Model liniowy

wybór podzbioru zmiennych

Agata Cieślik

24 marca 2021 r.

Wprowadzenie

- T. Hastie, R. Tibshirani, J. Friedman, *The elements of statistical learning*
- Rozdział 3: *Linear Methods for Regression*
- W szczególności: podrozdział 3.3 *Subset selection*

Regresja liniowa

Zakładamy, że $E(Y|X)$ jest funkcją liniową znanych zmiennych X (*zmiennych objaśniających*).

$$E(Y|X) = X\beta$$

$$E(Y|X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j$$

Model liniowy

$$Y = X\beta + \varepsilon, \varepsilon \sim N(0, \sigma^2)$$

- przy uwzględnieniu możliwych przekształceń zmiennych może obrazować bardzo szeroki zakres zależności - zamiast każdej rozpatrywanej zmiennej X_i możemy uwzgędnić jej przekształcenie: X_i^2, $\log(X_i)$...
- łatwy w interpretacji
- sprawdza się na niewielkich zbiorach danych, w przypadku danych rzadkich (*sparse data*) oraz w przypadku występowania wyjątkowo dużego szumu (*low signal-to-noise ratio*)

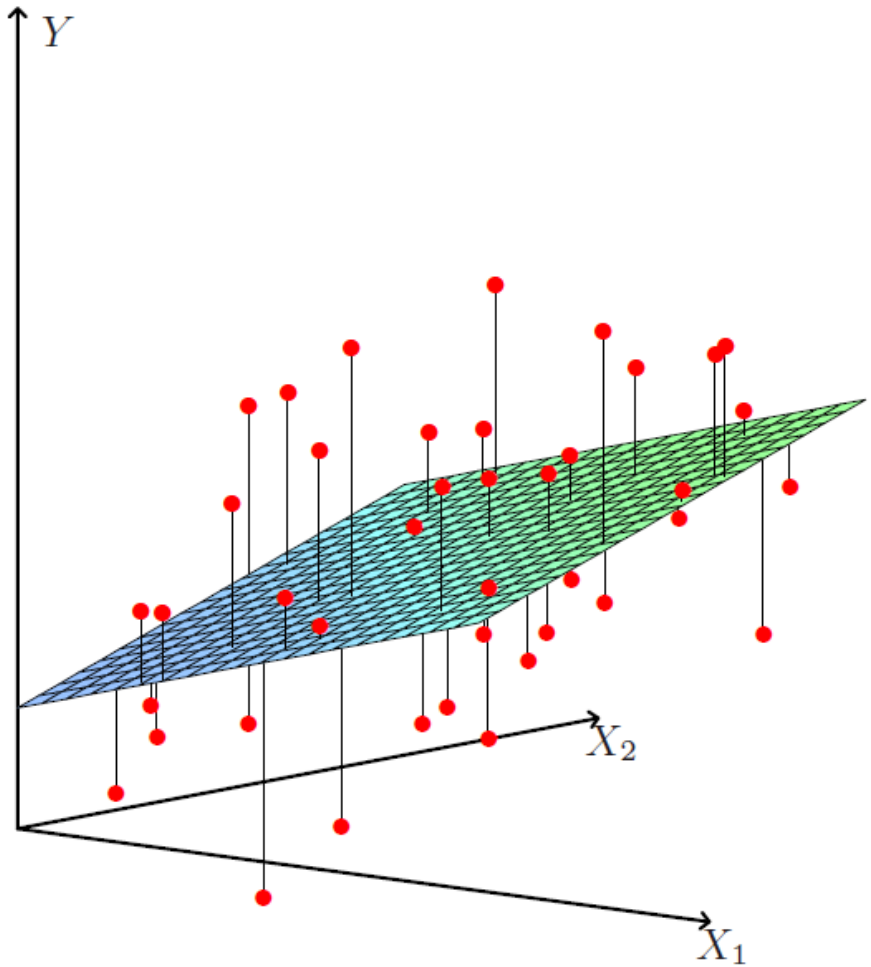
Model liniowy: estymacja

$$Y = X\beta + \varepsilon$$

