Wstęp do uczenia maszynowego

Regularyzacja, redukcja wymiaru

Ewa Szczurek + BW (modyfikacje)

bartek@mimuw.edu.pl
Instytut Informatyki
Uniwersytet Warszawski

22 kwietnia 2024

UNIWERSYTET
WARSZAWSKI

Metody regularyzacji estymacji współczynników

Przypomnienie

- Metoda najmniejszych kwadratów: minimalizacja RSS

$$RSS = \sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2.$$

- Dla modelu liniowego

$$RSS = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2.$$

Regresja grzbietowa (*ridge regression*)

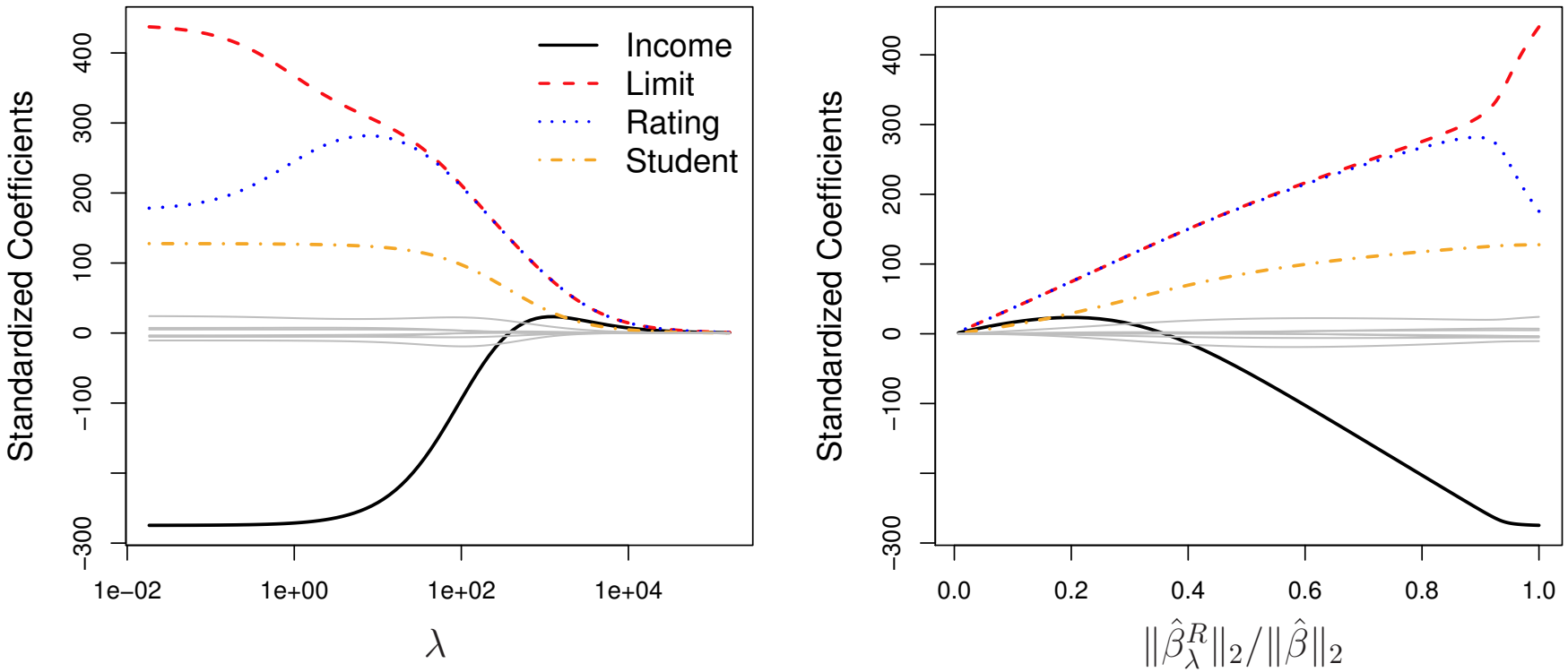
- Podobnie, plus **kara ściągająca** (*shrinkage penalty*). Minimalizacja wyrażenia

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2 = RSS + \lambda\sum_{j=1}^{p}\beta_j^2,$$

  gdzie $\lambda \geq 0$ jest **parametrem sterującym** metody (*tuning parameter*), dobieranym osobno.
- Parametr λ kontroluje jak silnie parametry są ściągane do 0.
  - Dla $\lambda = 0$ zwykła metoda najmniejszych kwadratów.
  - Dla $\lambda \to \infty$ współczynniki β_i kurczą się do zera.
- Nie ściągamy wyrazu wolnego β_0 - jest to predykcja średniej wartości zmiennej objaśnianej gdy wszystkie predyktory są równe 0.
- Gdy wszystkie kolumny X scentrowane wokół 0

$$\hat{\beta}_0 = \bar{y} = \sum_{i=1}^{n} y_i/n.$$

Regresja grzbietowa dla danych 'Credit'

