Metody regresji liniowej

K.Melka

Wprowadzenie

Metoda najmniejszych kwadratów

Przykład

Błąd średniokwadratowy

Selekcja zmiennych

Regresja grzbietowa (Ridge regression)

LASSO

Porównanie i możliwości rozszerzenia

Metody regresji liniowej

Melka Kamil

Uniwersytet Wrocławski

14.03.2022

Plan prezentacji

1 Wprowadzenie

2 Metoda najmniejszych kwadratów

3 Przykład

4 Błąd średniokwadratowy

5 Selekcja zmiennych

6 Regresja grzbietowa (Ridge regression)

7 LASSO

8 Porównanie i możliwości rozszerzenia

Model regresji liniowej

- Model regresji liniowej zakłada, że funkcja regresji $E(Y|X)$ jest liniowa względem zmiennych objaśniających $X_1, \ldots, X_p$.
- $f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j$.
- Zazwyczaj nieskomplikowany sposób liczenia i interpretacji otrzymanych parametrów.
- Dobre własności predykcyjne w przypadku małej liczby danych treningowych, niskiej proporcji sygnału do szumu lub rzadkich danych.

Rodzaj zmiennych objaśniających

Mimo, że model jest liniowy, to możemy dokonywać pewnych przekształceń na zmiennych objaśniających, które niekoniecznie są liniowe.

- Zmienne ilościowe.
- Przekształcenia typu: logarytm, pierwiastek kwadratowy lub kwadrat zmiennej.
- Rozszerzanie bazy, np. $X_2 = X_1^2$, $X_3 = X_1^3$ odpowiada za przedstawienie modelu w postaci wielomianowej.
- Zmienne jakościowe przedstawiane za pomocą "dummy variables".
- Interakcje między zmiennymi, np. $X_3 = X_1 \cdot X_2$.

Metoda najmniejszych kwadratów

Celem metody najmniejszych kwadratów jest dobranie takich wartości β_j, które minimalizują

$$RSS(\beta) = \sum_{i=1}^{N} (y_i - f(x_i))^2 = \sum_{i=1}^{N} \Big(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\Big)^2.$$

Ze statystycznego punktu widzenia jest sensowne, gdy wektory obserwacji (x_i, y_i) są niezależnymi losowaniami z populacji lub gdy y_i jest warunkowo niezależne względem znanych wartości x_i.

Geometria metody najmniejszych kwadratów

Metody regresji liniowej

K.Melka

Wprowadzenie

Metoda najmniejszych kwadratów

Przykład

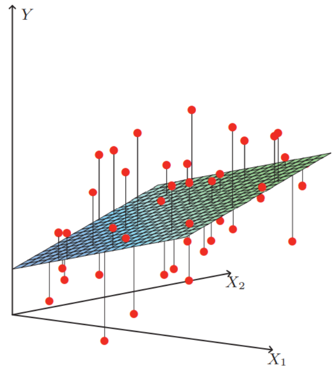
Błąd średniokwadratowy

Selekcja zmiennych

Regresja grzbietowa (Ridge regression)

LASSO

Porównanie i możliwości rozszerzenia



Rysunek: Regresja liniowa za pomocą metody najmniejszych kwadratów dla $X \in \mathbb{R}^2$, źródło: *Elements of statistical learning, fig. 3.1*

Postać estymatora $\hat{\beta}$

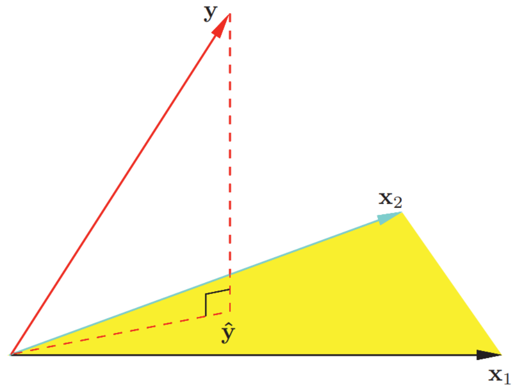
Niech $\mathbf{X}$ oznacza macierz rozmiaru $N \times (p+1)$, w której mamy wartości zmiennych objaśniających.
Niech $\mathbf{y}$ oznacza wektor wartości zmiennej objaśnianej długości N.
Wtedy:

- $RSS(\beta) = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta)$.
- $\frac{\partial RSS}{\partial \beta} = -2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta)$.
- $\frac{\partial^2 RSS}{\partial \beta \partial \beta^T} = 2\mathbf{X}^T\mathbf{X}$.
- $\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta) = 0$.
- $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$.

Geometria metody najmniejszych kwadratów

- $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$
- $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$
- Wektory $\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_p$ rozpinają podprzestrzeń przestrzeni $\mathbb{R}^N$.
- Metoda najmniejszych kwadratów minimalizuje $RSS(\beta) = ||\mathbf{y} - \mathbf{X}\beta||^2$.
- Zatem optymalne $\hat{\beta}$ sprawia, że wektor residuów $\mathbf{y} - \hat{\mathbf{y}}$ jest ortogonalny do wyżej wspomnianej podprzestrzeni.

Geometria metody najmniejszych kwadratów



Rysunek: Rysunek pokazujący ortogonalność wektora residuów do przestrzeni rozpiętej przez zmienne objaśniające, źródło: *Elements of statistical learning, fig. 3.2*

Własności estymatorów

Niech obserwacje y_i będą nieskorelowane o tej samej wariancji σ^2. Niech x_i będą wyznaczone. Wtedy macierz kowariancji wektora $\hat{\beta}$ ma postać:

$$\mathrm{Var}(\hat{\beta}) = (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2$$

$$\hat{\sigma}^2 = \frac{1}{N-p-1}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

Bardzo często zakładamy, że możemy zapisać zmienne objaśniające w postaci: $Y = \beta_0 + \sum_{j=1}^{p} X_j\beta_j + \varepsilon$, gdzie $\varepsilon \sim N(0, \sigma^2)$

Własności estymatorów

- $\hat{\beta} \sim N(\beta, (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2)$
- $(N-p-1)\hat{\sigma}^2 \sim \sigma^2\chi^2_{N-p-1}$
- $\hat{\beta}$ i $\hat{\sigma}^2$ są niezależnie statystycznie.
- $z_j = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{v_j}}$, gdzie v_j jest $j-$tym elementem na diagonali $(\mathbf{X}^T\mathbf{X})^{-1}$
- Przy $H_0 : \beta_j = 0$, z_j ma rozkład t_{N-p-1}.

Ogon rozkładów

Metody regresji liniowej

K.Melka

Wprowadzenie

Metoda najmniejszych kwadratów

Przykład

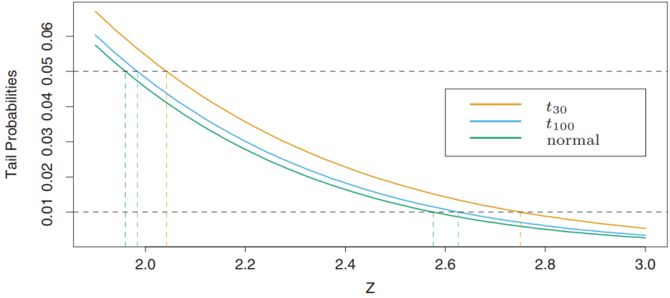
Błąd średniokwadratowy

Selekcja zmiennych

Regresja grzbietowa (Ridge regression)

LASSO

Porównanie i możliwości rozszerzenia



Rysunek: Ogony rozkładów t_{30}, t_{100} i $N(0,1)$, źródło: *Elements of statistical learning, fig. 3.3*