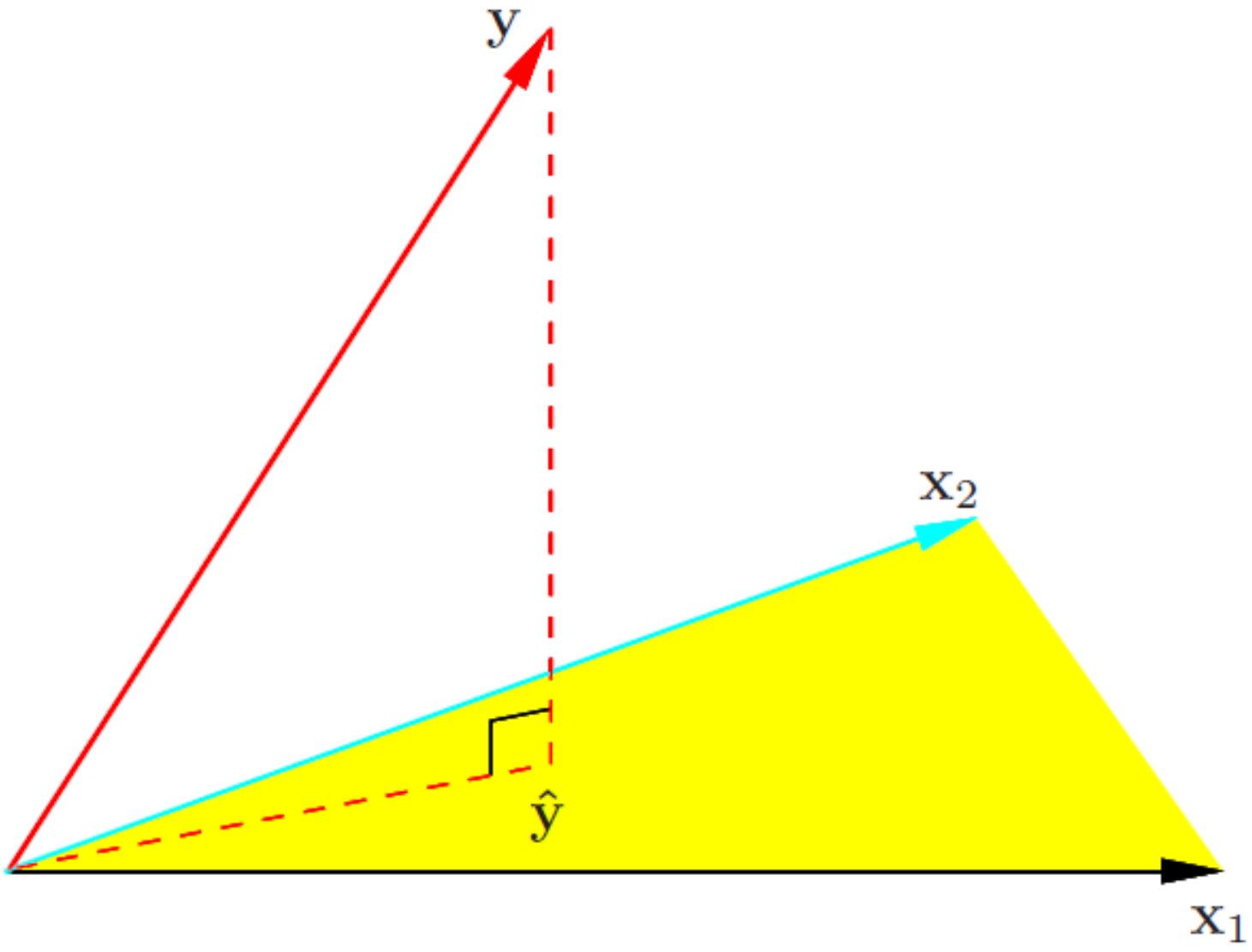
$$\hat{Y} = X\hat{\beta}$$

Metoda najmniejszych kwadratów: minimalizacja *RSS*

$$RSS(\beta) = \sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 = ||Y - X\beta||^2$$



Metoda najmniejszych kwadratów: rzut ortogonalny



Metoda najmniejszych kwadratów: estymator

Dla:

- (X_i, Y_i) - niezależnych, wylosowanych z populacji (lub, co najmniej, Y_i warunkowo niezależnych przy danych X_i)
- X - macierzy pełnego rzędu kolumn (stąd $X^T X$ jest dodatnio określona)

Mamy jedyne rozwiązanie:

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

Stąd predykcje:

$$\hat{Y} = X\hat{\beta} = X(X^T X)^{-1} X^T Y$$

gdzie $H = X(X^T X)^{-1} X^T$ tzw. *hat matrix* jest macierzą rzutu.

Przykład: dane dotyczące rozwoju raka prostaty

Predyktory:

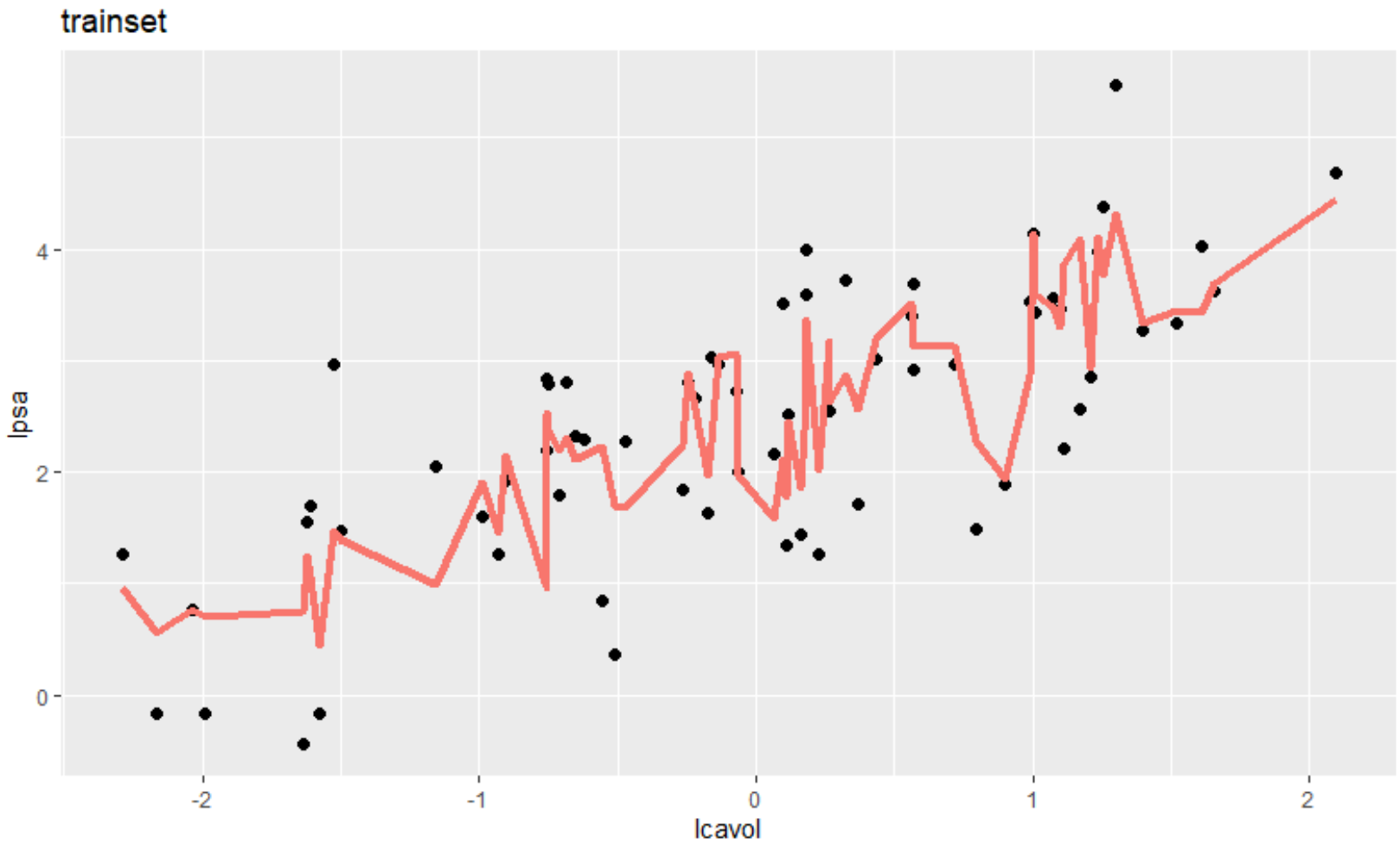
- **lcavol** - logarytm objętości raka (guza)
- **lweight** - logarytm wagi prostaty
- **age** - wiek
- **lbph**, **svi**, **lcp**, **gleason**, **pgg45** - wyniki różnych badań powiązanych z rozwojem choroby