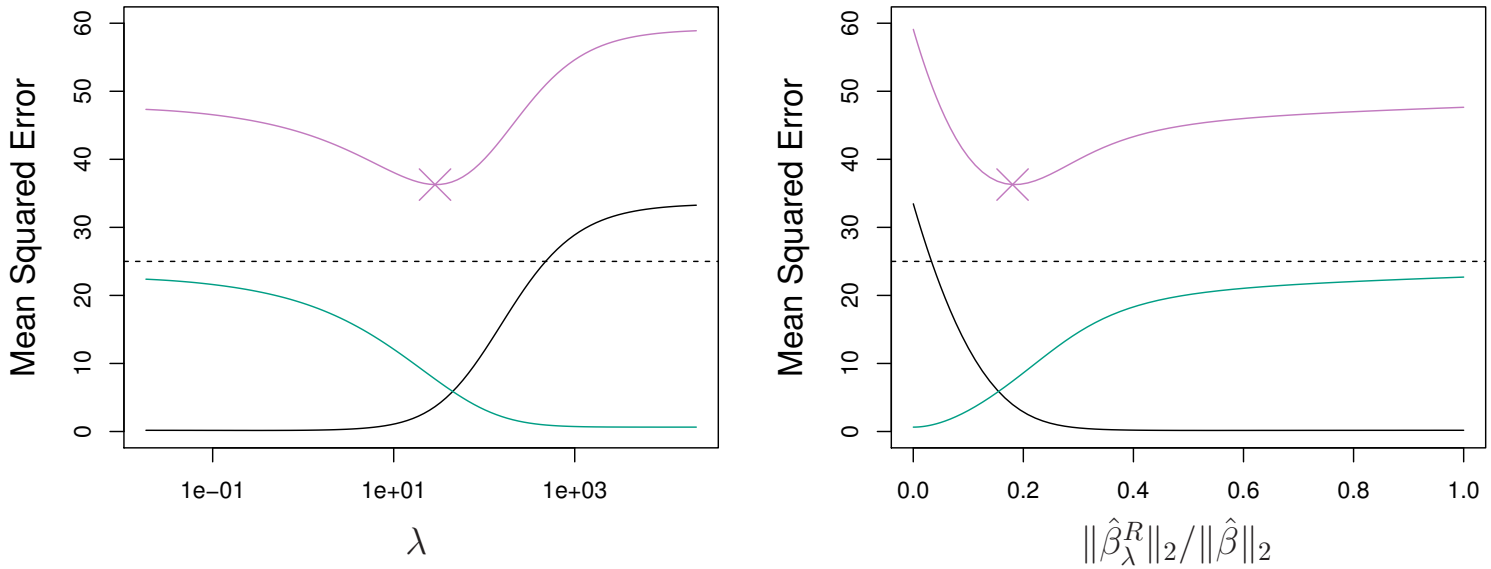


- $\hat{\beta}^R_\lambda$ wektor wyestymowany dla regresji grzbietowej i parametru λ
- $\hat{\beta}$ wektor wyestymowany dla metody najmniejszych kwadratów
- $||\beta||_2 = \sqrt{\sum_{j=1}^{p} \beta_j^2}$ (norma ℓ_2). Mierzy odległość β od 0.
- Przy rosnącym λ, $||\hat{\beta}^R_\lambda||_2$ zawsze maleje.

Uwaga: skalowanie predyktorów ma znaczenie

- W przypadku regresji liniowej: pomnożenie predyktora X_j przez stałą c zmniejsza estymowany parametr o $1/c$ ($X_j\hat{\beta}_j$ się nie zmienia)
- W przypadku regresji grzbietowej, z powodu kary ściągającej skalowanie predyktora X_j może znacząco zmienić zarówno estmowany parametr $\hat{\beta}_j$ jak i inne predyktory
  - Przykład: zmiana jednostek zarobków rocznych z PLN na 1000PLN
- Przed zastosowaniem regresji liniowej standaryzuje się predyktory tak aby były w tej samej skali.

Przewaga regresji grzbietowej nad metodą najmniejszych kwadratów (dane symulowane)



- **Czarny** – obciążenie; **zielony** – wariancja; **purpurowy** – MSE.
- Zwiększanie λ powoduje zmniejszenie elastyczności, a zatem zmniejszenie wariancji oraz zwiększenie obciążenia.
- MSE dla metody najmniejszych kwadratów ($\lambda = 0$) jest większe niż dla odpowiednio dobranego parametru $\lambda > 10$.
- Właściwy dobór λ jest krytycznie ważny.

Regresja grzbietowa a metoda najmniejszych kwadratów

- W powyższym przykładzie mieliśmy $n = 50$ obserwacji i $p = 45$ predyktorów.
- Kiedy liczba predyktorów p jest bliska liczbie obserwacji n to metoda najmniejszych kwadratów ma dużą wariancję – nieduża zmiana w danych treningowych może powodować dużą zmianę wyestymowanych parametrów.
- Gdy $p > n$ to metoda najmniejszych kwadratów przestaje działać (brak jednoznaczności estymacji), a regresja grzbietowa pozwala wybrać najlepszy model spośród wielu dających RSS=0.
- Grzbietowa regresja może być użyta do wyboru modelu (zamiast metod typu wyszukiwanie w przód) – pomijamy te predyktory, dla których współczynnik β_j jest mały. Tu konieczny jest jakiś dobór odcięcia.

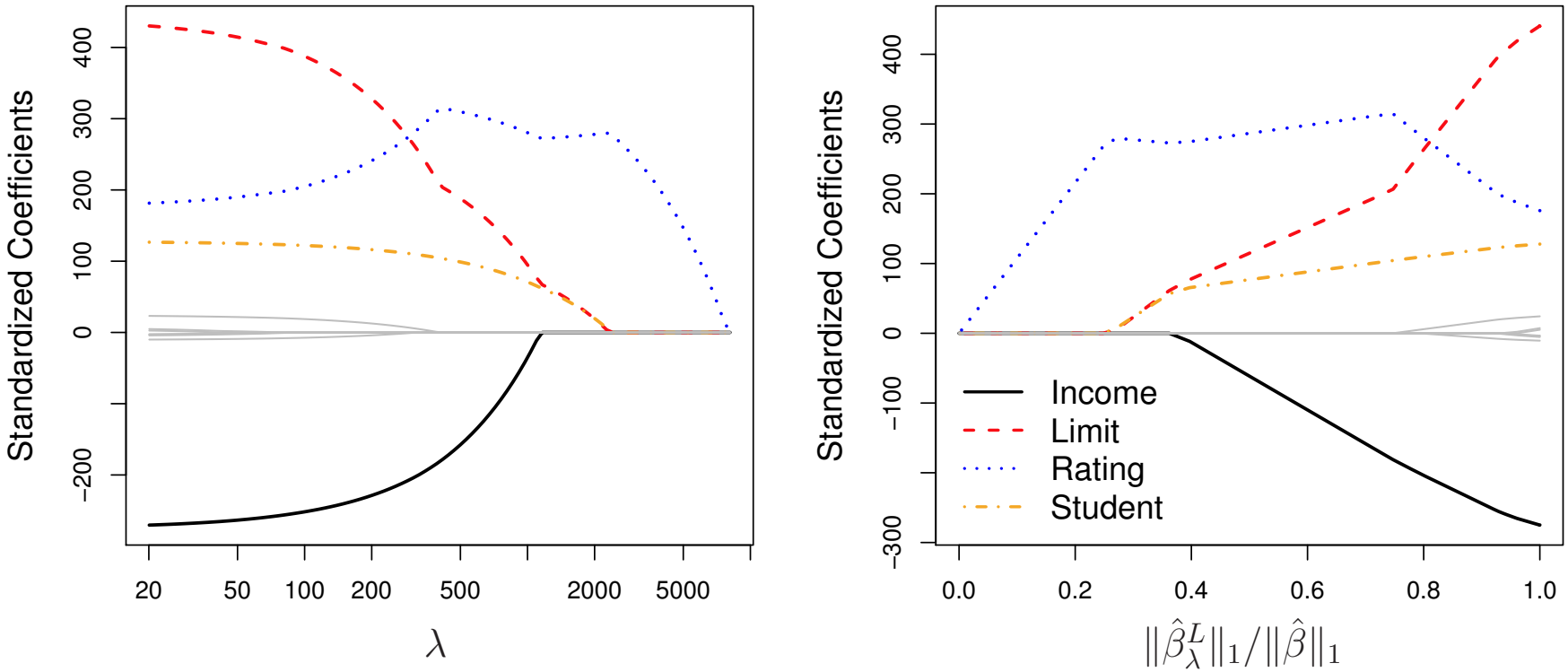
Metoda *lasso*

- Regresja grzbietowa tworzy modele zawierające wszystkie p predyktorów. Współczynniki β_j kurczą się ze wzrostem λ, ale nie muszą osiągać zera.
- **Metoda lasso** polega na minimalizacji wyrażenia

$$\sum_{i+1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| = RSS + \lambda\sum_{j=1}^{p}|\beta_j|,$$

  gdzie $\lambda \geq 0$ jest specjalnie dobranym parametrem sterującym.
- Tutaj $||\beta||_1 = \sum|\beta_j|$ jest normą ℓ_1.
- Podstawowa różnica między lasso a regresją grzbietową polega na tym, że dla dostatecznie dużych wartości λ estymacje niektórych parametrów β_j przyjmują wartości **równe 0**, jednocześnie zadając **selekcję** predyktorów.

