Uogólnione modele liniowe

Laboratorium nr 2

Model regresji logistycznej w R

Do dopasowania uogólnionych modeli liniowych w R służy funkcja `glm()`. Schemat użycia tej funkcji w przypadku regresji logistycznej jest następujący:

`nazwa=glm(zmienna_objasniana~zm_objasniajaca1+...+zm_objasniajaca_n,family=binomial)`. Wersja z `binomial(link=probit)` zamiast samego `binomial` daje regresję probitową. Pisząc `nazwa` lub `print(nazwa)` otrzymuje się wyniki; bardziej rozbudowane podsumowanie to `summary(nazwa)`. Dostaje się w szczególności tabelę *Coefficients*. Jej kolejne kolumny to wyestymowane współczynniki modelu, ich odchylenia standardowe, wartości statystyki testu Walda i p-wartości tego testu (hipotezą zerową jest nieistotność odpowiedniej zmiennej objaśniającej). Wiersze tabeli *Coefficients* odpowiadają kolejnym predyktorom i wyrazowi wolnemu dopasowywanego modelu.

Na obiekcie klasy `glm` można zastosować m.in.:

- `$coef` lub `$coefficients` - wektory oszacowań współczynników
- `residuals(nazwa)` lub `residuals(nazwa,"deviance")` - rezydua oparte na dewiacjach (uwaga: `nazwa$residuals` lub `residuals(nazwa,"working")` dają tzw. *working residuals*)
- `residuals(nazwa,"Pearson")` - rezydua Pearsona
- `residuals(nazwa,''response")` - roznica obserwacji i wartości dopasowanej
- `rstandard(nazwa)` - standaryzowane (studentyzowane) reszty oparte na dewiacjach
- `$df.residual` - liczba stopni swobody dla reszt
- `$fitted.values` lub `$fit` - dopasowane wartości
- funkcja `linear.predictors` wylicza oszacowanie $\log(\hat{\pi}/(1-\hat{\pi}))$
- `$family` - użyta funkcja łącząca
- `$deviance` i `$null.deviance` - wielkości wypisywane przez summary jako `Residual deviance` (dewiacja danego modelu) i `Null deviance` (dewiacja modelu minimalnego)
- `$cov.unscaled` – macierz kowariancji dla oszacowań współczynników
- przekątna macierzy daszkowej dana jest jako `hatvalues(nazwa)`

Testy oparte na dewiacjach uzyskać można za pomocą komendy `print(anova(nazwa, test=`Chi"))". W przypadku porównywania modeli model1 i model2 pisze się `print(anova(model1,model2, test=`Chi"))".

Do testowania dobrego dopasowania służyć może komenda
`print(1-pchisq(nazwa$deviance,nazwa$df.residual))`
(liczba stopni swobody $N-p$ dla rezyduów zadana jest jako `$df.residual`, a nie jako `$df`).

Przedziały ufności dla oszacowań współczynników dostać można za pomocą polecenia `confint(nazwa)`.

Do wybrania właściwych zmiennych do modelu można użyć funkcji
`step(nazwa,direction=c("both","backward","forward"),steps=...)`,
która znajduje model najlepiej dopasowany do danych jedną z metod z direction. Metoda `backward` to usuwanie najmniej istotnych zmiennych z modelu zawierającego wszystkie zmienne objaśniające aż wszystkie zmienne będą istotne. `forward` to dodawanie najbardziej istotnych zmiennych do modelu minimalnego (tylko wyraz wolny). Domyślnym kryterium oceny istotności zmiennych jest kryterium Akaike (AIC). `steps` oznacza maksymalną liczbę kroków.

Analogicznie, można używać wielokrotnie procedurę `drop1(nazwa,test='Chi')`, która z danego zbioru zmiennych odrzuca poszczególne zmienne i testuje (w oparciu o różnice dewiacji) hipotezy, że mniejsze modele są adekwatne.

Procedura `halfnorm` z biblioteki `faraway` (z cranu) daje wykres kwantylowy wartości bezwzględnych rezyduów z zaznaczonymi potencjalnymi obserwacjami odstającymi.

2.1 Rozważyć zbiór danych z $y = 0$, gdy $x = 10, 20, 30, 40$ (pojedyncza zmienna objaśniająca) i $y = 1$, gdy $x = 60, 70, 80, 90$. Dopasować do niego model regresji logistycznej. Wyestymować parametry modelu - zastanowić się nad wyjaśnieniem zaobserwowanych problemów.

2.2 Ustalić dwie liczby rzeczywiste β_1 i β_2. Wygenerować 10 wartości zmiennej x, np. z rozkładu jednostajnego na $[0, 1]$. Dla każdej z nich wyliczyć

$$\pi(x) = \frac{\exp(\beta_1 + \beta_2 x)}{1 + \exp(\beta_1 + \beta_2 x)},$$

a następnie dla każdego $\pi(x)$ wygenerować 10 niezależnych obserwacji z rozkładu $bin(1, \pi(x)$. Otrzymane dane zgrupować.

(a) Do otrzymanych danych dopasować model regresji logistycznej. Przeprowadzić test dopasowania. Obliczyć procent dewiacji wyjaśnianej przez model. Narysować odpowiednie wykresy rezyduów.

(b) Zaburzyć zmienną odpowiedzi poprzez dodanie szumu losowego z rozkładu $\mathcal{N}(0, 0.01)$. Przeprowadzić analizę otrzymanych danych analogiczną do tej z poprzedniego punktu.

2.3 Dla danych ze zbioru **bliss** dopasować model logistyczny $y \sim$ conc.

(a) Zbadać istotność conc według testu Walda i testu opartego na dewiacjach.

(b) Przeprowadzić test dopasowania modelu.

(c) Obliczyć procent dewiacji wyjaśnianej przez model.

(d) Dopasować większy model, z dwoma zmiennymi objaśniającymi conc i conc^2 i ocenić, czy wprowadzenie conc^2 jest uzasadnione.

2.4 Dopasować model logistyczny dla danych ze zbioru **bliss** w postaci rozwiniętej (danych indywidualnych). Porównać współczynniki, dewiację modelową oraz wartość statystyki opartej na dewiacjach (test ilorazu wiarogodności) dla testowania

H0: $y \sim$ const przeciwko HA: $y \sim$conc.

2.5 Zbiór **malaria** zawiera informację na temat liczby osób posiadających przeciwciała (Spositivie) 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) Używając modelu, oszacować wiek, dla którego prawdopodobieństwo dodatniego odczynu wynosi 1/4.

(c) Skonstruować przedział ufności dla prawdopodobieństwa dodatniego odczynu w wieku 20 lat. Można to zrobić np. za pomocą instrukcji
`predict(obiekt.glm,data.frame(Age=...),se.fit=T)`.

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