Zmienną objaśnianą jest **lpsa** - poziom antygenu specyficznego dla prostaty

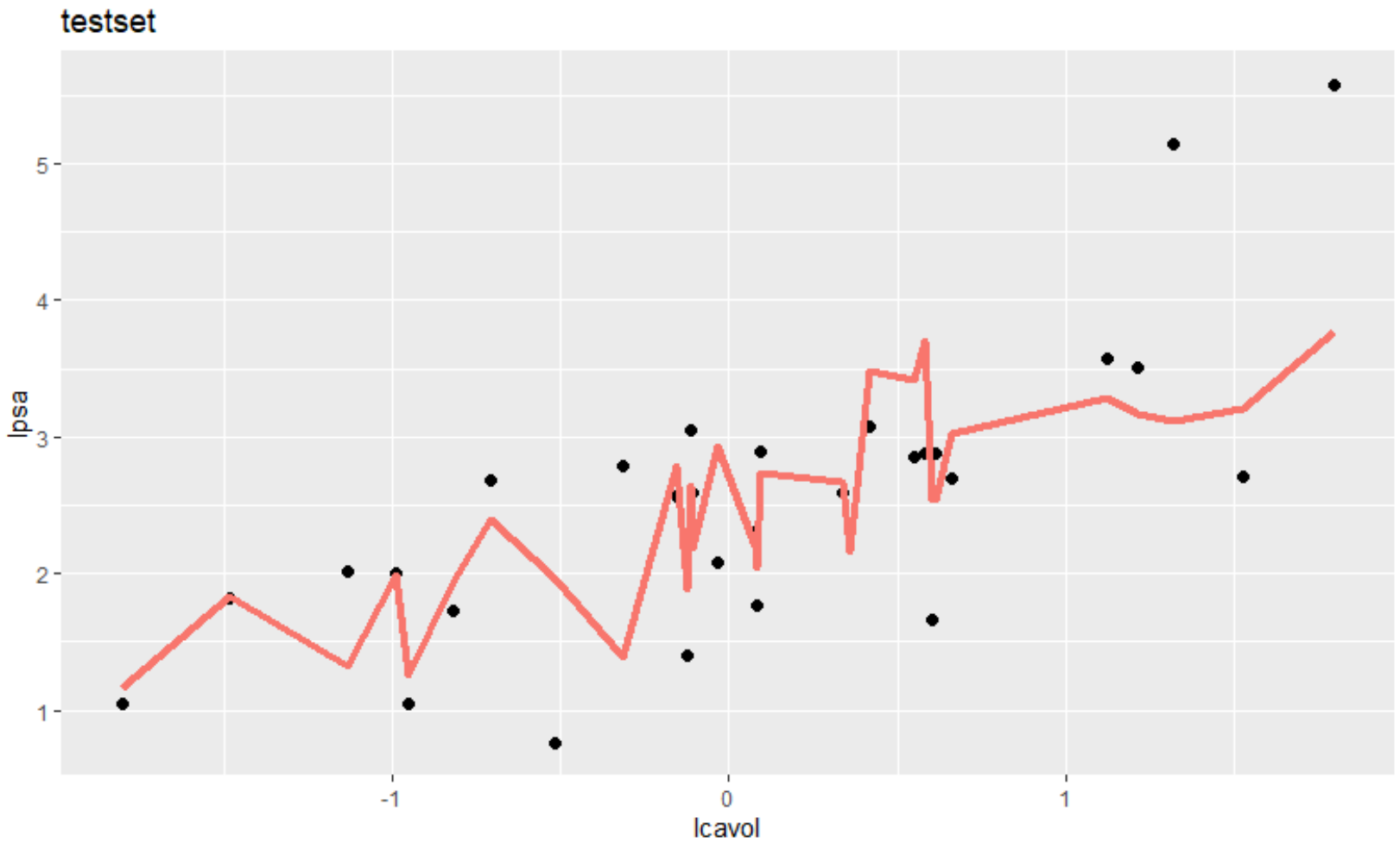
Przykład: dane dotyczące rozwoju raka prostaty