Lasso zastosowane do danych 'Credit'



W zależności od wielkości λ, lasso prowadzi do modeli opartych na 1, 2, 3, 4,... predyktorach.

Zjawisko selekcji predyktorów w metodzie lasso

Można pokazać, że dla każdego $\lambda \geq 0$ istnieje $s \geq 0$ takie, że

- metoda lasso jest równoważna problemowi minimalizacji po wektorach β wyrażenia

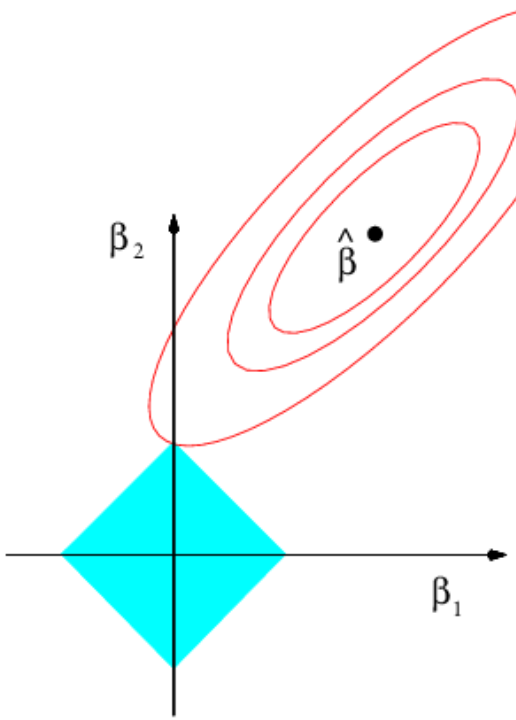
$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 \quad \text{przy warunku} \quad \sum_{j=1}^{p}|\beta_j| \leq s.$$

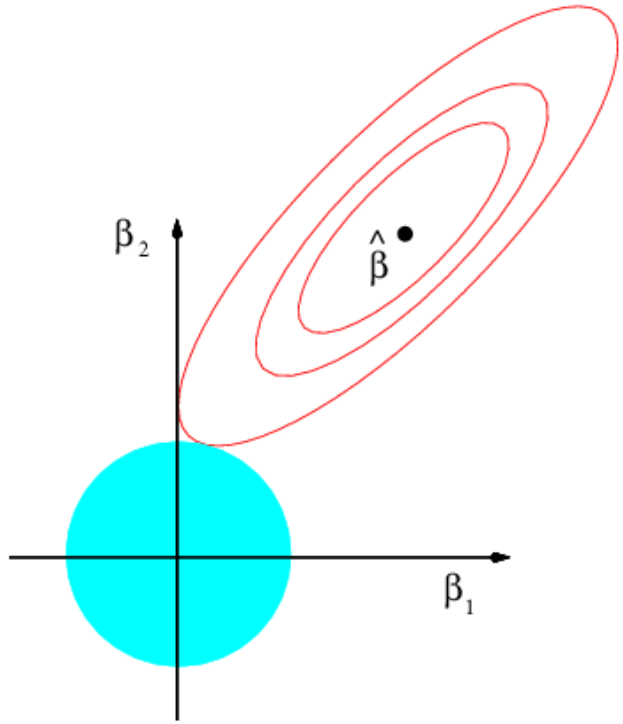
- metoda grzbietowej regresji jest równoważna problemowi minimalizacji po wektorach β wyrażenia

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 \quad \text{przy warunku} \quad \sum_{j=1}^{p}\beta_j^2 \leq s.$$

Zjawisko selekcji predyktorów w metodzie lasso dla $p = 2$

$\hat{\beta}$ – otrzymane z metody najmniejszych kwadratów. **Elipsy** opisują obszar o równym RSS.





Lasso z warunkiem $|\beta_1| + |\beta_2| \leq s$ (tutaj $\beta_1 = 0$ – wybieramy tylko drugi predyktor).

Regresja grzbietowa z waunkiem $\beta_1^2 + \beta_2^2 \leq s$. Oba współczynniki niezerowe.

Porównanie regresji grzbietowej z lasso

- Żadna z metod nie jest lepsza od drugiej we wszystkich sytuacjach.
- Lasso działa lepiej w sytuacjach, gdy zmienna objaśniana Y istotnie zależy tylko od małej liczby predyktorów.
- Regresja grzbietowa działa lepiej gdy Y zależy od dużej liczby predyktorów o współczynnikach w przybliżeniu podobnego rozmiaru.
- Liczba istotnych predyktorów nigdy nie jest znana z góry – trzeba stosować walidację krzyżową do wyboru modelu.
- W sytuacji gdy estymacje z metody najmniejszych kwadratów mają dużą wariancję, to zarówno lasso, jaki regresja grzbietowa zmniejszają wariancję predykcji kosztem zwiększenia obciążenia.
- Metoda lasso, w odróżnieniu od regresji grzbietowej, pozwala dokonywać selekcji istotnych predyktorów.

Uproszczony przykład działania lasso i regresji grzbietowej

- Przyjmujemy $n = p$, a dane treningowe to $(e_1, y_1), \ldots, (e_n, y_n)$, gdzie e_i to i-ty wektor jednostkowy. Dla uproszczenia przyjmujemy, że wyraz wolny $\beta_0 = 0$.
- Wówczas metoda najmniejszych kwadratów sprowadza się do minimalizacji

$$\sum_{j=1}^{p}(y_j - \beta_j)^2,$$

  co daje $\hat{\beta}_j = y_j$.
- Regresja grzbietowa polega tu na minimalizacji

$$\sum_{j=1}^{p}(y_j - \beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2,$$

