Problem doboru zmiennych w modelu liniowym

Kryteria oceny modelu

Kryterium AIC jest oparte na teorii informacji. Przypuśćmy że dane są generowane przez pewien proces (mechanizm) losowy f. Rozważmy dwa modele g_1 i g_2 mające reprezentować proces f. Gdybyśmy znali f, moglibyśmy obliczyć stratę informacji o f reprezentując f przez g_1 lub g_2 obliczając "odległość " Kullbacka-Leiblera

$$D_{KL}(f,g_i)=\int_{x\in X} f(x)\log\frac{f(x)}{g_i(x)}dx \quad i=1,2.$$

(ogólnie $D_{KL}(P,Q_i)=\int_{x\in X}\log\frac{P(dx)}{Q_i(dx)}P(dx),\ P<<Q_i$).

Z tych dwóch modeli wybieramy ten, który charakteryzuje się mniejszą stratą informacji. Niestety nie znamy modelu f. Jednak jak pokazał Akaike (1974) możemy poprzez jego kryterium AIC szacować, o ile więcej (lub mniej) informacji o f tracimy wybierając g_1 zamiast g_2. Tak więc ze skończonego zbioru modeli $g_i; i=1,...,M$ możemy wybrać ten, który charakteryzuje się najmniejszą stratą informacji o f. Niech $\Lambda=-2\ln L(\hat{\boldsymbol{\beta}},\hat{\sigma}^2)$ oznacza maksimum logarytmu funkcji wiarogodności a Φ liczbę wszystkich parametrów modelu. Wówczas

$$AIC=-2\ln L(\hat{\boldsymbol{\beta}},\hat{\sigma}^2)+2\Phi=n+n\log(2\pi)+n\log(\frac{RSS}{n})+2\Phi$$

W przypadku modelu liniowego z *p* zmiennymi objaśniającymi $\Phi=\dim\beta+1=p+1=k+2$, bo wymiar wektora $\boldsymbol{\beta}$ wynosi *p*=*k*+1 i jeszcze jeden parametr σ^2.

Oczywiście nie mamy żadnej gwarancji że najlepszy model z rozważanej klasy jest satysfakcjonujący.

Kryterium AIC ma ponadto charakter asymptotyczny i przy małej liczności zbioru obserwacji niezbędne są pewne korekty kryterium AIC. Zastępowane jest ono prze AICc, które ma różną postać dla różnych modeli. Generalnie AICc nakłada większą karę za dodatkowe parametry niż AIC. Podobne, lecz uzyskane w podejściu bayesowskim jest kryterium

$$BIC=-2\ln L(\hat{\boldsymbol{\beta}},\hat{\sigma}^2)+\ln(n)\ \Phi$$

W programie R używane są

- *GIC* (*Generalized Information Criterion*)

$$GIC=-2\ln L(\hat{\boldsymbol{\beta}},\hat{\sigma}^2)+h\cdot\Phi$$

gdzie h to pewien współczynnik, Φ -liczba parametrów w aktualnym modelu . Dwa specjalne przypadki *GIC* to kryterium Akaike (*AIC*) w którym $h=2$ oraz kryterium Schwartza w którym $h=\log(n)$ gdzie n oznacza liczbę obserwacji.

- Zmodyfikowany współczynnik R^2_{adj}
- Statystyka *Cp Mallowsa* $E\sum_{i=1}^{n}(y_i - E(y_i|\, X_i))^2 / \sigma^2$ szacowana jako

$$C_p(M) = \frac{RSS_p}{S^2} - n + 2p,$$

gdzie RSS_p to suma kwadratów reszt w modelu z p zmiennymi a S^2 to suma kwadratów reszt w modelu ze wszystkimi zmiennymi. Im mniejsza wartość C_p tym model lepszy. Statystykę C_p wyznacza funkcja ols_mallows_cp {olsrr}.

Idea konstrukcji wskaźnika C_p Mallowsa jest następująca. Mając próbę uczącą $(\mathbf{x}_i, y_i)\ i = 1,...,n$ oraz funkcję straty L możemy przyjąć kryterium jakości rozwiązania $\hat{f}$ rozważanego problemu

$$\frac{1}{n}\sum_{i=1}^{n} L(y_i, \hat{f}(\mathbf{x}_i)).$$

Takie przybliżenie ryzyka, a zatem prawdziwej mocy predykcyjnej rozwiązania $\hat{f}$ jest oczywiście przybliżeniem optymistycznym. Niech miarą optymizmu będzie wielkość

$$\text{op} = \frac{1}{n}\sum_{i=1}^{n} E_{y_i^*} L(y_i^*, \hat{f}(\mathbf{x}_i)) - \frac{1}{n}\sum_{i=1}^{n} E_{y_i} L(y_i, \hat{f}(\mathbf{x}_i)),$$