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

Przykład: dane dotyczące rozwoju raka prostaty



Przykład: dane dotyczące rozwoju raka prostaty



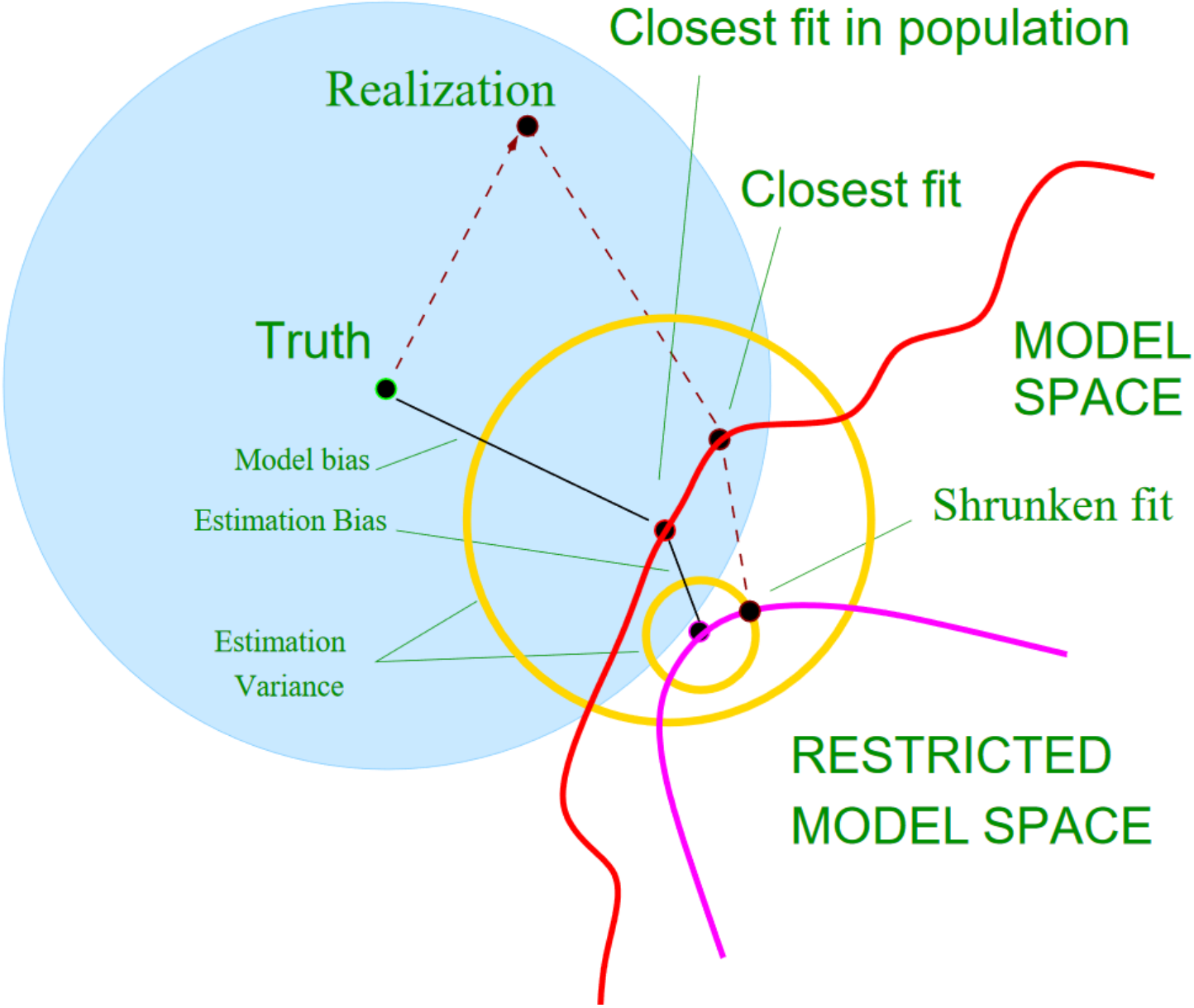
Estymator OLS: obciążenie i wariancja

Twierdzenie Gaussa - Markova mówi, że estymator OLS ma najmniejszą wariancję spośród estymatorów liniowych nieobciążonych. Na podstawie poniższej równości:

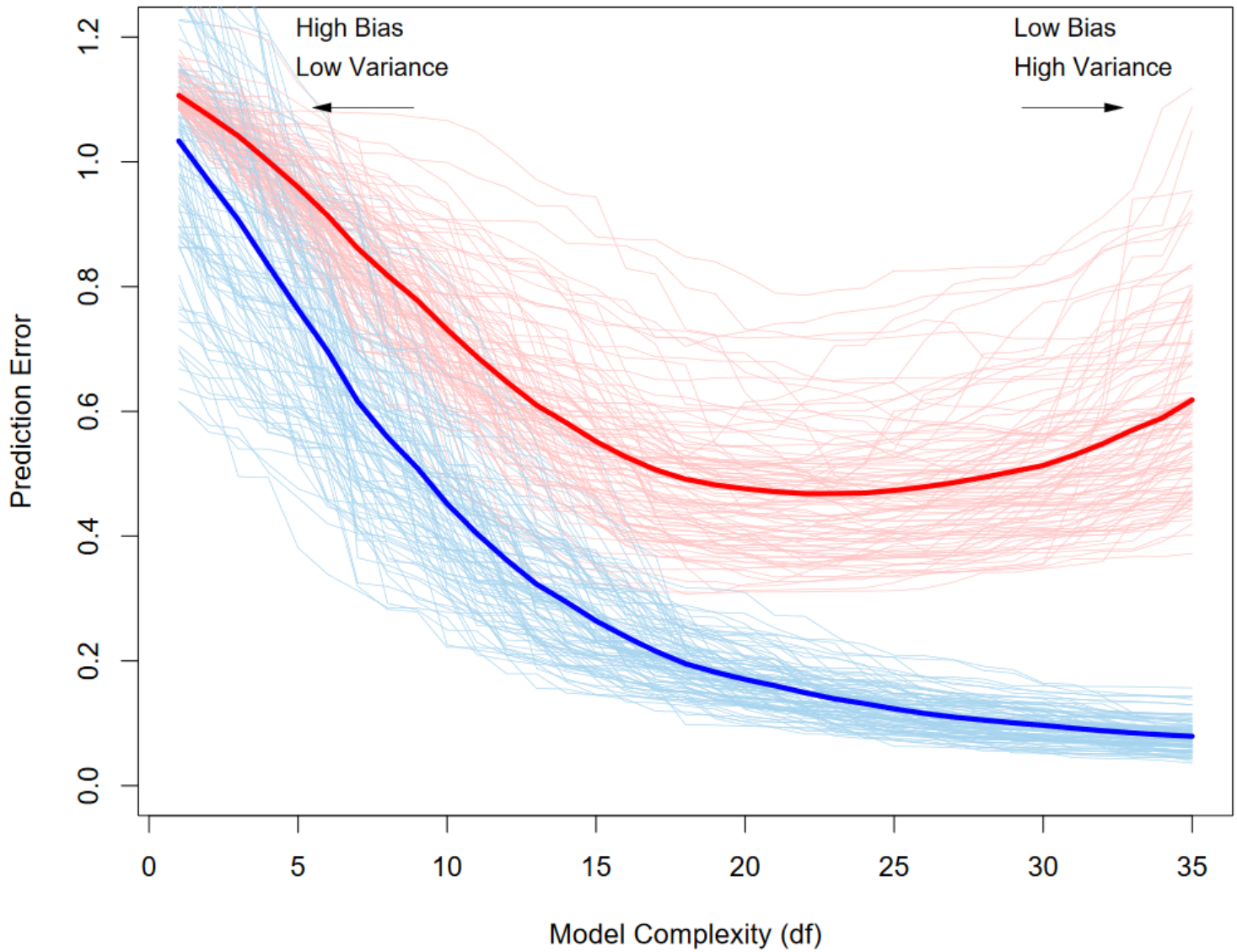
$$MSE(\hat{\beta}) = E(\hat{\beta} - \beta)^2 = Var(\hat{\beta}) + bias(\hat{\beta})^2$$

wnioskujemy, że najmniejsza wariancja implikuje również najmniejszy błąd w klasie modeli nieobciążonych.