  co daje

$$\hat{\beta}_j^R = y_j/(1 + \lambda).$$

Uproszczony przykład działania lasso i regresji grzbietowej

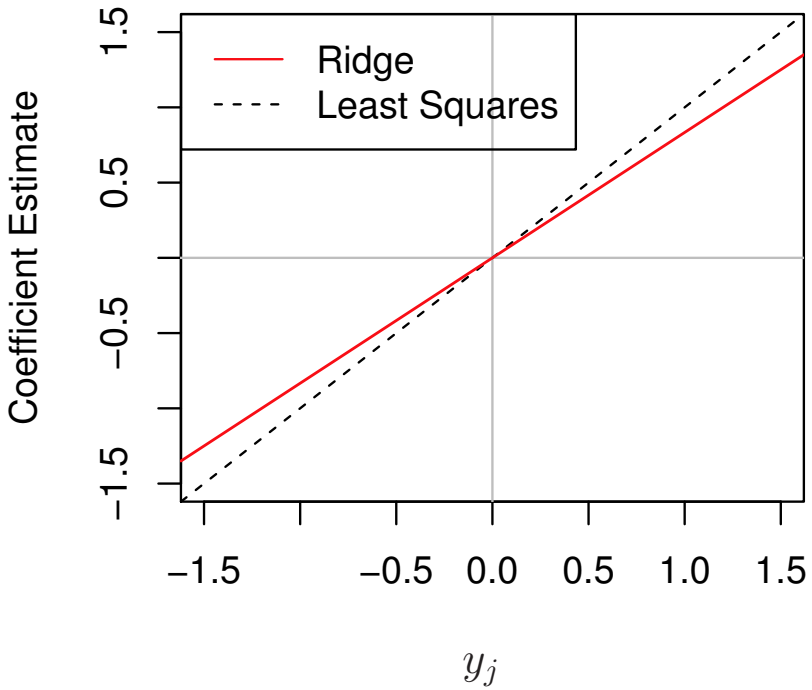
- Metoda lasso polega tu na minimalizacji

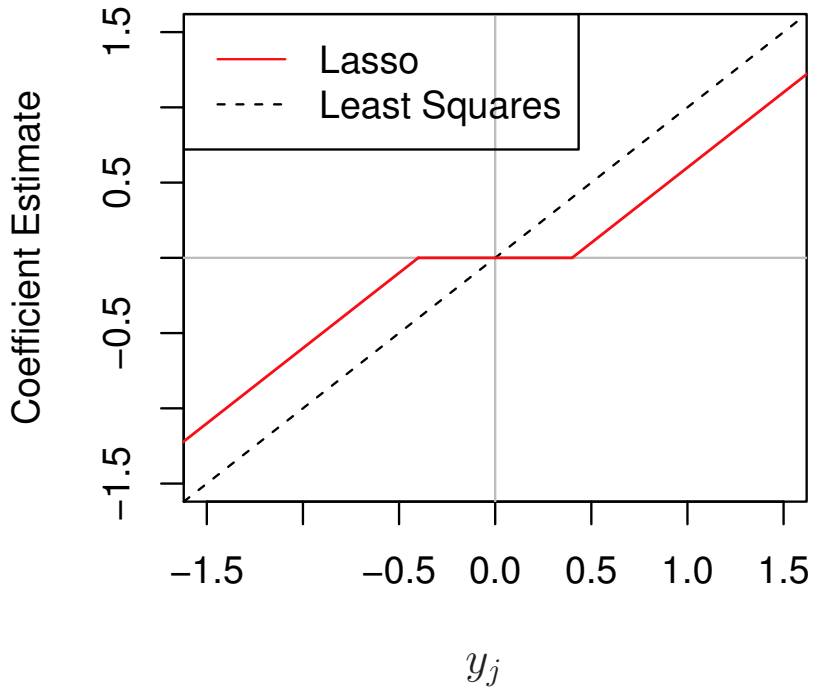
$$\sum_{j=1}^{p}(y_j - \beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|,$$

- co prowadzi do estymacji

$$\hat{\beta}_j^L = \begin{cases} y_j - \lambda/2, & \text{jeśli } y_j > \lambda/2; \\ y_j + \lambda/2, & \text{jeśli } y_j < -\lambda/2; \\ 0, & \text{jeśli } |y_j| \leq \lambda/2 \end{cases}$$

Bardzo różne metody kurczenia parametrów w regresji grzbietowej i w lasso ($\lambda = 1$)



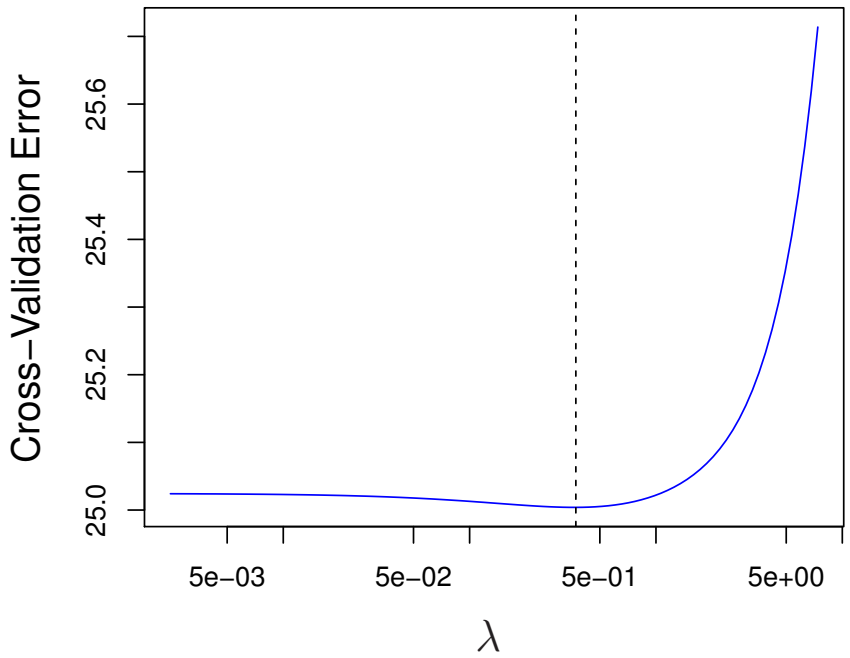


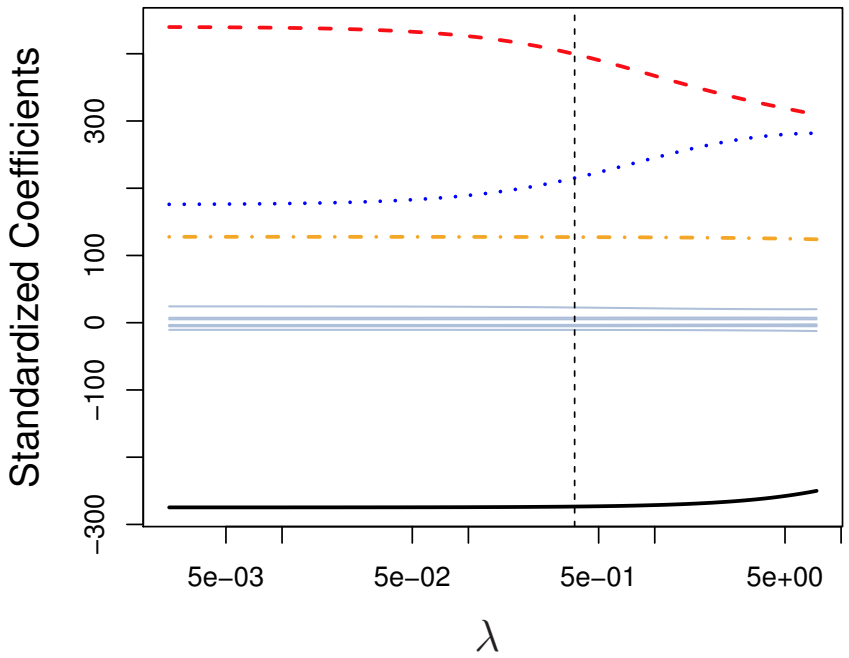
Jak wybrać parametr λ?