Testowanie grupy współczynników

$$F = \frac{(RSS_0 - RSS_1)/(p_1 - p_0)}{RSS_1/(N - p_1 - 1)},$$

gdzie RSS_1 to RSS dla modelu z $p_1 + 1$ parametrami, a RSS_0 to RSS dla zagnieżdżonego modelu z $p_0 + 1$ parametrami. Przy założeniu normalności błędów oraz przy H_0, że mniejszy model jest poprawny, to F ma rozkład $F_{p_1 - p_0, N - p_1 - 1}$

Przedział ufności dla β_j

$$(\hat{\beta}_j - z^{(1-\alpha)}\sqrt{v_j}\hat{\sigma}, \hat{\beta}_j + z^{(1-\alpha)}\sqrt{v_j}\hat{\sigma})$$

Macierz korelacji między zmiennymi objaśniającymi

Metody regresji liniowej

K.Melka

Wprowadzenie

Metoda najmniejszych kwadratów

Przykład

Błąd średniokwadratowy

Selekcja zmiennych

Regresja grzbietowa (Ridge regression)

LASSO

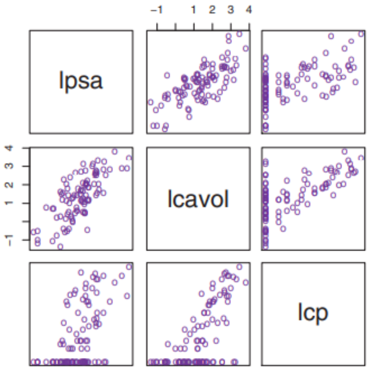
Porównanie i możliwości rozszerzenia

TABLE 3.1. *Correlations of predictors in the prostate cancer data.*

	lcavol	lweight	age	lbph	svi	lcp	gleason
lweight	0.300						
age	0.286	0.317					
lbph	0.063	0.437	0.287				
svi	0.593	0.181	0.129	−0.139			
lcp	0.692	0.157	0.173	−0.089	0.671		
gleason	0.426	0.024	0.366	0.033	0.307	0.476	
pgg45	0.483	0.074	0.276	−0.030	0.481	0.663	0.757

źródło: *Elements of statistical learning, tab. 3.1*

Scatterplot



Rysunek: Scatterplot między zmienną objaśnianą, a dwoma zmiennymi objaśniającymi, źródło: *Elements of statistical learning, fig. 1.1*

Parametry modelu liniowego

TABLE 3.2. *Linear model fit to the prostate cancer data. The Z score is the coefficient divided by its standard error (3.12). Roughly a Z score larger than two in absolute value is significantly nonzero at the $p = 0.05$ level.*

Term	Coefficient	Std. Error	Z Score
Intercept	2.46	0.09	27.60
lcavol	0.68	0.13	5.37
lweight	0.26	0.10	2.75
age	−0.14	0.10	−1.40
lbph	0.21	0.10	2.06
svi	0.31	0.12	2.47
lcp	−0.29	0.15	−1.87
gleason	−0.02	0.15	−0.15
pgg45	0.27	0.15	1.74

źródło: *Elements of statistical learning, tab. 3.2*

F-test

Przeprowadzamy F-test, aby sprawdzić, czy możemy wyrzucić z modelu nieistotne zmienne.

$$F = \frac{(32.81 - 29.43)/(9-5)}{29.43/(67-9)} = 1.67$$

$$\Pr(F_{4,58} > 1.67) = 0.17$$

Średni błąd predykcji

Średni błąd predykcji na zbiorze testowym wynosi 0.521. Z kolei, dla predykcji opartej na użyciu średniej wartości zmiennej objaśnianej ze zbioru treningowego, otrzymaliśmy 1.057. Oznacza to, że nasz model zredukował bazowy współczynnik błędu o około 50%.

Błąd średniokwadratowy

Metoda najmniejszych kwadratów estymuje parametry β z najmniejszą możliwą wariancją z pośród wszystkich nieobciążonych liniowych estymatorów.

