Rozszerzone Modele Liniowe

Wybór istotnych zmiennych w regresji logistycznej

Zainstaluj pakiety *devtools*, *bigstep* i *glmnet*. Zainstaluj pakiet dla logistycznego slope z repozytorium dr. Daniela Kucharczyka: *install_ github("dkucharc/glmSLOPE")*.

1. Wygeneruj macierz $X_{1000\times250}$, tak że jej elementy są niezależnymi zmiennymi losowymi z rozkładu $N(0, \sigma = 1)$. Następnie wygeneruj wektor zmiennej odpowiedzi zgodnie z modelem regresji logistycznej gdzie wektor współczynników wynosi $\beta = (c, c, c, c, c, 0, ..., 0)^T$, z $c = 9/\sqrt{n}$.

   a) Przeanalizuj te dane standardową funkcją *glm* i wybierz istotne zmienne w oparciu o p-wartości na poziomie istotności 0.05. Wyznacz liczbę prawdziwych i fałszywych odkryć i FDP. Wylicz $B(\hat{\beta}) = \frac{1}{p}\sum_{i=1}^{p}(\hat{\beta}_i - \beta_i)$, $SE(\hat{\beta}) = ||\hat{\beta} - \beta||^2$ i $SE(\hat{p}) = ||\hat{p} - p||^2$.

   b) Ustal optymalny model zgodnie z kryteriami *BIC*, *mBIC* i *mBIC2*. Dla każdego z tych kryteriów wyznacz liczbę prawdziwych i fałszywych odkryć i FDP. Następnie dla każego z tych kryteriów wyznacz estymatory współczynników regresji logistycznej wykorzystując funkcję *glm* w oparciu o odpowiednio zredukowaną macierz X. Wyznacz błędy estymacji $B(\hat{\beta})$, $SE(\hat{\beta})$ i $SE(\hat{p})$.

   c) Wyestymuj współczynniki w modelu regresji logistycznej za pomocą LASSO z walidacją krzyżową (funkcja *cv.glmnet* w bibliotece *glmnet*). Wyznacz liczbę prawdziwych i fałszywych odkryć, FDP i wszystkie trzy błędy estymacji.

   d) Ustal optymalny model za pomocą SLOPE z wektorem parametów wygładzających danych wzorem: dla $j \in \{1, \ldots, p\}$, $\lambda_j = \frac{\sqrt{n}}{2}\Phi^{-1}\left(1 - \frac{0.1j}{2p}\right)$. Wyznacz liczbę fałszywych i prawdziwych odkryć i FDP. Następnie wyznacz estymatory współczynników regresji logistycznej wykorzystując funkcję *glm* w oparciu o odpowiednio zredukowaną macierz X. Wyznacz błędy estymacji $B(\hat{\beta})$, $SE(\hat{\beta})$ i $SE(\hat{p})$.

   e) Powtórz a)-d) 20 razy i wyznacz średnie wartości wszystkich rozważanych charakterystyk. Krytycznie porównaj wszystkie testowane metody.

2. Wygeneruj wektor odpowiedzi w sytuacji gdy w modelu jest 50 istotnych predykatorów i $\beta_1 = \ldots = \beta_{50} = \frac{9}{\sqrt{n}}$. Powtórz punkty a)-d) z zadania 1 i krytycznie porównaj analizowane metody.

3. Powtórz zadania 1-2 w przypadku gdy wiersze macierzy X są niezależnymi wektorami losowymi z wielowymiarowego rozkładu normalnego $N(0, \Sigma)$ z macierzą kowariancji taką, że $\Sigma_{ii} = 1$ i $\Sigma_{ij} = 0.3$ dla $i \neq j$.

4. Wygeneruj macierz $X_{100\times25}$, tak że jej elementy są niezależnymi zmiennymi losowymi z rozkładu $N(0, \sigma = 1)$. Następnie wygeneruj wektor zmiennej odpowiedzi zgodnie z modelem regresji logistycznej gdzie wektor współczynników wynosi $\beta = (c, c, c, c, c, 0, ..., 0)^T$, z $c = 9/\sqrt{n}$.

   a) Wyznacz estymatory współczynników regresji w modelu regresji logistycznej za pomocą metody największej wiarogodności, metody Firtha i funkcji *cv.glmnet*.

   b) Dla wszystkich metod wyznacz liczbę fałszywych i prawdziwych odkryć, FDP, $B(\hat{\beta})$, $SE(\hat{\beta})$ i $SE(\hat{p})$. W przypadku estymatora największej wiarogodności i estymatora Firtha identyfikacje odkryć przeprowadź wykonując testy na poziomie istotności 0.05.

   c) Wyznacz istotne zmienne za pomocą *SLOPE* i *mBIC2*. Wyznacz liczbę fałszywych i prawdziwych odkryć i FDP. Następnie wyznacz estymatory współczynników regresji logistycznej wykorzystując funkcję *glm* w oparciu o odpowiednio zredukowaną macierz X i wyznacz błędy estymacji $B(\hat{\beta})$, $SE(\hat{\beta})$ i $SE(\hat{p})$.

   d) Powtórz punkty a)-c) 100 razy i wyznacz i krytycznie pońownaj uśrednione wyniki różnych charakerystyk dla wszystkich metod.

5. Powtórz zadanie 4 w przypadku gdy wiersze macierzy X są niezależnymi wektorami losowymi z wielowymiarowego rozkładu normalnego $N(0, \Sigma)$ z macierzą kowariancji taką, że $\Sigma_{ii} = 1$ i $\Sigma_{ij} = 0.3$ dla $i \neq j$.

Malgorzata Bogdan