Zaawansowane Modele Liniowe

Lista zadań
Uogólnienia Regresji Poissona

Symulacje

1. Testowanie hipotez brzegowych – rozkład statystyki χ^2 i T.
Na wykładzie pokazano że badanie hipotezy

$$H_0 : \text{dane pochodzą z rozkł. Poissona}$$

$$vs$$

$$H_1 : \text{dane pochodzą z rozkł. ujemnego dwumianowego}$$

można przeprowadzić za pomocą statystyki $\chi^2 = D(M_1) - D(M_2)$ lub statystyki $T = \frac{\hat{\alpha}}{\sqrt{Var(\hat{\alpha})}}$.

Wygeneruj losową macierz $X \in \mathbb{M}_{1000x2}$, t.że $X_{ij} \sim^{i.i.d} N(0, \sigma = 1/\sqrt{1000})$. Następnie wyznacz ciąg predyktorów liniowych $\eta = X\beta$ dla wektora $\beta = (3,3)$ i na ich podstawie wygeneruj 10000 niezależych replikacji wektora odpowiedzi y przy założeniu hipotezy zerowej. Dla każdej replikacji wektora odpowiedzi y dopasuj model regresji ujemnej dwumianowej (glm.nb()) i regresji Poissona(glm()) i wyznacz na ich podstawie ciąg statystyk χ^2 oraz $\hat{\alpha}$. Powinieneś w ten sposób uzyskać po 10000 realizacji każdej statystyki.
Realizacje statystyki χ^2 najwygodniej wyznaczyć ze wzoru

$$\chi^2 = -2\left((l_1(\hat{\beta}^{(1)}) - (l_2(\hat{\beta}^{(2)})\right)$$

gdzie $l_i()$ jest logarytmem funkcji wiarogodności dla modelu $i = 1,2$ obliczonym w punkcie odpowiadajacym estymatorowi $\hat{\beta}^{(i)}$ $(i = 1,2)$ (możesz użyć funkcji logLik() na odpowiednim modelu). Porównaj empiryczne rozkłady statystyk χ^2 z teoretycznym podanymi na wykładzie. Uzyskaj podobne wykresy do tych z wykładu. Narysuj histogram i dorysuj do niego gęstość ciągłej części rozkładu teoretycznego. Narysuj wykres kwantylowo-kwantylowy (nie chodzi o funkję qqnorm()!!!)

Estymator $\hat{\alpha}$ jest odwrotnością estymatora *theta* z summary() dla modelu regresji ujemnej dwumianowej. Uzyskaj podobne wykresy do tych z wykładu. Narysuj wykres kwantylowo-kwantylowy (skorzystaj z funkcji qqnorm()). Narysuj histogram i dorysuj do niego gęstość ciągłej części rozkładu teoretycznego (połowa rozkładu

normalnego). Aby wyznaczyć $\hat{\sigma}$ dla dodatniej części histogramu skorzystaj z relacji:

$$\hat{\sigma} \approx \frac{F^{-1}(0.75)}{\Phi^{-1}(0.75)}$$

gdzie $F^{-1}(q)$ jest kwantylem rzędu q.

Analiza danych

2. Pobierz z platformy e-learning plik z danymi "DebTrivedi" i wczytaj go do R.

Opis dotyczący danych:
Deb and Trivedi (1997) analyze data on 4406 individuals, aged 66 and over, who are covered by Medicare, a public insurance program. The objective is to model the demand for medical care — as captured by the number of physician/non-physician office and hospital outpatient visits—by the covariates available for the patients.

Na laboratorium będziemy chcieli zbadać związek pomiędzy liczbą wizyt w gabinecie lekarskim (zmienna zależna, kolumna "ofp") i zmiennymi niezależnymi opisujacymi pacjenta:

- "hosp" – liczba pobytów w szpitalu,
- "health" – zmienna opisująca subiektywny odczucie pacjenta o jego zdrowiu,
- "numchron" – liczba przewlekłych stanów chorobowych,
- "gender" – płeć
- "school" – liczba lat edukacji
- "privins" – indykator opisujący to czy pacjent ma dodatkowe prywatne ubezpieczenie zdrowotne.

3. Wstępna analiza.

- Narysuj histogram zmiennej "ofp". Czy wykres wskazuje na obecność zjawiska nadmiernej dyspersji i/lub inflacji w zerze?
- Ze względu na znaczącą liczbę zer wprowadź zmienną pomocniczą $f(ofp) = log(ofp+0.5)$ dzięki której łatwiej będzie zbadać wzajemne zależności pomiędzy "ofp" i regresorami (funkcja f jest monotoniczna, dlatego zachowuje uporządkowanie pomiędzy punktami).

- Narysuj (dla każdego regresora osobno) na jednym rysunku boxploty dla f(ofp) w rozbiciu ze względu na przyjmowane wartości przez dany regresor. Jeżeli dla danej wartości regresora będzie mało obserwacji, pogrupuj wartości regresora i wykonaj boxplot dla pogrupowanych wartości (np. pogrupy dla zm. "hosp": ”0”, ”1”, ”2”,"3 i więcej" itp.).
- Opisz uzyskane rysunki i wyciągnij wnioski dotyczące wpływu poszczególnych regresorów na zmienną wynikową.

4. Dopasuj 6 modeli do danych:

- Model Poissona – glm()
- Model ujemny dwumianowy – glm.nb()
- Model ZIPR – zeroinfl()
- Model ZINBR – zeroinfl()
- Model Poissona z barierą – hurdle()
- Model ujemny dwumianowy z barierą – hurdle()

Następnie, tam gdzie są ku temu przesłanki, zredukuj model o niepotrzebne zmienne. Każdą redukcję potwierdź odpowiednimi testami.

Stabelaryzuj otrzymane wyniki.
Niech każda kolumna odpowiada jednemu modelowi. W poszczególnych wierszach wypisz wartości uzyskanych estymatorów (jeżeli model nie generuje określoego estymatora pozostaw wolne miejse , np. model poissona nie ma parametrów związanych z nadmierną dyspersją i inflacją zer), liczbę parametrów w modelu, logarytm funkcji wiarogodności, AIC, BIC oraz oczekiwaną liczbę zer generowanych przez model (suma funkcji rozkładu prawdopodobieństwa obliczona w 0 dla wszystkich obserwacji).
Porównaj otrzymane wyniki dla różnych modeli i opisz wnioski.