- $MSE(\hat{\beta}) = E(\hat{\beta} - \beta)^2 = Var(\hat{\beta}) + [E(\hat{\beta}) - \beta]^2$

Widać jednak, że może istnieć obciążony estymator z mniejszym błędem średniokwadratowym.
Jakakolwiek metoda, która przesuwa współczynniki otrzymane z metody najmniejszych kwadratów bliżej 0, może w rezultacie być obciążonym estymatorem.
Błąd średniokwadratowy jest ściśle powiązany z oczekiwanym błędem predykcji:

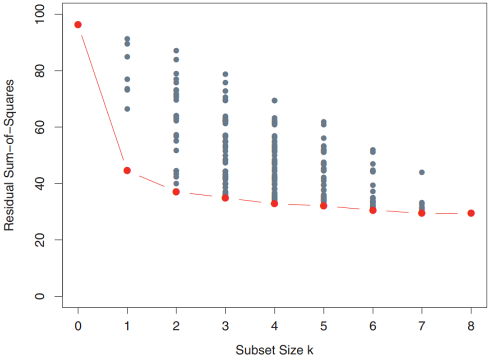
- $E(Y_0 - \hat{f}(x_0))^2 = \sigma^2 + E(x_0^T \hat{\beta} - f(x_0))^2 = \sigma^2 + MSE(\hat{f}(x_0))$.

Po co nam w ogóle selekcja zmiennych?

- Skuteczność predykcji: pozbywając się niektórych zmiennych (czyli zmieniając jej współczynniki na 0) możemy zwiększyć obciążenie, zmniejszając przy tym wariancję.
- Interpretacja: mniejsza liczba zmiennych pozwala nam na łatwiejszą i trafniejszą interpretację modelu.

Best-Subset Selection

Dla każdego $k \in \{0, 1, 2, \ldots, p\}$ znajdujemy podzbiór, dla którego otrzymujemy najmniejszą wartość RSS.

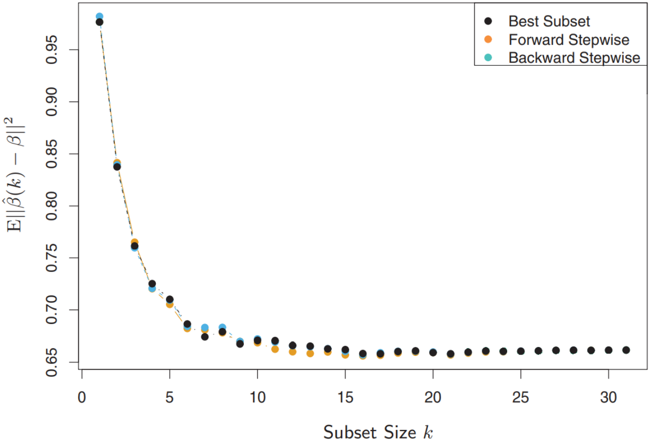


Rysunek: Best-Subset Selection dla danych o raku prostaty, źródło: *Elements of statistical learning, fig. 3.5*

Forward- i Backward-Stepwise Selection

- W Forward-Stepwise Selection startujemy z modelu z interceptem i z każdym krokiem dodajemy zmienną, która najmocniej poprawia dopasowanie.
- W Backward-Stepwise Selection startujemy z pełnego modelu i z każdym krokiem odejmujemy zmienną, która ma najmniejszy wpływ na model.