Stosując metodę walidacji krzyżowej:

- dla siatki różnych wartości λ obliczamy błąd walidacji krzyżowej dla tego wyboru λ.
- wybieramy tę wartość λ, dla której ten błąd jest najmniejszy
- trenujemy model już na całych danych, z wybranym λ.

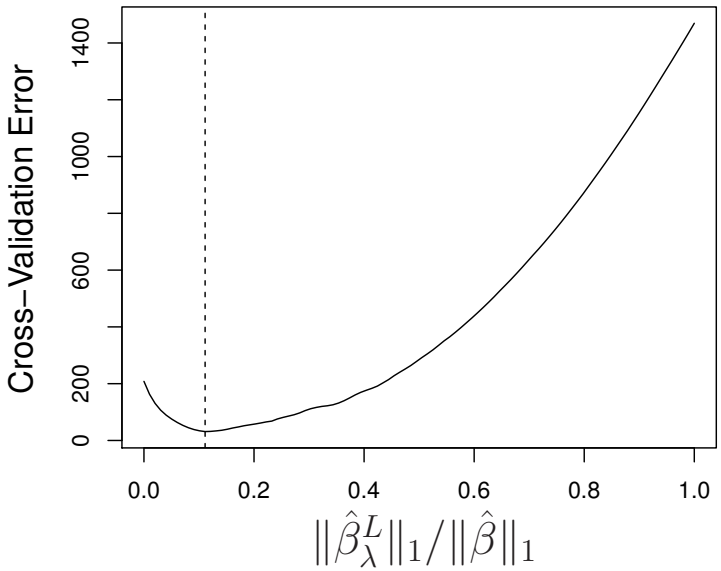
Przykład: wybór λ dla danych 'Credit' oraz regresji grzbietowej

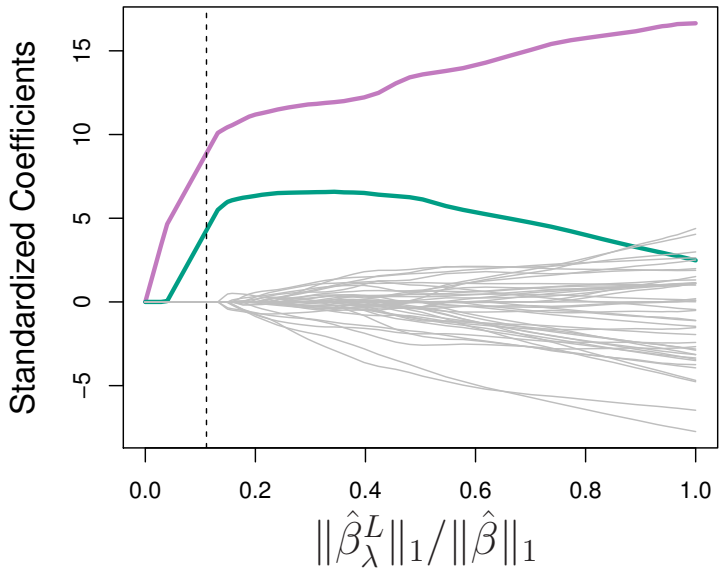




- Pionowa linia przerywana odpowiada minimum błędu CV.
- Minimum nie jest wyraźne, końcowy model zbliżony do otrzymanego metodą najmniejszych kwadratów

Wybór λ dla metody lasso





- Dla tej wartości λ model wybrany metodą lasso oparty jest na dwóch predyktorach (współczynniki pozostałych tu się zerują).
- Istotne predyktory (kolorowe wykresy) to **sygnał** a nieistotne (szare) to **szum**.
- Estymacja metodą najmniejszych kwadratów (prawy koniec wykresu) nadaje dużą wartość tylko jednemu z dwu predyktorów stanowiących sygnał.

Metody redukcji wymiaru

Ogólna metodologia redukcji wymiaru (1)

Dotąd opisane metody kontrolowania wariancji polegały na:

- wyborze podzbioru istotnych predyktorów (zmiennych), lub
- zastosowaniu ściągania parametrów do zera.
- wszystkie te metody operowały na (pod)-zbiorze oryginalnych predyktorów $X_1, X_2, \ldots, X_p$.
- Redukcja wymiaru polega na zastąpieniu oryginalnych predyktorów $X_1, X_2, \ldots, X_p$ liniowymi kombinacjami $Z_1, Z_2, \ldots, Z_M$, gdzie $M \leq p$ oraz dla $1 \leq m \leq M$

$$Z_m = \sum_{j=1}^{p} \phi_{jm} X_j,$$

gdzie $\phi_{1m}, \phi_{2m}, \ldots, \phi_{pm}$ są pewnymi stałymi (parametrami) metody.

Ogólna metodologia redukcji wymiaru (2)

- Stosując metodę najmniejszych kwadratów znajdujemy parametry $\theta_0, \theta_1, \ldots, \theta_M$ modelu regresji liniowej

$$y_i = \theta_0 + \sum_{m=1}^{M} \theta_m z_{im} + \epsilon_i, \qquad i = 1, \ldots, n$$

  gdzie $z_{i1}, \ldots, z_{iM}$ jest otrzymane z wartości $x_{i1}, \ldots, x_{ip}$ przez zastosowanie powyższego przekształcenia liniowego.

