wyksztalcenie2  fac.wyksztalcenie3  fac.wyksztalcenie4
         -1.17676            -0.30578             0.61582
```

Dlatego w ostatecznym modelu uwzględnimy tylko te trzy zmienne.

```
Call:
glm(formula = czy.ksiazki0 ~ fac.plec + fac.wiek + fac.wyksztalcenie,
   family = binomial())

Deviance Residuals:
   Min       1Q   Median       3Q      Max
-2.7514   0.2185   0.3388   0.5262   0.9637

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)         2.21314    0.52956   4.179 2.93e-05 ***
fac.plec1           0.93329    0.22595   4.131 3.62e-05 ***
fac.wiek2          -0.51039    0.24064  -2.121   0.0339 *
fac.wiek3          -0.03878    0.31324  -0.124   0.9015
fac.wyksztalcenie2 -1.17676    0.59088  -1.992   0.0464 *
fac.wyksztalcenie3 -0.30578    0.55517  -0.551   0.5818
```

```
fac.wyksztalcenie4   0.61582     0.58332    1.056    0.2911
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

   Null deviance: 841.82  on 1283  degrees of freedom
Residual deviance: 753.86  on 1277  degrees of freedom
AIC: 767.86

Number of Fisher Scoring iterations: 6
```

Na wykresach (rysunek 3.3) zaprezentowano, jak na cały model wpływają poszczególne zmienne.

3.3.3. Model badający zainteresowanie książkami – zmienna *czy.ksiazki*10

Zobaczmy teraz jak zmieni się nasz model, jeśli do grupy osób nieczytających książek dołączymy tych, którzy przeczytali 10 lub mniej książek w ciągu roku. Znowu zaczynamy od pełnego modelu:

```
Call:
glm(formula = czy.ksiazki10 ~ fac.plec + fac.wiek + fac.wyksztalcenie +
   fac.miejscowosc + fac.liczbaosob + fac.stan, family = binomial())

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)       -0.01779    0.41744  -0.043   0.9660
fac.plec1          0.50780    0.12786   3.972 7.14e-05 ***
fac.wiek2         -0.17994    0.16164  -1.113   0.2656
fac.wiek3          0.37425    0.20915   1.789   0.0736 .
fac.wyksztalcenie2 -1.00515    0.38970  -2.579   0.0099 **
fac.wyksztalcenie3 -0.22415    0.31765  -0.706   0.4804
fac.wyksztalcenie4  0.34710    0.32218   1.077   0.2813
fac.miejscowosc2   -0.13773    0.14541  -0.947   0.3436
fac.miejscowosc3   -0.15437    0.14644  -1.054   0.2918
fac.liczbaosob2    -0.03156    0.28739  -0.110   0.9125
fac.liczbaosob3    -0.00700    0.31911  -0.022   0.9825
fac.stan2          -0.31523    0.18153  -1.737   0.0825 .
fac.stan3          -0.51537    0.17116  -3.011   0.0026 **
fac.stan4          -0.58198    0.39453  -1.475   0.1402
---
```

Tym razem zmienne płeć, wykształcenie i stan cywilny wpływają istotnie na nasz model.

Funkcja `step()` dodaje do tego okrojonego modelu zmienną wiek:

```
Call:  glm(formula = czy.ksiazki10 ~ fac.plec + fac.wiek + fac.wyksztalcenie +
   fac.stan, family = binomial())

Coefficients:
      (Intercept)           fac.plec1           fac.wiek2           fac.wiek3
          -0.1040              0.5013              -0.1954              0.3514
fac.wyksztalcenie2  fac.wyksztalcenie3  fac.wyksztalcenie4           fac.stan2
          -1.0060             -0.2398              0.3255              -0.3274
        fac.stan3           fac.stan4
          -0.5183             -0.6004
```

Postępując jak poprzednio, sprawdźmy jak wykluczone zmienne wpływają na zainteresowanie czytelnictwem w modelach, w których są jedynymi zmiennymi objaśniającymi.

Miejscowość:

```
Call:
glm(formula = czy.ksiazki10 ~ fac.miejscowosc, family = binomial())

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      -0.13292    0.09588  -1.386    0.166
fac.miejscowosc2 -0.22085    0.13816  -1.598    0.110
fac.miejscowosc3  0.03050    0.13609   0.224    0.823
```

Liczba osób w gospodarstwie:

```
Call:
glm(formula = czy.ksiazki10 ~ fac.liczbaosob, family = binomial())

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)       0.1214     0.2466   0.492    0.623
fac.liczbaosob2  -0.3486     0.2547  -1.368    0.171
fac.liczbaosob3  -0.2586     0.2814  -0.919    0.358
```

Widzimy więc, że obydwie zmienne możemy odrzucić, skoro okazały się nieistotne w modelach regresji, w których pełniły rolę jedynych zmiennych objaśniających.

Nasz ostateczny model jest złożony ze zmiennych: *fac.plec*, *fac.wiek*, *fac.wyksztalcenie*, *fac.stan*.

```
Call:
glm(formula = czy.ksiazki10 ~ fac.plec + fac.wiek + fac.wyksztalcenie +
    fac.stan, family = binomial())

Deviance Residuals:
   Min       1Q   Median       3Q      Max
-1.6541  -1.0449  -0.7014   1.1574   1.9867

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)         -0.1040     0.3018  -0.345  0.73025
fac.plec1            0.5013     0.1264   3.966 7.32e-05 ***
fac.wiek2           -0.1954     0.1567  -1.247  0.21224
fac.wiek3            0.3514     0.2036   1.726  0.08441 .
fac.wyksztalcenie2  -1.0060     0.3885  -2.589  0.00961 **
fac.wyksztalcenie3  -0.2398     0.3168  -0.757  0.44922
fac.wyksztalcenie4   0.3255     0.3199   1.018  0.30884
fac.stan2           -0.3274     0.1778  -1.842  0.06553 .
fac.stan3           -0.5183     0.1630  -3.180  0.00147 **
fac.stan4           -0.6004     0.3925  -1.530  0.12608
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1768  on 1283  degrees of freedom
Residual deviance: 1662  on 1274  degrees of freedom
AIC: 1682

Number of Fisher Scoring iterations: 4
```

Wpływ poszczególnych zmiennych na *czy.ksiazki*10 jest zaprezentowany na rysunku 3.4.

3.4. Diagnostyka modeli

W tym rozdziale przetestujemy modele, które ostatecznie otrzymałyśmy, pod względem dopasowania do naszych danych. Dla każdego modelu sprawdzimy 4 kryteria:

1. różnica dewiancji,
2. statystyka R_P^2,
3. R_N^2 Nagelkerkego,
4. statystyka Pearsona.

Różnica dewiancji służy do oceny poprawności modelu, podczas gdy R_P^2, R_N^2 i statystyka Pearsona mówią o dopasowaniu modelu do danych.

Zaczniemy od modelu dla zmiennej *czy.kino*. Oto otrzymany przez nas wynik:

```
Call:
glm(formula = czy.kino ~ fac.wiek + fac.wyksztalcenie + fac.miejscowosc,
family = binomial())
Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)          1.3186     0.3737   3.529 0.000417 ***
fac.wiek2           -0.8619     0.2083  -4.138  3.5e-05 ***
fac.wiek3           -0.8966     0.2507  -3.577 0.000347 ***
fac.wyksztalcenie2  -0.3825     0.4317  -0.886 0.375519
fac.wyksztalcenie3   0.4728     0.3932   1.202 0.229244
fac.wyksztalcenie4   1.5066     0.4087   3.687 0.000227 ***
fac.miejscowosc2     0.2600     0.1868   1.392 0.164051
fac.miejscowosc3     0.4817     0.2031   2.372 0.017705 *
---
   Null deviance: 1140.8  on 1283  degrees of freedom
Residual deviance: 1033.2  on 1276  degrees of freedom
AIC: 1049.2
```

Różnica dewiancji zerowej i resztowej wynosi: $G^2_{H_0} - G^2_{H_1} = 1140.8 - 1033.2 = 107.6$, gdzie według H_0 wszystkie współczynniki przy zmiennych objaśniających są równe 0, natomiast według H_1 przy zmiennych objaśniających stoją współczynniki wyliczone przez funkcję `glm()`. Tak zadana statystyka ma asymptotyczny rozkład χ^2 z $1283 - 1276 = 7$ stopniami swobody. Obliczamy: $\mathbb{P}(G^2_{H_0} - G^2_{H_1} > 107.6) \approx 0$. Czyli odrzucamy H_0, więc według tego kryterium nasz model jest istotnie lepszy niż model pusty.

Teraz sprawdźmy wartość R_P^2. Wiedząc, że $R_P^2 = 1 - \frac{G^2_{H_1}}{G^2_{H_0}}$ dla hipotez jak wyżej, obliczamy: $R_P^2 = 1 - \frac{1033.2}{1140.8} = 0.094$. Zobaczmy jeszcze ile wynosi R_N^2 Nagelkerkego. Obliczamy w pakiecie R za pomocą odpowiedniej funkcji, że $R_N^2 = 0.137$. Nie są to duże liczby, jednak ciężko jest uzyskać wysoką wartość R_P^2 w dziedzinach społecznych, gdzie na odpowiedzi wpływa ogromna ilość czynników, a znamy tylko niewielką część z nich, Tak więc niskie wartości R_P^2 i R_N^2 nie są podstawą do stwierdzenia, że nasz model jest niedostatecznie dopasowany do danych.

Sprawdźmy jeszcze jak wygląda statystyka Pearsona X^2. Według H_0 zbiór wyestymowanych współczynników jest wystarczający, natomiast według H_1 powinniśmy użyć modelu wysyconego. X^2 policzymy korzystając z funkcji pakietu R opisanej w rozdziale 2. Wynosi ona 1278,48. Asymptotycznie ta statystyka zbiega do rozkładu χ^2 z 1276 stopniami swobody, stąd $\mathbb{P}(X^2 > 1278.48) = 0.475$. Oznacza to, że nie możemy odrzucić hipotezy zerowej o poprawności modelu, więc możemy przyjąć, że nasz model jest dobrze dopasowany do naszych danych.

Przejdziemy teraz do modelu dla zmiennej *czy.ksiazki*0:

```
Call:
glm(formula = czy.ksiazki0 ~ fac.plec + fac.wiek + fac.wyksztalcenie,
   family = binomial())

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)         2.21314    0.52956   4.179 2.93e-05 ***
fac.plec1           0.93329    0.22595   4.131 3.62e-05 ***
fac.wiek2          -0.51039    0.24064  -2.121   0.0339 *
fac.wiek3          -0.03878    0.31324  -0.124   0.9015
fac.wyksztalcenie2 -1.17676    0.59088  -1.992   0.0464 *
fac.wyksztalcenie3 -0.30578    0.55517  -0.551   0.5818
fac.wyksztalcenie4  0.61582    0.58332   1.056   0.2911
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

   Null deviance: 841.82  on 1283  degrees of freedom
Residual deviance: 753.86  on 1277  degrees of freedom
AIC: 767.86
```

Obliczamy różnicę dewiancji: $G^2_{H_0} - G^2_{H_1} = 841.82 - 753.86 = 87.96$. Liczba stopni swobody rozkładu χ^2 wynosi $1283 - 1277 = 6$. Zatem $\mathbb{P}(G^2_{H_0} - G^2_{H_1} > 87.96) \approx 0$, czyli model jest poprawny.

Dalej sprawdzamy, że $R^2_P = 1 - \frac{753.86}{841.82} = 0.104$, a $R^2_N = 0.138$. Znowu 0.104 i 0,138% nie są dużymi liczbami, jednak, biorąc pod uwagę charakter naszych danych, dopuszczalnymi.

Obliczamy w R statystykę Pearsona X^2, wynosi ona 1270,36 dla rozkładu o 1277 stopniach swobody. Podobnie jak poprzednio znajdujemy $\mathbb{P}(X^2 > 1270.36) = 0.547$ i stwierdzamy, że nie możemy odrzucić hipotezy o poprawności naszego modelu, czyli według tego kryterium zmienne wraz z ocenami współczynników są dostatecznie dopasowane do danych.

I ostatni model, czyli dla zmiennej *czy.ksiazki*10:

```
Call:
glm(formula = czy.ksiazki10 ~ fac.plec + fac.wiek + fac.wyksztalcenie +
   fac.stan, family = binomial())

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)         -0.1040     0.3018  -0.345  0.73025
fac.plec1            0.5013     0.1264   3.966 7.32e-05 ***
fac.wiek2           -0.1954     0.1567  -1.247  0.21224
fac.wiek3            0.3514     0.2036   1.726  0.08441 .
fac.wyksztalcenie2  -1.0060     0.3885  -2.589  0.00961 **
fac.wyksztalcenie3  -0.2398     0.3168  -0.757  0.44922
fac.wyksztalcenie4   0.3255     0.3199   1.018  0.30884
fac.stan2           -0.3274     0.1778  -1.842  0.06553 .
fac.stan3           -0.5183     0.1630  -3.180  0.00147 **
fac.stan4           -0.6004     0.3925  -1.530  0.12608
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

   Null deviance: 1768  on 1283  degrees of freedom
Residual deviance: 1662  on 1274  degrees of freedom
AIC: 1682
```

Różnica dewiancji wynosi $G^2_{H_0} - G^2_{H_1} = 1768 - 1662 = 106$ przy $1283 - 1274 = 9$ stopniach swobody. Stąd $\mathbb{P}(G^2_{H_0} - G^2_{H_1} > 106) \approx 0$, więc według tego kryterium nasz model jest poprawny.

Liczymy $R^2_P = 1 - \frac{1662}{1768} = 0.06$, jednak zobaczmy jeszcze, jak wygląda R^2_N Nagelkerkego. Postępując jak poprzednio mamy, że $R^2_N = 0.106$.

Na koniec policzymy statystykę Pearsona X^2, wynosi ona 1285,19 przy 1274 stopniach swobody. $\mathbb{P}(X^2 > 1285.19) = 0.407$, czyli hipoteza zerowa, mówiąca, że model jest poprawny, nie może zostać odrzucona, co mieliśmy nadzieję uzyskać.

3.5. Interpretacja modeli

Otrzymane za pomocą funkcji `glm()` oceny współczynników w modelach można łatwo zinterpretować po przekształceniu ich funkcją $exp(x)$. Po wykonaniu takiej operacji reprezentują one wzrost lub spadek szansy w odniesieniu do poziomu bazowego.

Zaczniemy od interpretacji modelu opisującego zmienne różnicujące zainteresowanie kinem. Oto otrzymane oceny współczynników wyrażone w terminach szansy, uzyskane za pomocą funkcji `exp(coef())`:

```
       (Intercept)          fac.wiek2          fac.wiek3
         3.7380972          0.4223658          0.4079534
fac.wyksztalcenie2 fac.wyksztalcenie3 fac.wyksztalcenie4
         0.6821323          1.6044290          4.5113986
  fac.miejscowosc2   fac.miejscowosc3
         1.2969024          1.6187689
```

W przypadku tego modelu istotnymi zmiennymi objaśniającymi są wiek, wykształcenie oraz wielkość miejscowości.

Poziomem bazowym stworzonej przez nas zmiennej *fac.wiek* jest przedział wiekowy 16-25 lat. W naszym modelu wiek wpływa ograniczająco na częstotliwość chodzenia do kina – dla obu kolejnych przedziałów wiekowych (26-40 lat oraz 41-50 lat) szansa maleje o ok. 60% w stosunku do poziomu bazowego.

Z modelu wynika, że szansa zainteresowania kinem jest najmniejsza dla osób z wykształceniem zasadniczym zawodowym (o ok. 32% mniejsza niż dla osób z wykształceniem podstawowym), większa dla osób z wykształceniem średnim (o ok. 60% większa niż dla osób z wykształceniem podstawowym), a zdecydowanie największa dla grupy osób z wykształceniem wyższym i studentów – aż 4,5 razy większa niż dla grupy osób z wykształceniem podstawowym.

Zgodnie z oczekiwaniami z naszego modelu wynika, że osoby mieszkające w miastach więcej korzystają z nośnika kultury, jakim jest kino, niż osoby mieszkające w miejscowościach wiejskich. Szansa zainteresowania kinem w małych i dużych miastach wzrasta o odpowiednio 30% i 62% w stosunku do szansy na wsiach.

Przejdźmy do modelu związanego ze zmienną *czy.ksiazki*0. Przypomnijmy, że ta zmienna uznaje za czytelników książek osoby, które w ostatnim roku przeczytały co najmniej jedną książkę. Oto jak przedstawiają się eksponenty ocen kolejnych współczynników:

```
       (Intercept)          fac.plec1          fac.wiek2          fac.wiek3
         9.1443757          2.5428687          0.6002627          0.9619586
fac.wyksztalcenie2 fac.wyksztalcenie3 fac.wyksztalcenie4
         0.3082763          0.7365457          1.8511682
```

W przypadku tego modelu istotnymi zmiennymi objaśniającymi są płeć, wiek i wykształcenie.

Kobiety mają 2,5 razy większą szansę przeczytania książki niż mężczyźni.

Według modelu ludzie młodzi są najaktywniejszą grupą czytelników. Wraz z przejściem do drugiego przedziału wiekowego (26-40 lat) szansa maleje o 40%, a w trzecim przedziale (41-50 lat) osiąga prawie tak wysoki poziom jak w pierwszym (jest o tylko 4% niższa).

Zgodnie z oczekiwaniami najbardziej aktywną czytelniczo z badanych grup okazały się osoby z wykształceniem wyższym, a najmniej aktywną osoby po szkole zasadniczej zawodowej. W odniesieniu do poziomu bazowego tej zmiennej (osób z wykształceniem podstawowym) osiągają odpowiednio o 85% większą i 69% mniejszą szansę przeczytania przynajmniej jednej książki. Dość zaskakującym jest wynik mówiący, że dla osób z wykształceniem średnim szansa ta jest o 26% mniejsza niż dla osób z najniższym wykształceniem.

Przejdźmy do modelu związanego ze zmienną *czy.ksiazki*10. Przypomnijmy, że ta zmienna uznaje za czytelników książek osoby, które w ostatnim roku przeczytały co najmniej 10 książek. Oto jak przedstawiają się eksponenty ocen kolejnych współczynników:

```
      (Intercept)          fac.plec1          fac.wiek2          fac.wiek3
        0.9011803          1.6508767          0.8224651          1.4209970
fac.wyksztalcenie2 fac.wyksztalcenie3 fac.wyksztalcenie4          fac.stan2
        0.3656689          0.7868143          1.3847256          0.7207667
         fac.stan3          fac.stan4
        0.5955334          0.5485664
```

W tym modelu istotnymi zmiennymi objaśniającymi są te same, co dla poprzedniego modelu oraz dodatkowo stan cywilny.

Podobnie jak w poprzednim modelu badane kobiety okazały się bardziej aktywnymi czytelnikami od mężczyzn, ale różnica nie jest już tak wyraźna – szansa dla kobiet jest o 65% większa.

W odróżnieniu od modelu opartego na zmiennej *czy.ksiazki*0, w tym modelu największą szansę przeczytania co najmniej 10 książek w ciągu roku mają osoby po 40 roku życia – jest ona o 42% większa niż dla osób młodych. Podobnie jak w tamtym modelu, najmniej czytającą grupą okazały się osoby w wieku 26-40 lat.

Oceny współczynników stojących przy zmiennych odpowiadających za wykształcenie są zbliżone do tych uzyskanych w poprzednim modelu. Ponownie najbardziej aktywnymi czytelnikami okazują się osoby z wyższym wykształceniem (szansa o 38% większa niż dla osób z wykształceniem podstawowym), a osoby z wykształceniem zasadniczym zawodowym i średnim mają szansę przeczytania co najmniej 10 książek w ciągu roku odpowiednio o 63% i 21% mniejszą niż grupa będąca poziomem bazowym.

Dodatkową istotną zmienną okazał się stan cywilny. Według modelu najaktywniejszymi czytelnikami są ludzie w stanie wolnym. Wraz ze zmianą stanu cywilnego szansa przeczytania co najmniej 10 książek w ciągu roku maleje. Dla osób w związkach, małżeństw i rozwodników/wdowców szansa maleje odpowiednio o 28%, 40% i 45%.

W tablicy 3.3 znajduje się zestawienie ocen współczynników budowanych modeli po przekształceniu ich funkcją $exp(x)$, czyli wzrostu lub spadku szansy w odniesieniu do poziomu bazowego dla kolejnych poziomów zmiennych w modelach. Znak „–” oznacza, że dana zmienna nie została uwzględniona w ostatecznym modelu. Kropki i gwiazdki przy ocenach współczynników odpowiadają rzędowi wielkości p-wartości w teście, w którym hipotezą zerową jest nieistotność danej zmiennej. Przypomnijmy dokładniej ich znaczenie:

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

zmienna	poziom	**model *czy.kino***	**model *czy.ksiazki*0**	**model *czy.ksiazki*10**
wiek	26-40 lat	0,42 ***	0,60 *	0,82
	41-50 lat	0,41 ***	0,96	1,42 .
płeć	kobiety	–	2,54 ***	1,65 ***
wykształcenie	zawodowe	0,68	0,30 *	0,37 **
	średnie	1,60	0,74	0,79
	wyższe/studenci	4,51 ***	1,85	1,38
wielkość miejscowości	małe miasto	1,30	–	–
	duże miasto	1,62 *		
stan cywilny	partner	–	–	0,72 .
	małżonek			0,60 **
	rozwodnik/wdowiec			0,55

Tablica 3.3: Zestawienie ocen współczynników (przekształconych funkcją $exp(x)$) w modelach związanych ze zmiennymi *czy.kino*, *czy.ksiazki*0 i *czy.ksiazki*10

3.6. Porównanie modeli i wnioski

W punktach przedstawimy wnioski wynikające z porównania zbudowanych modeli.

- Na wstępie przypomnijmy, że **rozpatrywana próba jest reprezentatywna dla populacji polskich aktywnych internautów, a nie dla populacji wszystkich Polaków**. Jak już wspomniałyśmy w rozdziale 3.1, wynika to ze specyfiki badania, którego głównym celem było przeanalizowanie wpływu nieformalnych obiegów treści kulturowych na rozwój kultury. Na pewno fakt ten zawyża wielkość odsetka osób, które czytają książki i chodzą do kina. Jeśli dana osoba ma warunki materialne pozwalające na stały dostęp do Internetu, to z pewnością stać ją też na regularne wizyty w kinie oraz na kupno książek.

- **Na przewidywany poziom zainteresowania kinem istotny wpływ ma miejsce zamieszkania.** Ten wynik jest dość intuicyjny, można go uzasadnić tym, że w miejscowościach wiejskich dostęp do tego typu rozrywki jest utrudniony, natomiast duże miasta oferują mnogość kin i szeroki repertuar filmów. Jednocześnie ludzie z dużych miast prowadzą inny tryb życia, być może są bardziej otwarci na tego typu nośnik kultury. Zawężenie grupy badanych osób do aktywnych internautów sprawia, że wyestymowane zainteresowanie kinem w miejscowościach wiejskich w stosunku do miast może być wyższe niż w rzeczywistości. Gdyby zbadać grupę reprezentatywną dla wszystkich Polaków, czyli do badanej grupy dołączyć w szczególności osoby mieszkające we wsiach, o niskim statusie majątkowym, nie dysponujące stałym dostępem do Internetu, różnica w szansie zainteresowania kinem prawdopodobnie pogłębiłaby się.