Porównanie metod



Rysunek: $Y = X^T\beta + \varepsilon$; $N = 300$; $p = 31$; $\beta_1, \ldots, \beta_{10} \sim N(0, 0.4)$, pozostałe równe 0; $\varepsilon \sim N(0, 6.25)$ źródło: *Elements of statistical learning, fig. 3.6*

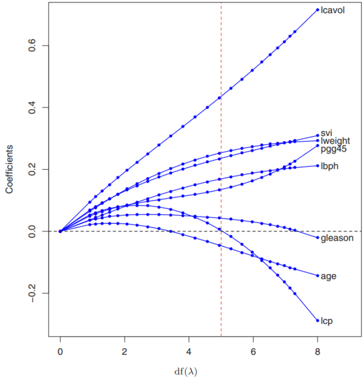
Ridge regression

- $\hat{\beta}^R = \text{argmin}_\beta \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}.$
- $\lambda \geq 0$ - parametr regularyzacyjny: im większa wartość, tym większy stopień obciążenia współczynników w stronę 0.
- $\hat{\beta}^R = \text{argmin}_\beta \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2$, pod warunkiem $\sum_{j=1}^{p} \beta_j^2 \leq t$.

Postać macierzowa

- $RSS(\lambda) = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta) + \lambda\beta^T\beta.$
- $\hat{\beta}^R = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$
- W przypadku ortonormalnym $\hat{\beta}^R = \hat{\beta}^{LS}/(1+\lambda).$

Ridge regression

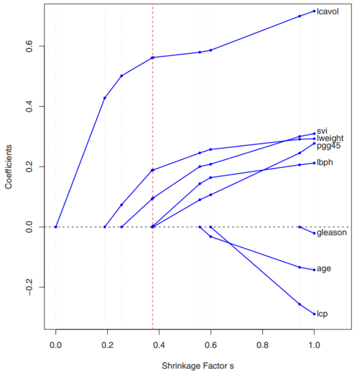


Rysunek: Współczynniki obliczone przy pomocy regresji grzbietowej w zależności od wartości λ; czerwoną linią zaznaczono wartość optymalną wyznaczoną przez kroswalidację, źródło: *Elements of statistical learning, fig. 3.8*

LASSO

- $\hat{\beta}^L = \text{argmin} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2$, pod warunkiem $\sum_{j=1}^{p} |\beta_j| \leq t$
- $\hat{\beta}^L = \text{argmin} \left\{ \frac{1}{2} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}.$
- Gdy $t > \sum_{j=1}^{p} |\hat{\beta}_j^{LS}|$, to $\hat{\beta}_j^L = \hat{\beta}_j^{LS}$ dla każdego j.

LASSO

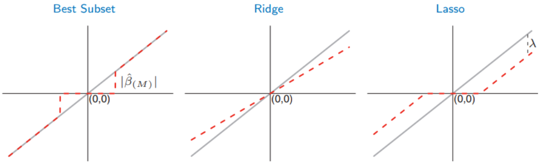


Rysunek: Współczynniki obliczone przy pomocy LASSO w zależności od wartości $s = t / \sum_{j=1}^{p} \hat{\beta}_j$; czerwoną linią zaznaczono wartość optymalną wyznaczoną przez kroswalidację, źródło: *Elements of statistical learning, fig. 3.10*

Estymatory β w zależności od wartości $\hat{\beta}^{LS}$

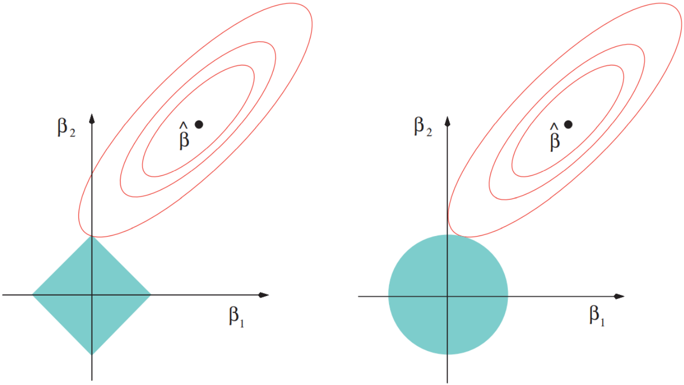
TABLE 3.4. *Estimators of β_j in the case of orthonormal columns of* $\mathbf{X}$. *M and λ are constants chosen by the corresponding techniques;* sign *denotes the sign of its argument (± 1), and x_+ denotes "positive part" of x. Below the table, estimators are shown by broken red lines. The $45°$ line in gray shows the unrestricted estimate for reference.*