- W oryginalnym zadaniu regresji liniowej stosujemy metodę najmniejszych kwadratów (estymując parametry $\beta_0, \beta_1, \ldots, \beta_p$) do wyrażenia

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i, \qquad i = 1, \ldots, n$$

Ogólna metodologia redukcji wymiaru (3)

- Pojęcie **redukcji wymiaru** polega na zastąpieniu estymowania $p+1$ parametrów $\beta_0, \ldots, \beta_p$ estymacją $M+1$ parametrów $\theta_0, \ldots, \theta_M$. Czyli wymiar problemu został zredukowany z $p+1$ do $M+1$.
- Przy odpowiednim doborze stałych ϕ_{jm} taka redukcja wymiaru często prowadzi do poprawienia jakości predykcji.
- Podejście to można traktować jako zadanie regresji liniowej z nałożonymi *więzami*:

$$\sum_{m=1}^{M} \theta_m z_{im} = \sum_{m=1}^{M} \theta_m \sum_{j=1}^{p} \phi_{jm} x_{ij} = \sum_{j=1}^{p} \sum_{m=1}^{M} \theta_m \phi_{jm} x_{ij} = \sum_{j=1}^{p} \beta_j x_{ij}.$$

Zatem więzy nałożone na β_j wyglądają następująco:

$$\beta_j = \sum_{m=1}^{M} \theta_m \phi_{jm}.$$

Dwa kroki metody redukcji wymiaru

1. Przekształcamy predyktory $X_1, \ldots, X_p$ w nowy zestaw predyktorów $Z_1, \ldots, Z_M$ ($M \leq p$).
2. Stosujemy regresję liniową do nowych predyktorów $Z_1, \ldots, Z_M$.

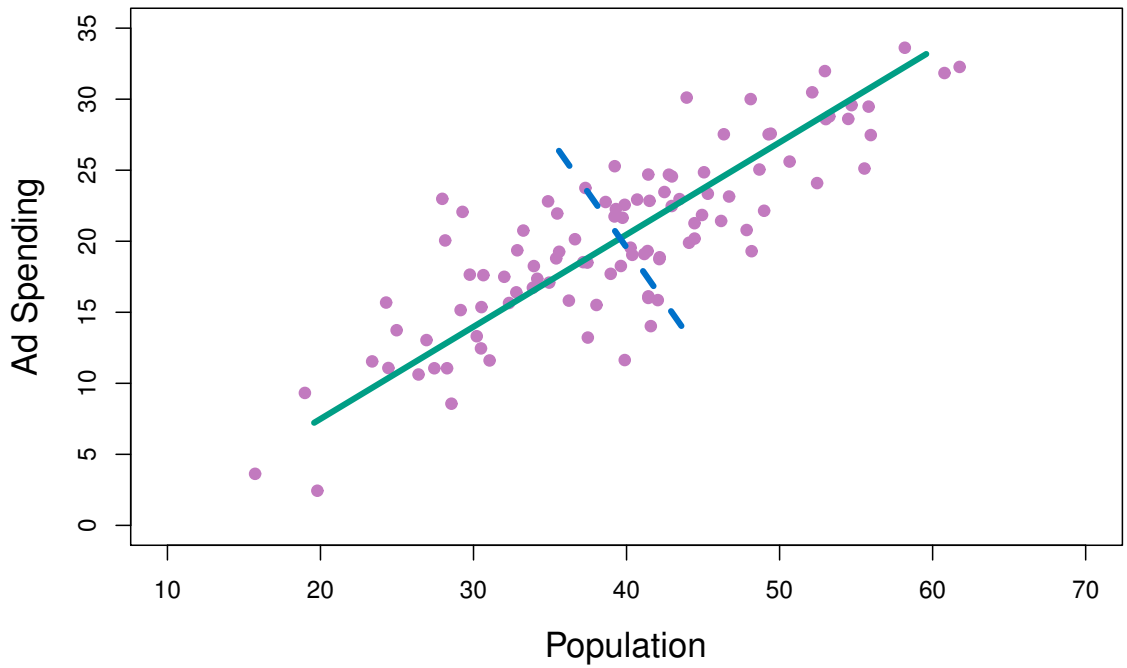
- Wybór predyktorów $Z_1 \ldots, Z_M$ może byc dokonywany na różne sposoby.
- Przykłady:
  - Regresja składowych głównych (*Prinicpal Components Regression*) - omówimy poniżej
  - Metoda częściowych najmniejszych kwadratów (*Partial Least Squares*)

Analiza składowych głównych

Principal component analysis (PCA)

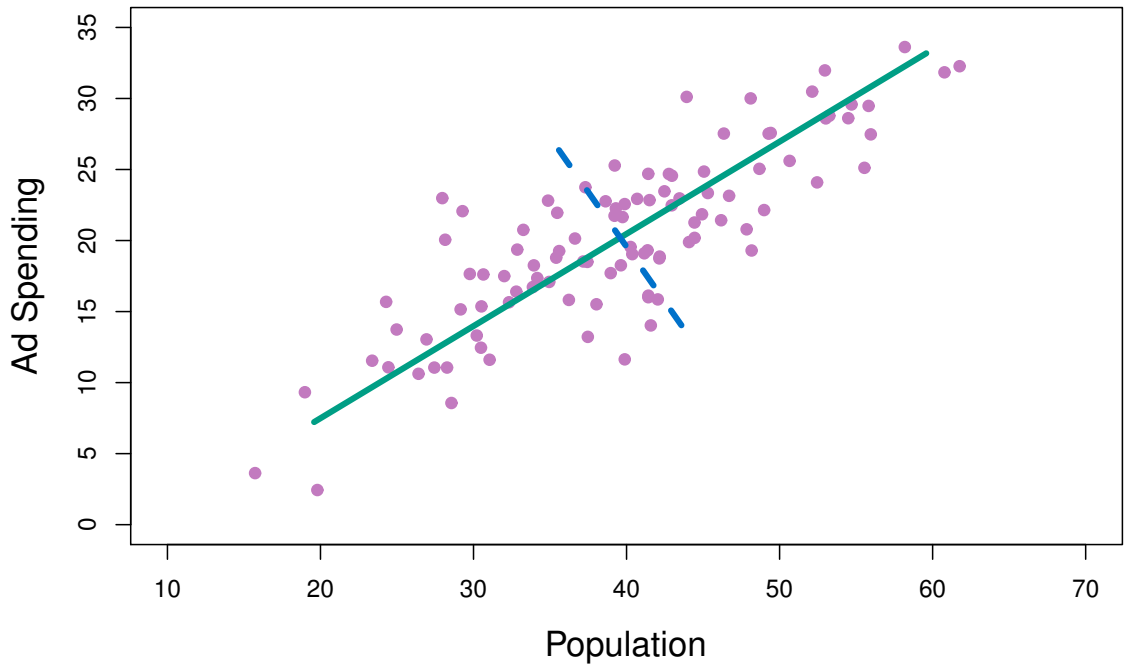
- Określa kierunki o największej zmienności w danych
- Kolejne kierunki są do siebie ortogonalne

Przykład: 100 miast



- Dla każdego miasta para wartości: ad (nakłady na reklamę) i pop (liczba mieszkańców)
- Zielona linia: 1sza składowa główna, niebieska: 2ga składowa główna