Obciążenie, wariancja i złożoność modelu



Obciążenie, wariancja i złożoność modelu



Metody redukcji wariancji

Autorzy *Elements of Statistical Learning* wyróżniają dwa główne sposoby redukcji wariancji:

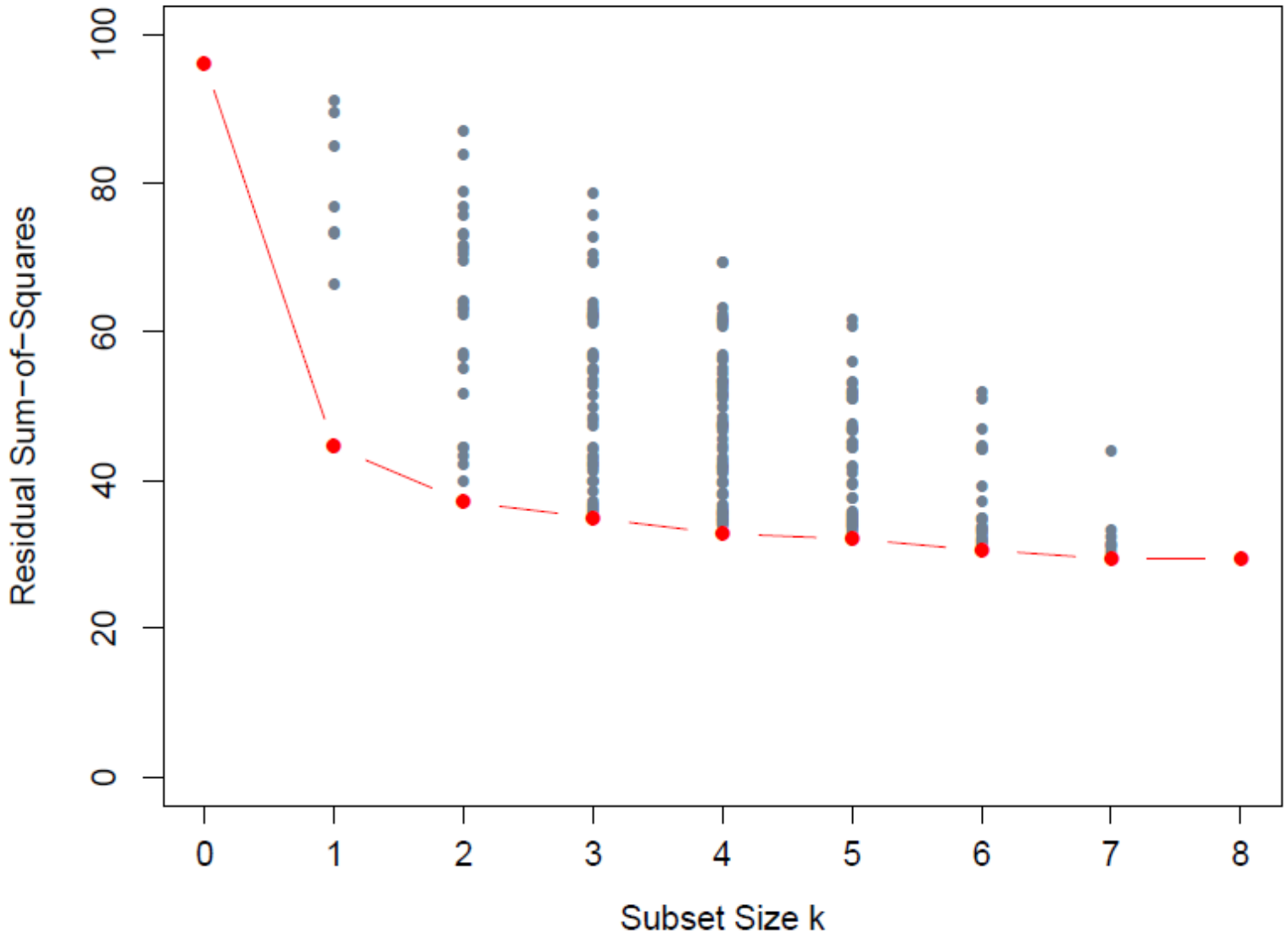
- użycie podzbioru zmiennych (*subset selection*, ograniczenie ich liczby to niezbędnego minimum
- metody regularyzacyjne (*shrinkage methods*) typu LASSO

Warto podkreślić, że wyznaczenie podzbioru zmiennych często nie jest równoważne wyłonieniu zmiennych istotnych - zmienne istotne, ale mające stosunkowo nieduży wpływ na predykcję mogą zostać usunięte w celu zbicia liczby estymowanych parametrów i, co za tym idzie, wariancji.

Best subset

Najbardziej podstawową, ale, w większości rzeczywistych przypadków, mało wydajną metodą wyboru najlepszego podzbioru zmiennych jest skonstruowanie modeli dla wszystkich możliwych kombinacji, porównanie ich dopasowania i wybranie spośród tych kombinacji optymalnej.

