Uogólnione modele liniowe

Laboratorium nr 3

3.1 Zbiór **malaria** zawiera informację na temat liczby osób posiadających przeciwciała (Spositive) pośród wszystkich badanych osób (Number) w danej grupie wiekowej (Age). (Przeciwciała produkowane przez organizm jako ochrona przed malarią pozostają w organizmie także po wyzdrowieniu i są wykrywane przez test serologiczny – osoby z przeciwciałami mają dodatni wynik testu serologicznego.)

(a) Dopasować model regresji logistycznej używając wieku jako jedynej zmiennej objaśniającej.

(b) Dopasować model regresji liniowej dla logitów proporcji z wagami `n(proporcja)(1-proporcja)` i sprawdzić, że ten model i model z poprzedniego punktu dają bardzo zbliżone wyniki. Dlaczego tak jest?

(c) Używając modelu logistycznego, oszacować wiek, dla którego prawdopodobieństwo dodatniego odczynu wynosi 1/4.

(d) Skonstruować przedział ufności dla prawdopodobieństwa dodatniego odczynu w wieku 20 lat.

(e) Narysować wykres frakcji przypadków dodatniego odczynu serologicznego w zależności od wieku wraz z dopasowaną krzywą.

3.2 Zbiór **finance** zawiera dane dotyczące kondycji finansowej 46 przedsiębiorstw na podstawie czterech wskaźników finansowych.

(a) Dopasować model logistyczny. Przetestować hipotezę, że zbiór zawiera zmienne istotne i obliczyć procent dewiacji wyjaśnianej przez model.

(b) Za pomocą instrukcji `drop1` dokonać sekwencyjnego usunięcia z modelu nieistotnych zmiennych. Porównać mniejszy model z modelem wyjściowym. Obliczyć procent dewiacji wyjaśnianej.

(c) Za pomocą instrukcji `step` dokonać sekwencyjnego usunięcia z modelu nieistotnych zmiennych. Porównać mniejszy model z modelem wyjściowym. Obliczyć procent dewiacji wyjaśnianej.

(d) Rozpatrzyć rezydua oparte na dewiacjach. Wyliczyć studentyzowane rezydua i narysować ich wykres kwantylowy.

(e) Wyrzucić obserwacje potencjalnie odstające, dopasować powtórnie model i obliczyć dla niego procent dewiacji wyjaśnionej.

3.3 (Test Hosmera-Lemeshowa) Zbiór **HosLemData** zawiera zmienną objaśnianą y i zmienną objaśniającą x.

(a) Dopasować model regresji logistycznej do danych z **HosLemData** i zastanowić się nad możliwością zbadania jakości dopasowania za pomocą testu opartego na dewiancjach i testu Pearsona.

(b) Zaimplementować test Hosmera-Lemeshowa z liczbą grup $g = 10$. Co z niego wynika?

(c) Przeprowadzić na tych samych danych testy Hosmera-Lemeshowa z $g = 9$, $g = 11$ i $g = 12$. Co z nich wynika?

3.4 Rozważamy zbiór **kyphosis**.

(a) Dopasować dwa modele: `g: kyphosis~Age + Number +Start` oraz $g2$, który jest modelem g powiększonym o kwadraty zmiennych objaśniających z g.

(b) Czy model g zawiera istotne zmienne?

(c) Za pomocą procedury `step` usunąć zmienne nieistotne z modelu $g2$, tworząc tym samym model $g1$.

(d) Porównać wartości AIC dla modeli g i $g1$.

(e) Przeprowadzić test $H_0 : g$ kontra $H_1 : g1$.

(f) Przeprowadzić test Hosmera-Lemeshowa dla modelu g.