gdzie $y_i^*\ = i = 1,...,n$ oznaczają nowe obserwacje zmiennej odpowiedzi niezależne od y_i i odpowiadające tym samym punktom próby uczącej $\mathbf{x}_i$ traktowanym jako punkty ustalone. Model $\hat{y}_i = \hat{f}(\mathbf{x}_i)$ jest dopasowany do próby uczącej $(\mathbf{x}_i, y_i)\ i = 1,...,n$ i dlatego optymizm przyjmuje zwykle wartości dodatnie. Można udowodnić, że przy ogólnych założeniach i kwadratowej funkcji straty (możliwe są też inne postacie funkcji straty)

$$\text{op} = \frac{2}{n}\sum_{i=1}^{n} Cov(\hat{y}_i, y_i).$$

Łatwo zauważyć, że im bardziej model jest dopasowany do próby uczącej tym większy będzie optymizm. Jeżeli prawdziwy model jest postaci $y = f(\mathbf{x}_i) + \varepsilon$ gdzie funkcję f można zapisać jako liniową kombinację p znanych funkcji bazowych oraz ε jest błędem losowym o zerowej wartości oczekiwanej i wariancji σ_ε^2, to $\sum_{i=1}^{n} Cov(\hat{y}_i, y_i) = p\sigma_\varepsilon^2$. Ostatecznie zamiast minimalizować

$\frac{1}{n}\sum_{i=1}^{n} L(y_i, \hat{f}(\mathbf{x}_i))$ bardziej uzasadniona jest minimalizacja wskaźnika

$$\frac{1}{n}\sum_{i=1}^{n} L(y_i, \hat{f}(\mathbf{x}_i)) + \frac{2}{n} p\sigma_\varepsilon^2,$$

znanego jako wskaźnik C_p Mallowsa. W języku R używana jest równoważna postać

$$C_p = \frac{RSS_p}{S^2} - n + 2p$$,t

RSS_p= deviance() model z p zmiennymi objaśniającymi
s<-summary_fullmodel$sigma #ocena sigma dla pełnego modelu dostępna dla obiektu klasy summary.lm

Uzupełnienia dotyczące kryteriów wyboru najlepszego modeli można znaleźć w pliku AIC BIC.pdf dostępnego w MS Teams

Wybór najlepszego modelu $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \varepsilon$

Algorytm A (Best model)- przeszukujemy wszystkie 2^k modeli

1. Niech M_0 oznacza model zerowy, który nie zawiera predyktorów. Ten model po prostu przewiduje średnią próbki dla każdej obserwacji.
2. Dla i=1,…,k
   a) Dopasować wszystkie $\binom{k}{i}$ modeli z i predyktorami
   b) Wybierz najlepszy spośród tych $\binom{k}{i}$modeli i nazwij go M_i. Za najlepszy uważany jest tu model z najmniejszym RSS lub równoważnie największym R^2.
3. Wybierz jeden najlepszy model spośród M_0, . . . ,M_k przy użyciu zweryfikowanego krzyżowo błędu przewidywania, C_p (AIC), BIC lub skorygowanego R^2.

Krok 2 algorytmu A redukuje problem wyboru najlepszego modelu spośród 2^k do wyboru jednego z k+1 modeli.

W kroku 3 wybieramy najlepszy model spośród k+1 modeli stosując jedno z kryteriów CV, AIC, BIC, C_p i R^2_{adj} .

Algorytm B -Regresja krokowa forward

1. Niech M_0 oznacza model zerowy, który nie zawiera predyktorów.
2. Dla i=0,…,k-1
   a) Rozważyć wszystkie k-i modeli które dodają jeden predyktor do M_i
   b) Wybierz najlepszy spośród tych k-i modeli i nazwij go M_{i+1}. Za najlepszy uważany jest tu model z najmniejszym RSS lub równoważnie największym R^2.
3. Wybierz jeden najlepszy model spośród M_0, . . . ,M_k przy użyciu zweryfikowanego krzyżowo błędu przewidywania, Cp , AIC, BIC lub skorygowanego R^2.

Algorytm C- Regresja krokowa backward

1. Niech M_k oznacza model pełny, który zawiera wszystkie predyktory.
2. Dla $i=k,k-1,\ldots, 1$
   a) Rozważyć wszystkie k modeli które zawierają wszystkie oprócz jednego predyktora M_i dla wszystkich i-1 predyktorów
   b) Wybierz najlepszy spośród tych i modeli i nazwij go M_{i-1}. Za najlepszy uważany jest tu model z najmniejszym RSS lub równoważnie największym R^2.
3. Wybierz jeden najlepszy model spośród $M_0, \ldots, Mk_p$ przy użyciu zweryfikowanego krzyżowo błędu przewidywania, Cp , AIC, BIC lub skorygowanego R^2.

Przykłady- Ćwiczenie 8 Best Model