Przykład: ponownie rak prostaty



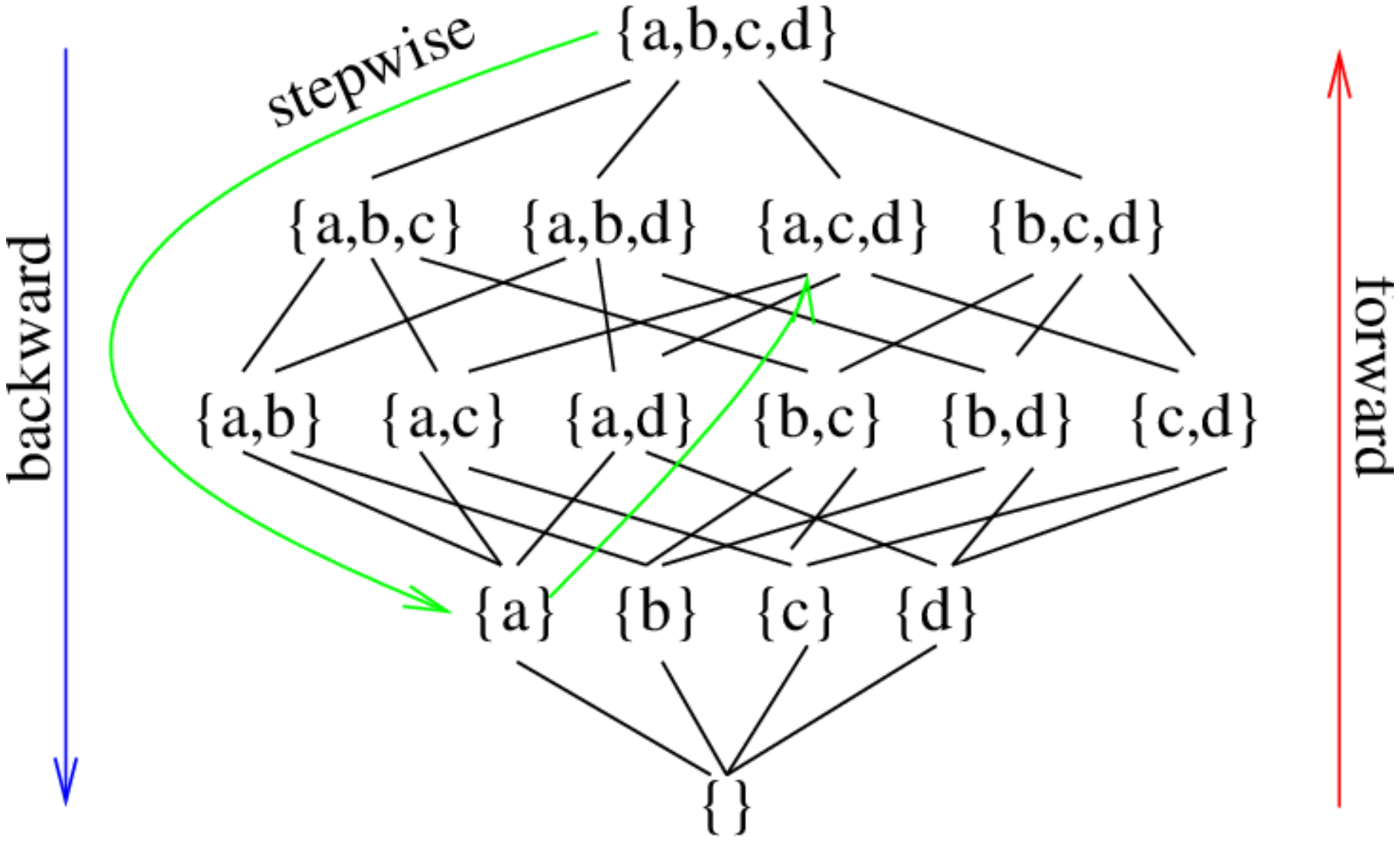
Forward stepwise (forward selection)

1. Zaczynamy od modelu zerowego z samym wyrazem wolnym (*interceptem*)
2. W każdym kroku sprawdzamy efekt **dodania** każdej z pozostałych zmiennych i dodajemy jedną z nich - tą, która w największym stopniu poprawia dopasowanie modelu
3. Procedura kończy się w momencie, gdy dodanie żadnej z kolejnych zmiennych nie poprawia znacząco dopasowania modelu
4. Efektem końcowym jest ciąg modeli opartych na zagnieżdżonych podzbiorach zmiennych rozmiaru od 1 do k

Backward stepwise (backward selection/elimination)

1. Zaczynamy od modelu pełnego z włączonymi wszystkimi
2. W każdym kroku sprawdzamy efekt **usunięcia** każdej z pozostałych w modelu zmiennych i usuwamy jedną z nich - tą, która w namniejszym stopniu pogarsza dopasowanie modelu
3. Procedura kończy się w momencie, gdy usunięcie którejkolwiek ze zmiennych pozostałych w modelu pogarsza znacząco dopasowanie
4. Efektem końcowym jest ciąg modeli opartych na zagnieżdżonych podzbiorach zmiennych rozmiaru od k do 1

Forward, backward, stepwise..

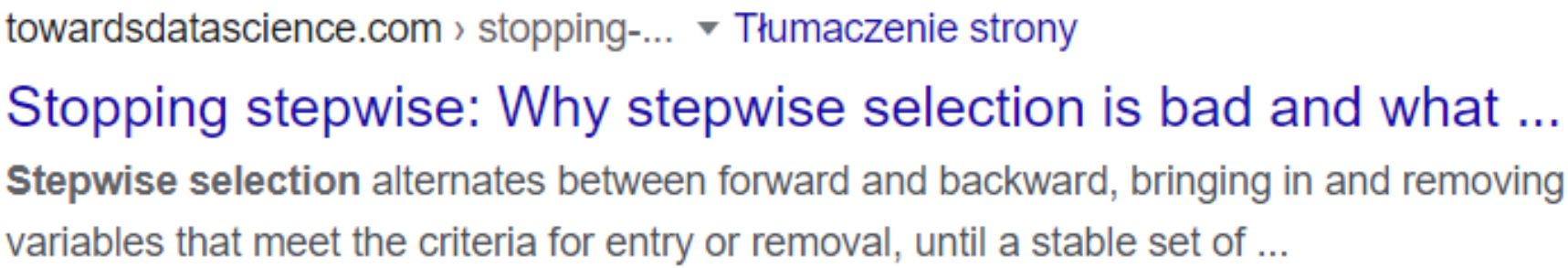


Kryteria w metodach typu *stepwise*

- test F (porównujący dopasowanie modelu i modelu w nim zagnieżdżonego), test t (istotności zmiennej w modelu)
- kryteria informacyjne: AIC, BIC...
- R^2, C_p
- *FDR* - false discovery rate

Krytyka

- wielokrotnie wykonywane testy służące za kryterium eliminacji lub dodawania są obciążone i stosowanie ich jako obiektywnej miary oceny bez uwzględnienia kontekstu modelu według niektórych jest nadużyciem (zjawisko tzw. *data dredging*)
- modele uzyskane procedurami typu *stepwise* mają tendencję do bycia nadmiernie uproszczonymi



Forward stagewise

Forward stagewise prezentuje zupełnie inne podejście niż metody typu stepwise - nie wybiera między najlepszymi dopasowaniami, nie korzysta z estymacji OLS. Jest metodą iteracyjną, w której stopniowo zwiększamy współczynniki zmiennych najbardziej skorelowanych z aktualnym wektorem wartości resztowych (*residuals*).