Estimator	Formula
Best subset (size M)	$\hat{\beta}_j \cdot I(\lvert\hat{\beta}_j\rvert \geq \lvert\hat{\beta}_{(M)}\rvert)$
Ridge	$\hat{\beta}_j/(1+\lambda)$
Lasso	$\mathrm{sign}(\hat{\beta}_j)(\lvert\hat{\beta}_j\rvert - \lambda)_+$



źródło: *Elements of statistical learning, tab. 3.4*

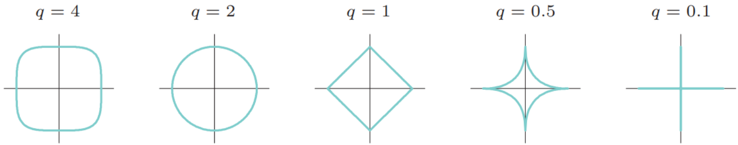
Geometria LASSO i regresji grzbietowej



Rysunek: Zaznaczone obszary to kolejno $|\beta_1| + |\beta_2| \leq t$ i $\beta_1^2 + \beta_2^2 \leq t^2$; czerwone elipsy to kontury funkcji błędu względem metody najmniejszych kwadratów, źródło: *Elements of statistical learning, fig. 3.11*

Generalizacja regresji grzbietowej i LASSO

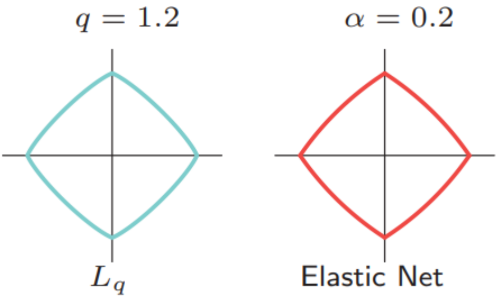
$$\hat{\beta} = \text{argmin}_{\beta}\left\{\sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|^q\right\}, q \geq 0$$



Rysunek: Kontury stałej wartości $\sum_j |\beta_j|^q$ dla danych wartości q, źródło: *Elements of statistical learning, fig. 3.12*

Elastic net

$$\lambda \sum_{j=1}^{p} \left(\alpha \beta_j^2 + (1-\alpha)|\beta_j| \right)$$



Rysunek: Porównanie kary elastic net i $\sum_j |\beta_j|^q$ dla $q = 1.2$, źródło: *Elements of statistical learning, fig. 3.13*

Porównanie wyników dla danych o raku prostaty

TABLE 3.3. *Estimated coefficients and test error results, for different subset and shrinkage methods applied to the prostate data. The blank entries correspond to variables omitted.*

Term	LS	Best Subset	Ridge	Lasso
Intercept	2.465	2.477	2.452	2.468
lcavol	0.680	0.740	0.420	0.533
lweight	0.263	0.316	0.238	0.169
age	−0.141		−0.046	
lbph	0.210		0.162	0.002
svi	0.305		0.227	0.094
lcp	−0.288		0.000	
gleason	−0.021		0.040	
pgg45	0.267		0.133	
Test Error	0.521	0.492	0.492	0.479
Std Error	0.179	0.143	0.165	0.164

źródło: *Elements of statistical learning, tab. 3.3*

Wprowadzenie

Metoda najmniejszych kwadratów

Przykład

Błąd średniokwadratowy

Selekcja zmiennych

Regresja grzbietowa (Ridge regression)

LASSO

Porównanie i możliwości rozszerzenia

Dziękuję za uwagę.