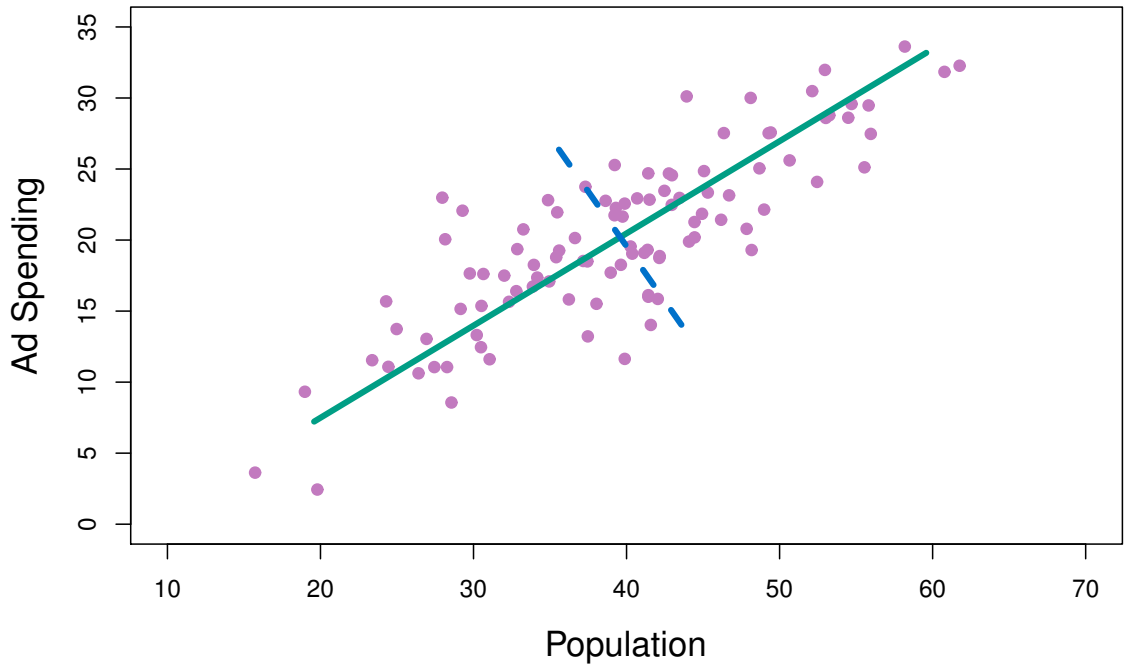
Pierwsza składowa główna



- $Z_1 = 0.839(\text{pop} - \overline{\text{pop}}\,\mathbb{1}) + 0.544(\text{ad} - \overline{\text{ad}}\,\mathbb{1})$,
  gdzie $\mathbb{1}$ jest wektorem złożonym z samych jedynek (wymiaru 100)
- Tutaj $\phi_{11} = 0.839$ in $\phi_{21} = 0.544$ nazywane są *ładunkami*
- i-te wejście wektora Z_1:

$$z_{i1} = 0.839(\text{pop}_i - \overline{\text{pop}}) + 0.544(\text{ad}_i - \overline{\text{ad}})$$

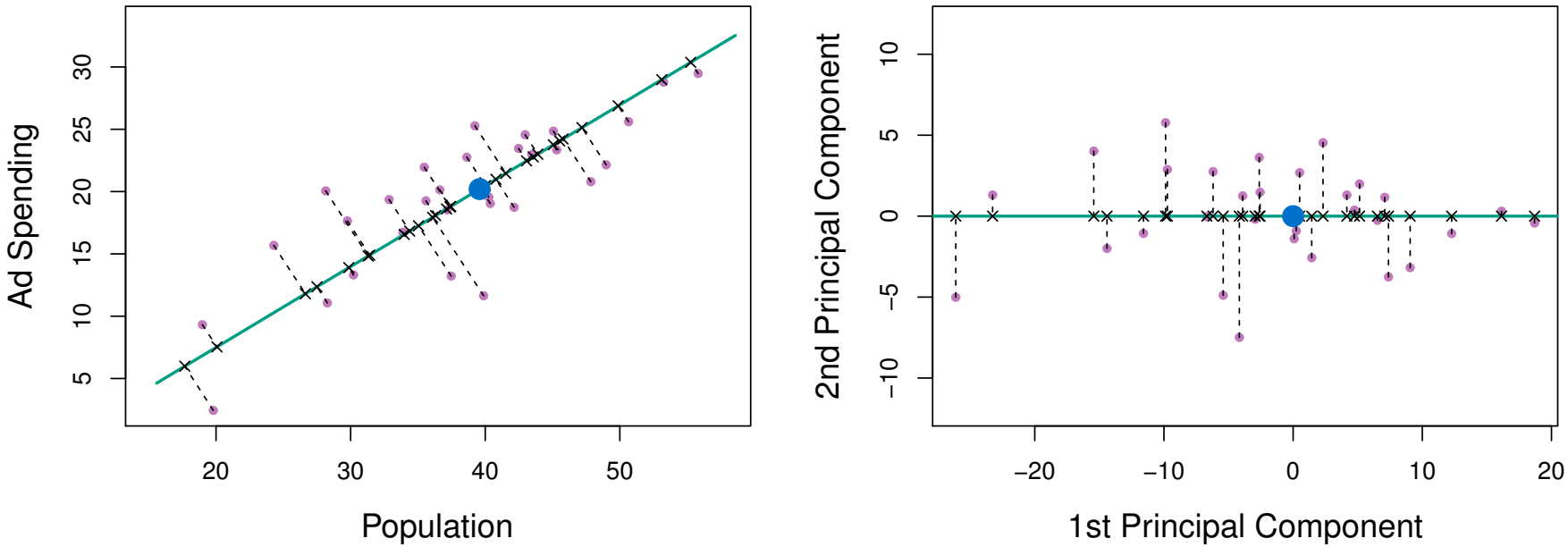
Pierwsza składowa główna



- Z_1 to taka kombinacja liniowa pop i ad, dla której wariancja $\text{Var}(\phi_{11}(\text{pop} - \overline{\text{pop}}) + \phi_{21}(\text{ad} - \overline{\text{ad}}))$ jest największa, przy warunku $\phi_{11}^2 + \phi_{21}^2 = 1$.
- Warunek jest potrzebny, bo wariancję możnaby dowolnie zwiększyć zwiększając ϕ_{11} lub ϕ_{21}