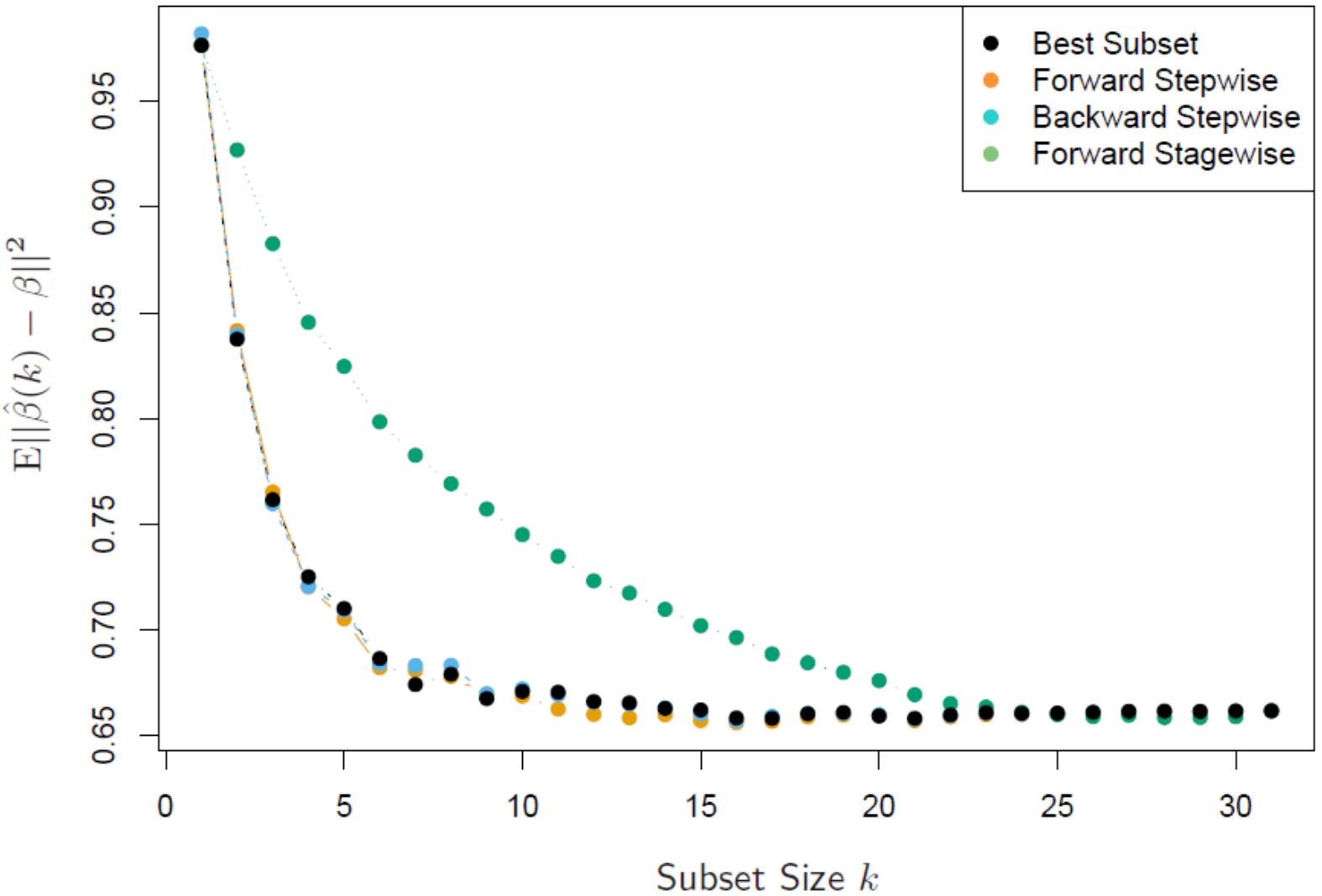
Forward stagewise: kroki

1. Startujemy z modelu ze współczynnikami wszystkich zmiennych równymi 0. Rozważamy $\hat{\mu}$ - estymator Y w danej iteracji. Przed rozpoczęciem algorytmu centrujemy wszystkie zmienne.
2. W każdym kroku tworzymy wektor korelacji $c(\hat{\mu})$ zmiennych X z aktualnymi wartościami resztowymi $Y - \hat{\mu}$. Wybieramy zmienną z największą korelacją.
3. Powiększamy wartość $\hat{\mu}$ dodając wartość wybranej zmiennej X_j przemnożoną przez mały współczynnik, tj:

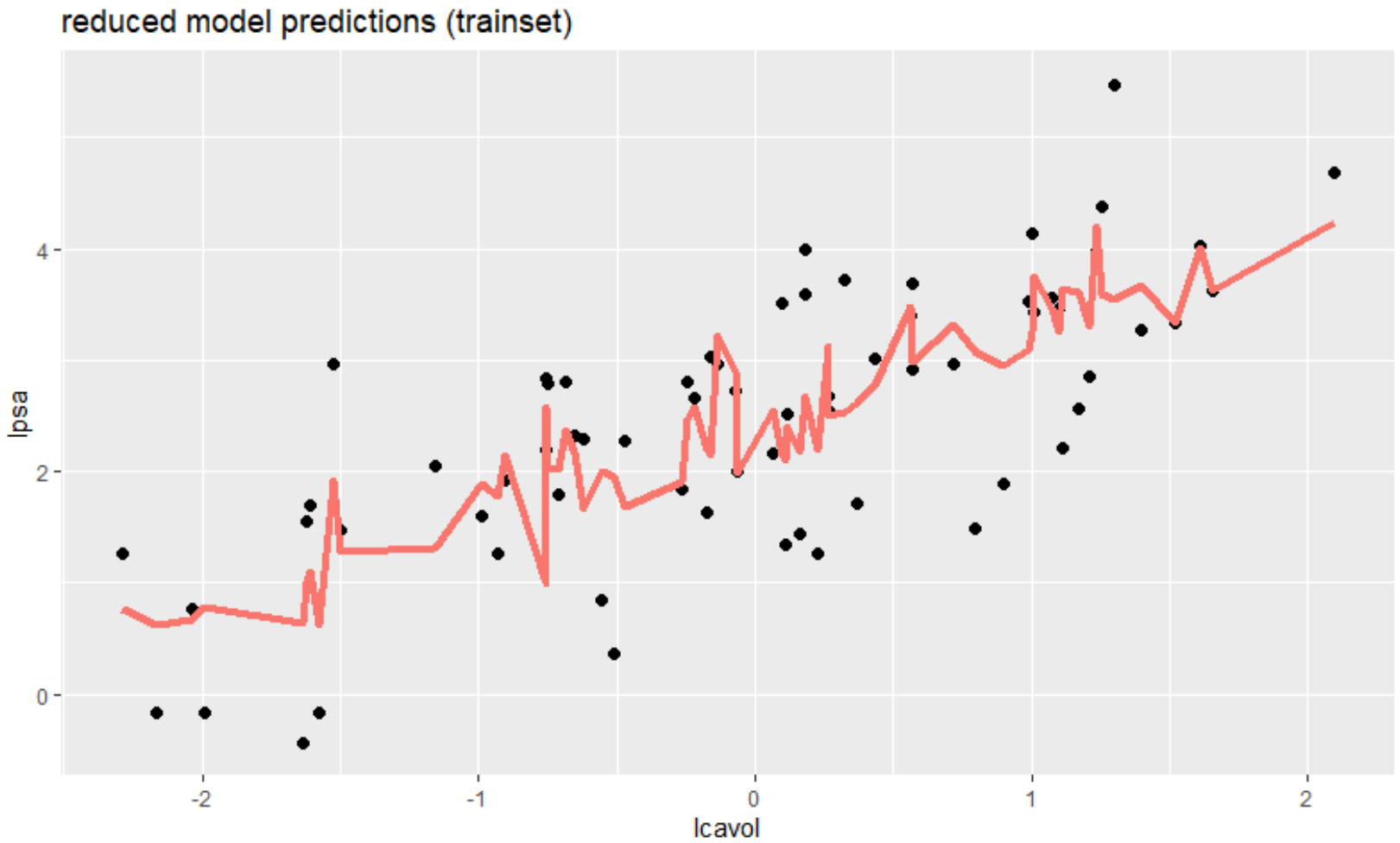
$$\hat{\mu} \rightarrow \hat{\mu} + \epsilon \cdot sign(c_j(\hat{\mu})) \cdot X_j$$