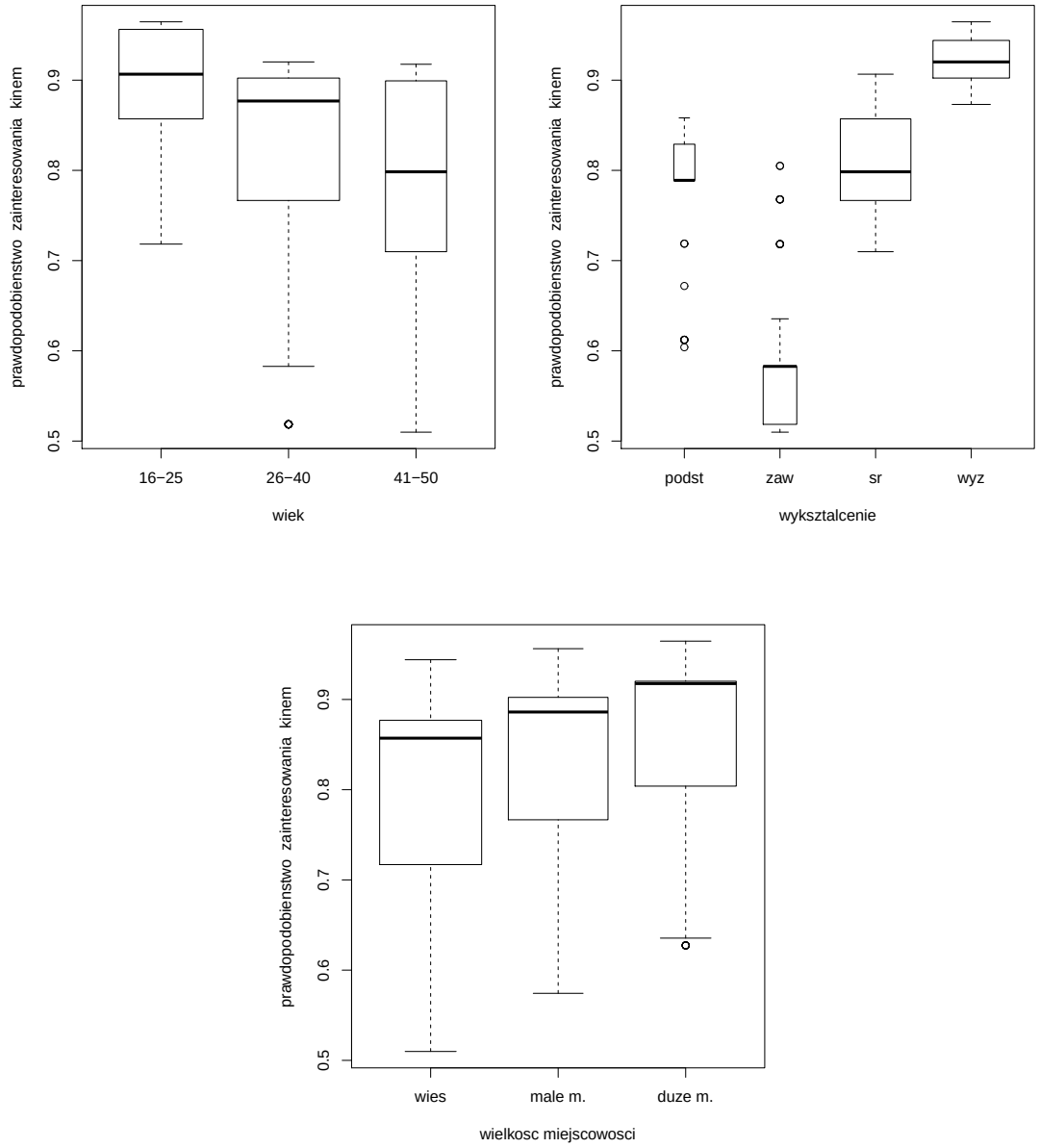
- **W badaniu zmiennej *czy.ksiazki*10 w odróżnieniu od *czy.ksiazki*0 istotny okazał się stan cywilny**. Przy takim rozróżnieniu najwięcej czytającą grupą badanych okazali się ludzie będący w stanie wolnym. Nie jest to dużym zaskoczeniem - ludzie nie zaangażowani w bliskie relacje z jedną osobą, nie posiadający pochłaniającego życia rodzinnego, mają więcej czasu dla siebie, na swoje pasje, przyjemności. Aczkolwiek otrzymany wynik mówiący, że najmniej sięgają po książki rozwodnicy i wdowcy, już

nie jest tak samo oczywisty. Niewątpliwie nie można ich postrzegać jako wolnych, choć takimi w większości są. Być może to, że nie czytają dużo książek wiąże się z nagłą zmianą w życiu. Muszą więcej czasu poświęcić na rzeczy, którymi, być może, do tej pory niewiele lub nawet w ogóle się nie zajmowali. Zauważmy, że **liczebność grupy rozwodników i wdowców jest dużo niższa od liczebności pozostałych grup** (widać to na rysunku 3.4), przez co wyniki dotyczące tej grupy mogą znacząco odbiegać od rzeczywistości.

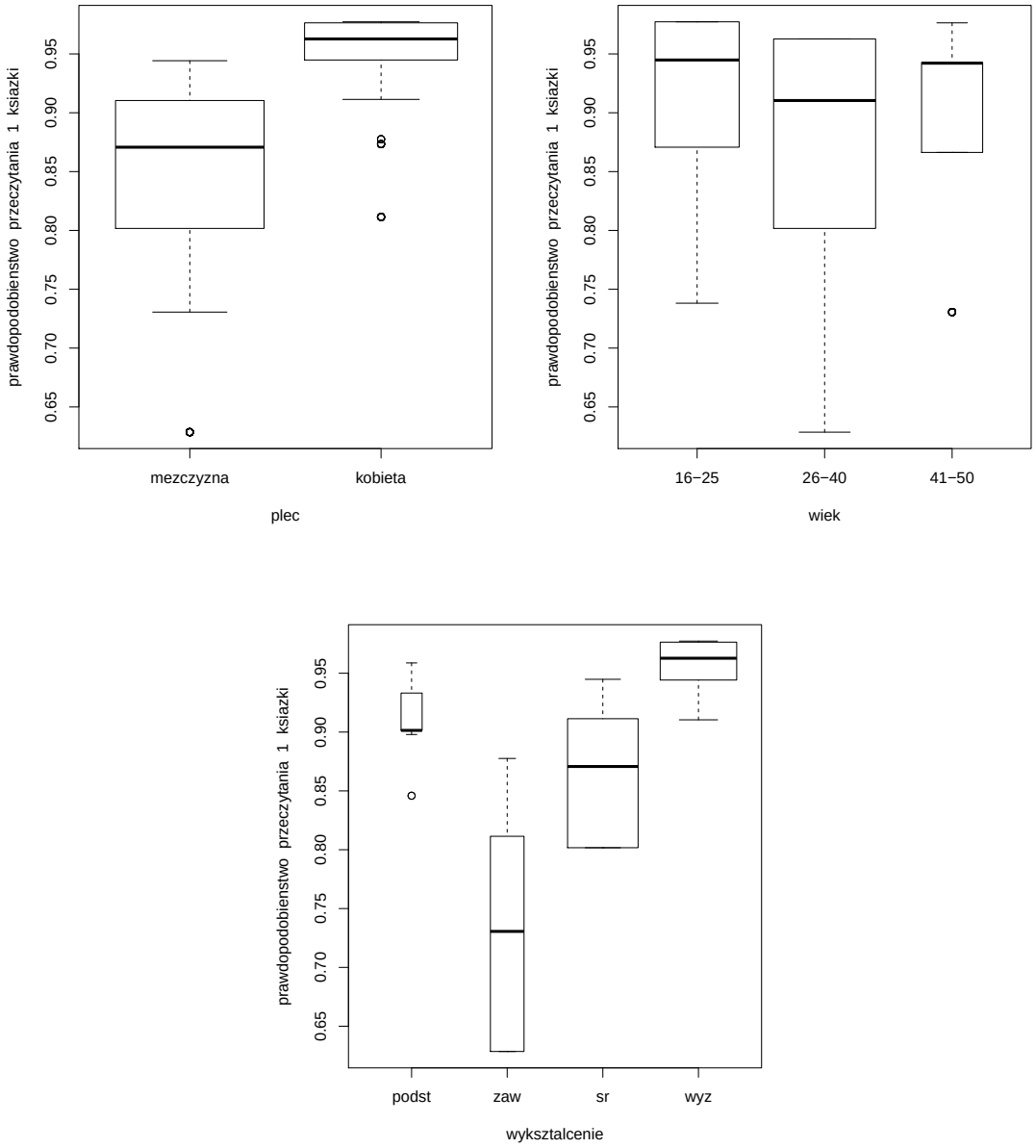
- Kolejnym ciekawym spostrzeżeniem jest to, że we wszystkich trzech modelach bardzo **podobnie plasują się uwzględnione przez sondaż etapy edukacji**. Zdecydowanie najmniej korzystają z badanych nośników kultury osoby z wykształceniem zasadniczym zawodowym, a najwięcej studenci i osoby, które ukończyły uczelnię wyższą. Pierwsza obserwacja prawdopodobnie wynika z tego, że osoby, które decydują się na ukończenie szkoły zawodowej mają nierzadko trudną sytuację materialną, bądź nie szczycą się wysokimi wynikami w nauce. Obydwie te rzeczy w większości nie sprzyjają ani czytaniu książek, ani częstemu chodzeniu do kina. Natomiast studenci czytają dużo naukowych książek, wymaganych przez swój kierunek, ale również w ramach relaksu więcej chodzą do kina. Podobnie grupa ludzi, którzy już pracują, w ten sposób spędza wolny czas, odpoczywa przy czytaniu książek i oglądaniu filmów. Jeśli chodzi o pozostałe dwie badane grupy, czyli osoby z wykształceniem podstawowym i średnim, to ta pierwsza więcej czyta, natomiast druga więcej chodzi do kina. Jest to być może spowodowane tym, że osoby z wykształceniem podstawowym mające niewiele więcej niż 16 lat, są zazwyczaj w liceum lub technikum. W obydwu instytucjach czyta się sporo lektur, ale też podręczników do poszczególnych przedmiotów. Uczniowie szkół ponadgimnazjalnych mają często mniej czasu na chodzenie do kina, zwłaszcza te osoby, które przygotowują się do matury. Brak dostępu do stałego łącza internetowego może być skorelowany z utrudnionym dostępem do edukacji, co mogło sprawić iż pewna grupa osób z niskim wykształceniem nie została uwzględniona w badanej próbie internautów, przez co otrzymane zróżnicowanie we wskaźnikach zainteresowania czytelnictwem i oglądaniem filmów w kinie między osobami z wykształceniem niskim a wysokim zmniejszyło się.

- Kolejnym uzyskanym przez nas wynikiem jest **niskie zainteresowanie czytelnictwem w grupie ludzi w wieku 26-40 lat w stosunku do innych grup** – zarówno w analizie zmiennej *czy.ksiazki*0, jak i zmiennej *czy.ksiazki*10 szansa zainteresowania czytaniem książek była dla tej grupy mniejsza niż dla grupy młodszej i dla grupy starszej. Być może jest to spowodowane tym, że ten wiek jest najbardziej aktywnym okresem w życiu człowieka – jest to na ogół okres intensywnego rozwoju kariery zawodowej i życia rodzinnego. Mała ilość czasu wolnego może powodować brak zainteresowania czytelnictwem. **Otrzymany wynik tłumaczy, dlaczego zdecydowałyśmy się na faktoryzację zmiennej *wiek***. Gdybyśmy tego nie zrobiły, model regresji wymusiłby monotoniczną, a nawet liniową zależność szansy od wieku. Z otrzymanego modelu wynika, że ta zależność nie jest liniowa.