gdzie ϵ jest ustalonym parametrem, wielkością kroku.

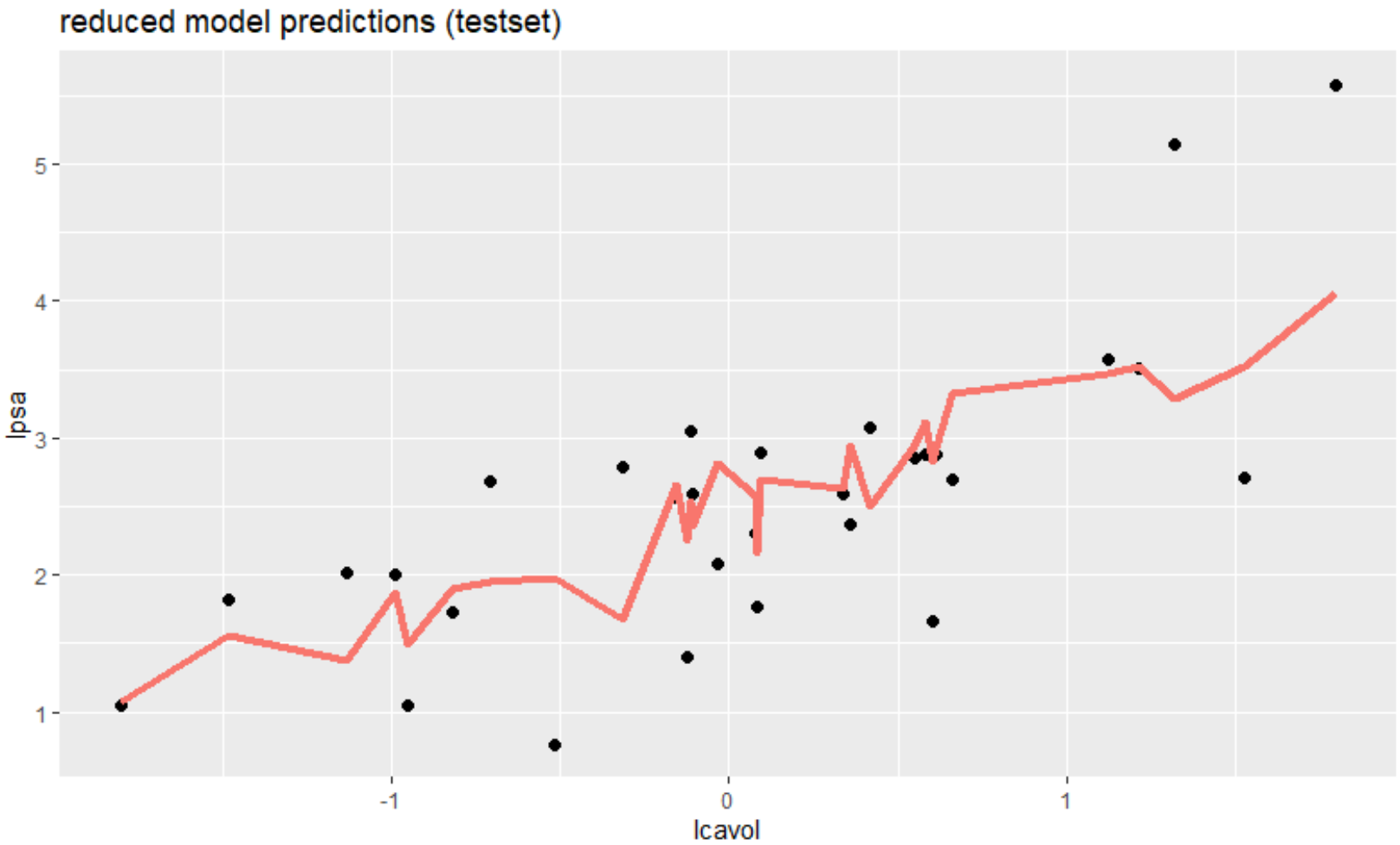
Porównanie


Best Subset
Forward Stepwise
Backward Stepwise
Forward Stagewise
$\mathrm{E}||\hat{\beta}(k) - \beta||^2$
0.65
0.70
0.75
0.80
0.85
0.90
0.95
0
5
10
15
20
25
30
Subset Size k

Dane o raku prostaty: model zredukowany



Dane o raku prostaty: model zredukowany



Dane o raku prostaty: metody doboru podzbioru vs. metody alternatywne

Term	LS	Best Subset	Ridge	Lasso	PCR	PLS
Intercept	2.465	2.477	2.452	2.468	2.497	2.452
lcavol	0.680	0.740	0.420	0.533	0.543	0.419
lweight	0.263	0.316	0.238	0.169	0.289	0.344
age	−0.141		−0.046		−0.152	−0.026
lbph	0.210		0.162	0.002	0.214	0.220
svi	0.305		0.227	0.094	0.315	0.243
lcp	−0.288		0.000		−0.051	0.079
gleason	−0.021		0.040		0.232	0.011
pgg45	0.267		0.133		−0.056	0.084
Test Error	0.521	0.492	0.492	0.479	0.449	0.528
Std Error	0.179	0.143	0.165	0.164	0.105	0.152