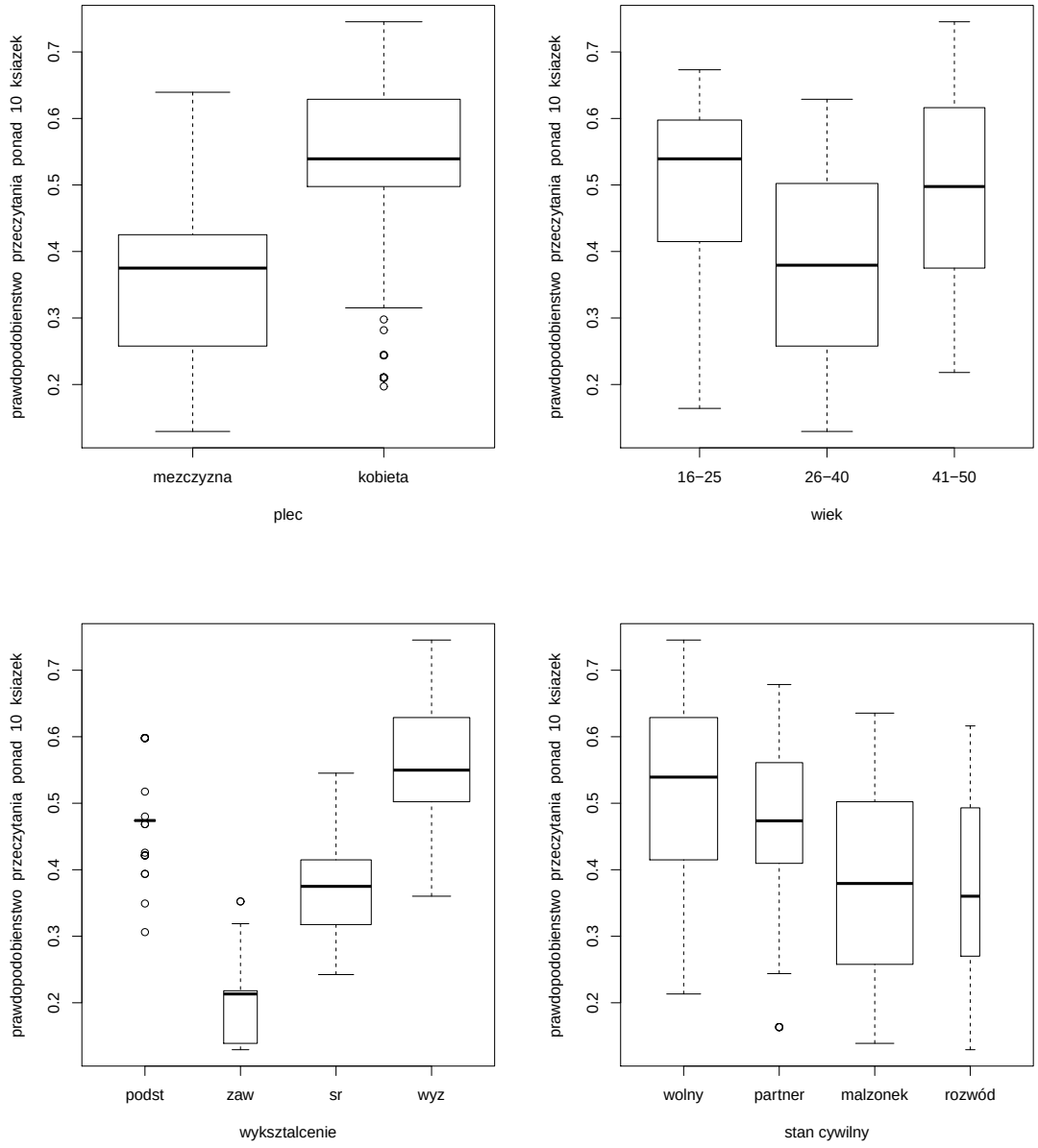
- Dość zaskakującym wynikiem okazała się tak **istotna i wyraźna zależność zainteresowania czytelnictwem od płci**. Zarówno przy analizie zmiennej *czy.ksiazki*0, jak i zmiennej *czy.ksiazki*10 badane kobiety wykazały dużo większe zamiłowanie do czytania książek niż mężczyźni. Trudno jest podać sensowną interpretację i prawdopodobne przyczyny takiego wyniku bez narażania się na zarzuty o stronniczość względem którejś z płci. Zatem powstrzymamy się od komentarza, aby uniknąć wniosków, które mogłyby zostać uznane za krzywdzące.



Rysunek 3.2: Wykresy pudełkowe dla prawdopodobieństwa zainteresowania kinem (zmienna objaśniana *czy.kino*) w zależności od poszczególnych poziomów zmiennych objaśniających $fac.wiek$, $fac.wyksztalcenie$ i $fac.miejscowosc$; grubość każdego z pudełek jest proporcjonalna do pierwiastka z liczebności danej grupy



Rysunek 3.3: Wykresy pudełkowe dla prawdopodobieństwa zainteresowania czytelnictwem (zmienna objaśniana $czy.ksiazki0$) w zależności od poszczególnych poziomów zmiennych objaśniających $fac.plec$, $fac.wiek$ i $fac.wyksztalcenie$; grubość każdego z pudełek jest proporcjonalna do pierwiastka z liczebności danej grupy



Rysunek 3.4: Wykresy pudełkowe dla prawdopodobieństwa zainteresowania czytelnictwem (zmienna objaśniana *czy.ksiazki*10) w zależności od poszczególnych poziomów zmiennych objaśniających $fac.plec$, $fac.wiek$, $fac.wyksztalcenie$ i $fac.stan$; grubość każdego z pudełek jest proporcjonalna do pierwiastka z liczebności danej grupy

Podsumowanie

W pracy tej przeprowadziłyśmy analizę danych rzeczywistych dotyczących nośników kultury, jakimi są kino i książki. Modelowanie oparłyśmy na regresji logistycznej, której zarys, wraz ze sposobem estymowania współczynników, przedstawiłyśmy w rozdziale teoretycznym.

Regresja służy do opisu zależności pomiędzy zmiennymi. W szczególności regresja logistyczna modeluje zmienną binarną mówiącą, czy dane zdarzenie nastąpiło, czy nie, w zależności od innych, dowolnych zmiennych. Nasza analiza dotyczyła zainteresowania czytaniem książek i chodzeniem do kina w zależności od płci, wieku, wykształcenia, miejsca zamieszkania, liczby osób w gospodarstwie domowym oraz stanu cywilnego. Ze zmiennych objaśniających stworzyłyśmy wektory jakościowe, co w przypadku płci, wykształcenia, miejsca zamieszkania i stanu cywilnego jest w miarę naturalne, natomiast w przypadku wieku i liczby osób w gospodarstwie nie wiedziałyśmy, czy istnieje reguła opisująca wzrost lub spadek zainteresowania nośnikami kultury wraz ze wzrostem wieku. Co do liczby osób, okazała się ona w każdym z modeli nieistotnym czynnikiem, natomiast jeśli chodzi o wiek, wprowadzenie takiego podziału okazało się słuszne, gdyż zależność od wieku okazała się niemonotoniczna w modelach regresji opisujących zainteresowanie książkami. Zmienne binarne stworzyłyśmy w oparciu o liczbę przeczytanych książek i wyjść do kina w ostatnim roku. Ponieważ nie byłyśmy do końca przekonane, co uznać za czytanie książek, stworzyłyśmy dwie zmienne – jedna z nich za osobę czytającą uznaje taką, która przeczytała co najmniej jedną książkę, a druga taką, która przeczytała ich powyżej dziesięciu w ostatnim roku. Za pomocą odpowiednich funkcji, opisanych w rozdziale 2, udało nam się zbudować trzy modele, które w zadowalającym stopniu wyznaczają zależność pomiędzy istotnymi zmiennymi. Diagnoza modeli polegająca na testowaniu hipotez o ich poprawności i dopasowaniu do danych wyszła pozytywnie.

Otrzymane wyniki zinterpretowałyśmy odwołując się do własnych doświadczeń i obserwacji, na co pozwolił wybrany przez nas temat i charakter danych. Być może nie wszystkie wysnute przez nas wnioski są trafne, główną przyczyną może być ograniczona liczba czynników wpływających na zmienne objaśniane, do których nie miałyśmy dostępu. Niemniej jednak w większości wyniki są intuicyjnie poprawne, co może przemawiać za niewielkim odstępstwem od rzeczywistości.

Dodatek A

Kody pakietu R użyte w pracy

A.1. Wczytanie i transformacja danych

Wczytanie danych pobranych ze strony internetowej:

```
> obiegi = read.table("C:/sciezka_do_pliku/dane_sondaz.csv", sep=";",
  header=T)
```

Wczytanie i binaryzacja odpowiedzi na pytanie o częstotliwość chodzenia do kina:

```
> kino = obiegi[,248]
> czy.kino=c(1:1284)
> for (i in 1:1284) if (kino[i]<5) czy.kino[i]=1 else czy.kino[i]=0
```

Wczytanie i dwa sposoby binaryzacji odpowiedzi na pytanie o liczbę przeczytanych książek:

```
> ileksiazek = obiegi[,160]
> czy.ksiazki0=c(1:1284)
> if (ileksiazek[i]>0) czy.ksiazki0[i]=1 else czy.ksiazki0[i]=0
> czy.ksiazki10=c(1:1284)
> if (ileksiazek[i]>10) czy.ksiazki10[i]=1 else czy.ksiazki10[i]=0
```

Wczytanie i faktoryzacja zmiennej *plec*:

```
> plec = obiegi[,152]
> fac.plec=factor(factor(plec, labels=c("1","0")))
```

Wczytanie zmiennej *rokurodzenia*, utworzenie zmiennej *wiek* oraz jej faktoryzacja:

```
> rokurodzenia = obiegi[,151]
> wiek = 2011 - rokurodzenia
> fac.wiek=factor(factor(wiek, labels=c("1","1","1","1","1","1","1","1",
  "1","1","2","2","2","2","2","2","2","2","2","2","2","2","2","2",
  "3","3","3","3","3","3","3","3","3","3")))
```

Wczytanie zmiennej *wyksztalcenie* i jej faktoryzacja:

```
> wyksztalcenie = obiegi[,159]
> fac.wyksztalcenie=factor(factor(wyksztalcenie, labels=c("1","2","3","4")))
```

Wczytanie zmiennej *wielkoscmiejscowosci* i jej faktoryzacja:

```
> wielkoscmiejscowosci = obiegi[,156]
> fac.miejscowosc=factor(factor(wielkoscmiejscowosci, labels=c("1","2","2",
  "2","2","3","3","3","3")))
```

Wczytanie zmiennej *liczbaosob* (liczba osób w gospodarstwie domowym) i jej faktoryzacja:

```
> liczbaosob=obiegi[,346]
> fac.liczbaosob=factor(factor(liczbaosob, labels=c("1","2","2","2","3","3",
  "3","3")))
```

Wczytanie zmiennej *stan* (stan cywilny) i jej faktoryzacja:

```
> stan=obiegi[,380]
> fac.stan=factor(factor(stan,labels=c("1","2","3","4")))
```

A.2. Budowa modelu związanego z kinem

Użycie funkcji `glm()` i `summary()` do budowy pierwszego modelu związanego z kinem:

```
> regresjakino1 = glm(formula = czy.kino ~ fac.plec+fac.wiek+
  fac.wyksztalcenie+fac.miejscowosc+fac.liczbaosob+fac.stan,
  family=binomial())
> summary(regresjakino1)
```

Użycie funkcji `step()` do korekty pierwszego modelu związanego z kinem:

```
> step(regresjakino1, direction="backwards")
```

Budowa modeli opartych na pojedynczych zmiennych objaśniających:

```
> regresjakino.plec = glm(formula = czy.kino ~ fac.plec, family=binomial())
> summary(regresjakino.plec)

> regresjakino.liczbaosob = glm(formula = czy.kino ~ fac.liczbaosob,
  family=binomial())
> summary(regresjakino.liczbaosob)

> regresjakino.stan = glm(formula = czy.kino ~ fac.stan, family=binomial())
> summary(regresjakino.stan)
```

Budowanie kolejnych modeli związanych z kinem:

```
> regresjakino2 = glm(formula = czy.kino ~ fac.plec+fac.wiek+
  fac.wyksztalcenie+fac.miejscowosc+fac.stan, family=binomial())
> summary(regresjakino2)

> regresjakino3 = glm(formula = czy.kino ~ fac.plec+fac.wiek+
  fac.wyksztalcenie+fac.miejscowosc+fac.stan+fac.stan:fac.wiek+
  fac.plec:fac.wyksztalcenie+fac.stan:fac.miejscowosc+
  fac.wiek:fac.wyksztalcenie, family=binomial())
> summary(regresjakino3)

> regresjakino4 = glm(formula = czy.kino ~ fac.wiek+fac.wyksztalcenie+
  fac.miejscowosc+fac.stan+fac.stan:fac.wiek+fac.stan:fac.miejscowosc,
  family=binomial())
> summary(regresjakino4)

> regresjakino5 = glm(formula = czy.kino ~ fac.wiek+fac.wyksztalcenie+
  fac.miejscowosc, family=binomial())
> summary(regresjakino5)
```

A.3. Budowa modelu związanego ze zmienną *czy.ksiazki*0

Budowa pierwszego modelu związanego ze zmienną *czy.ksiazki*0:

```
> regresjaksiazki0a<-glm(formula=czy.ksiazki0~fac.plec+fac.wiek+
  fac.wyksztalcenie+ fac.miejscowosc+fac.liczbaosob+fac.stan,
  family=binomial())
> summary(regresjaksiazki1)
```

Budowa modeli opartych na pojedynczych zmiennych objaśniających:

```
> regresjaksiazki0.miejscowosc<-glm(formula=czy.ksiazki0~fac.miejscowosc,
  family=binomial())
> summary(regresjaksiazki0.miejscowosc)

> regresjaksiazki0.liczbaosob<-glm(formula=czy.ksiazki0~fac.liczbaosob,
  family=binomial())
> summary(regresjaksiazki0.liczbaosob)

> regresjaksiazki0.stan<-glm(formula=czy.ksiazki0~fac.stan, family=binomial())
> summary(regresjaksiazki0.stan)
```

Budowa kolejnych modeli związanych ze zmienną *czy.ksiazki*0:

```
> regresjaksiazki0b <- glm(formula = czy.ksiazki0 ~ fac.plec+fac.wiek+
  fac.wyksztalcenie+fac.stan+fac.wiek:fac.stan, family=binomial())
> summary(regresjaksiazki0b)

> regresjaksiazki0c<-glm(formula=czy.ksiazki0~fac.plec+fac.wiek+
  fac.wyksztalcenie, family=binomial())
> summary(regresjaksiazki0c)
```

A.4. Budowa modelu związanego ze zmienną *czy.ksiazki*10

Budowa pierwszego modelu związanego ze zmienną *czy.ksiazki*10:

```
> regresjaksiazki10a<-glm(formula=czy.ksiazki10~fac.plec+fac.wiek+
  fac.wyksztalcenie+ fac.miejscowosc+fac.liczbaosob+fac.stan,
  family=binomial())
> summary(regresjaksiazki10a)
```

Budowa modeli opartych na pojedynczych zmiennych objaśniających:

```
> regresjaksiazki10.miejscowosc<-glm(formula=czy.ksiazki10~fac.miejscowosc,
  family=binomial())
> summary(regresjaksiazki10.miejscowosc)

> regresjaksiazki10.liczbaosob<-glm(formula=czy.ksiazki10~fac.liczbaosob,
  family=binomial())
> summary(regresjaksiazki10.liczbaosob)
```

Budowa ostatecznego modelu związanego ze zmienną *czy.ksiazki*10:

```
> regresjaksiazki10b<-glm(formula=czy.ksiazki10~fac.plec+fac.wiek+
  fac.wyksztalcenie+fac.stan, family=binomial())
> summary(regresjaksiazki10b)
```

A.5. Tworzenie wykresów

Nadanie nowych wartości poszczególnym poziomom zmiennych objaśniających w celu otrzymania etykiet na wykresach – ponowne użycie funkcji `factor()`:

```
> fac.plec=factor(factor(plec, labels=c("mezczyzna","kobieta")))
> fac.wiek=factor(factor(wiek, labels=c("16-25","16-25","16-25","16-25",
  "16-25","16-25","16-25","16-25","16-25","16-25","26-40","26-40","26-40",
  "26-40","26-40","26-40","26-40","26-40","26-40","26-40","26-40","26-40",
  "26-40","26-40","26-40","41-50","41-50","41-50","41-50","41-50","41-50",
  "41-50","41-50","41-50","41-50")))
> fac.wyksztalcenie=factor(factor(wyksztalcenie, labels=c("podst","zaw","sr",
```

```
  "wyz")))
> fac.miejscowosc=factor(factor(wielkoscmiejscowosci, labels=c("wies",
  "male m.","male m.","male m.","male m.","duze m.","duze m.","duze m.",
  "duze m.")))
> fac.stan=factor(factor(stan,labels=c("wolny","partner","malzonek",
  "rozwod")))
```

Tworzenie wykresów z rysunku 3.1:

```
> interaction.plot(fac.stan,fac.wiek,czy.kino,lwd=2)
> interaction.plot(fac.stan,fac.miejscowosc,czy.kino,lwd=2)
```

Tworzenie wykresów z rysunku 3.2:

```
> plot(fac.wiek, regresjakino5$fitted.value, xlab="wiek",
  ylab="prawdopodobienstwo   zainteresowania   kinem", varwidth = TRUE)

> plot(fac.wyksztalcenie, regresjakino5$fitted.value,
  xlab="wyksztalcenie",
  ylab="prawdopodobienstwo   zainteresowania   kinem", varwidth = TRUE)

> plot(fac.miejscowosc, regresjakino5$fitted.value,
  xlab="wielkosc miejscowosci",
  ylab="prawdopodobienstwo   zainteresowania   kinem", varwidth = TRUE)
```

Tworzenie wykresów z rysunku 3.3:

```
> plot(fac.plec, regresjaksiazki0c$fitted.value, xlab="plec",
  ylab="prawdopodobienstwo   przeczytania   1   ksiazki", varwidth = TRUE)

> plot(fac.wiek, regresjaksiazki0c$fitted.value, xlab="wiek",
  ylab="prawdopodobienstwo   przeczytania   1   ksiazki", varwidth = TRUE)

> plot(fac.wyksztalcenie, regresjaksiazki0c$fitted.value,
  xlab="wyksztalcenie",
  ylab="prawdopodobienstwo   przeczytania   1   ksiazki", varwidth = TRUE)
```

Tworzenie wykresów z rysunku 3.4:

```
> plot(fac.plec, regresjaksiazki10b$fitted.value, xlab="plec",
  ylab="prawdopodobienstwo   przeczytania   ponad   10   ksiazek",
  varwidth = TRUE)

> plot(fac.wiek, regresjaksiazki10b$fitted.value, xlab="wiek",
  ylab="prawdopodobienstwo   przeczytania   ponad   10   ksiazek",
  varwidth = TRUE)

> plot(fac.wyksztalcenie, regresjaksiazki10b$fitted.value,
  xlab="wyksztalcenie",
  ylab="prawdopodobienstwo   przeczytania   ponad   10   ksiazek",
  varwidth = TRUE)

> plot(fac.stan, regresjaksiazki10b$fitted.value, xlab="stan cywilny",
  ylab="prawdopodobienstwo   przeczytania   ponad   10   ksiazek",
  varwidth = TRUE)
```

A.6. Diagnostyka modeli

Wczytanie pakietu do obliczenia R^2:

```
> library(rms)
```

Obliczanie R^2 dla modelu *regresjakino*5:

```
> lrm(regresjakino5)
> lrmregresjakino5<-lrm(regresjakino5)
> lrmregresjakino5$stats
```

Obliczanie R^2 dla modelu *regresjaksiazki*0*c*:

```
> lrm(regresjaksiazki0c)
> lrmregresjaksiazki0c<-lrm(regresjaksiazki0c)
> lrmregresjaksiazki0c$stats
```

Obliczanie R^2 dla modelu *regresjaksiazki*10*b*:

```
> lrm(regresjaksiazki10b)
> lrmregresjaksiazki10b<-lrm(regresjaksiazki10b)
> lrmregresjaksiazki10b$stats
```

Obliczanie statystyki Pearsona X^2 dla modelu *regresjakino*5:

```
> sum(residuals(regresjakino5,type="pearson")^2)
```

Obliczanie statystyki Pearsona X^2 dla modelu *regresjaksiazki*0*c*:

```
> sum(residuals(regresjaksiazki0c,type="pearson")^2)
```

Obliczanie statystyki Pearsona X^2 dla modelu *regresjaksiazki*10*b*:

```
> sum(residuals(regresjaksiazki10b,type="pearson")^2)
```

Bibliografia

[1] Przemysław Biecek, *Przewodnik po pakiecie R*, Oficyna Wydawnicza GiS, Wrocław 2011.

[2] Jeff Gill, *Generalized Linear Models: A Unified Approach*, Sage Publications, Floryda 2001.

[3] R. I. Jennrich, P. F. Sampson, *Newton-Raphson and Related Algorithms for Maximum Likehood Variance Component Estimation*, Technometrics, vol. 18, nr 1, 1976, 11–17.

[4] Christopher Manning, *Logistic regression with R*, `http://nlp.stanford.edu/manning/courses/ling289/logistic.pdf`, 2007.

[5] Wojciech Niemiro, *Statystyka*, 2011.

[6] Simon J. Sheather, *A Modern Approach to Regression with R*, Springer, Texas 2008.

[7] *Dokumentacja pakietu R*, `http://finzi.psych.upenn.edu/R/library/stats/html/`, dostęp dnia 13.05.2012r.

[8] Internetowa encyklopedia *Wikipedia*, `http://pl.wikipedia.org/wiki/Regresja_(statystyka)`, dostęp dnia 2.06.2012r.