Inna interpretacja pierwszej składowej głównej

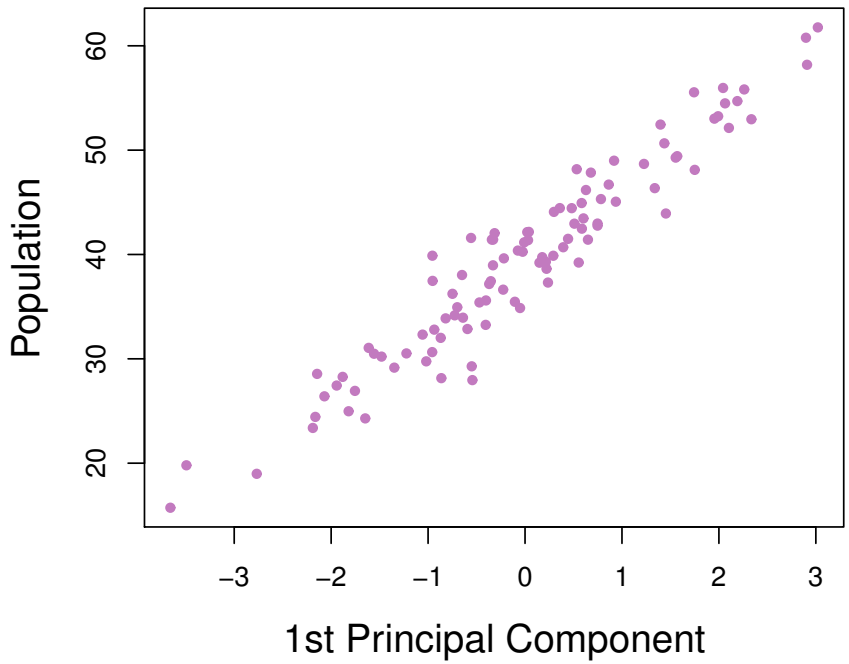
Wyznacza prostą położoną najbliżej danych (suma kwadratów odległości poszczególnych punktów od prostej jest najmniejsza).

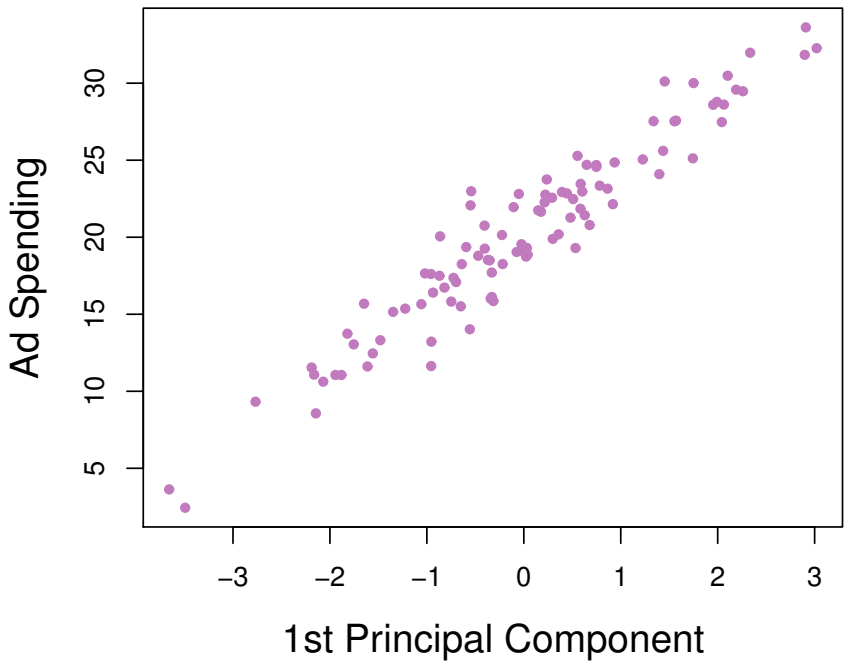


Punkt reprezentujący średnie zaznaczony na niebiesko.

Po prawej: Oś x to teraz kierunek Z_1. Kolejne obserwacje mają tu wartości z_{i1} (i-te wejścia wektora Z_1).

Dla miast pierwsza składowa główna zawiera większość istotnej informacji o danych





PCA ogólniej

- Niech X będzie macierzą o kolumnach $x_1, \ldots, x_p$ stanowiących standaryzowane wartości predyktorów.
- Jeśli mamy wyznaczone składowe główne przez wektory $Z_1, \ldots, Z_{m-1}$, to m-ty kierunek jest wyznaczony przez

$$\max_{\phi} \text{Var}(X\phi), \text{ pod warunkiem}$$

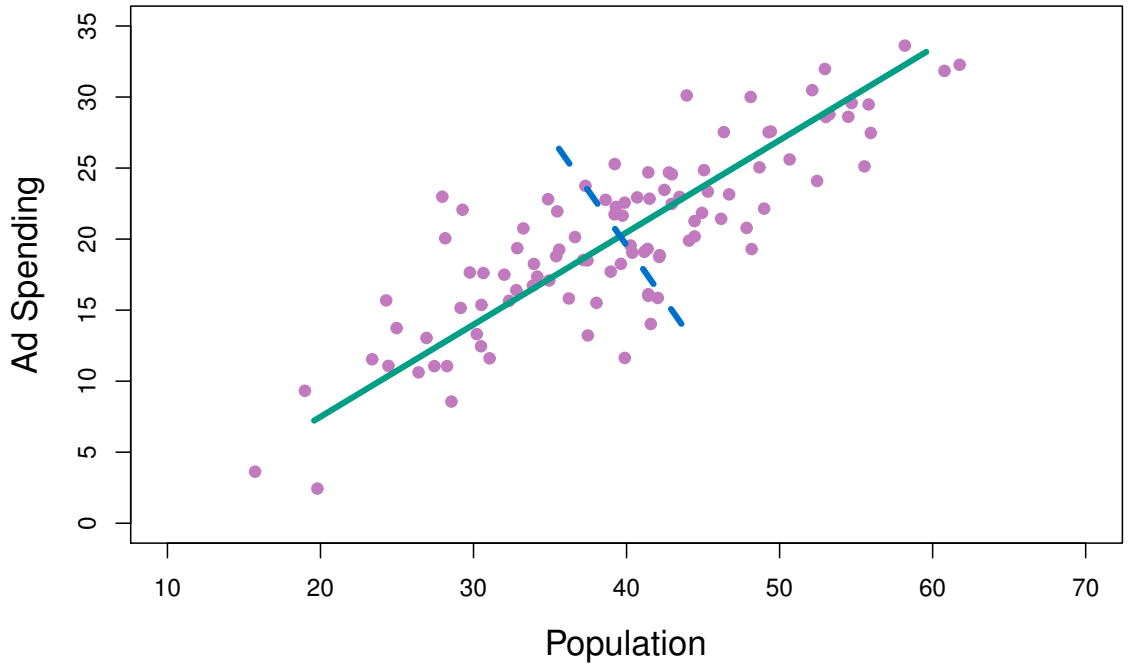
$$||\phi||_2 = 1, \ (X\phi)^T Z_i = 0, \ \text{dla } i = 1, \ldots, k-1$$

czyli jest to wektor $Z_m = X\phi$ będący

- liniową kombinacją predyktorów spełniającą warunek $||\phi||_2 = 1$
- wektorem ortogonalnym do wszystkich pozostałych kierunków Z_i (czyli nieskorelowanym z pozostałymi składowymi głównymi)

Druga składowa główna Z_2

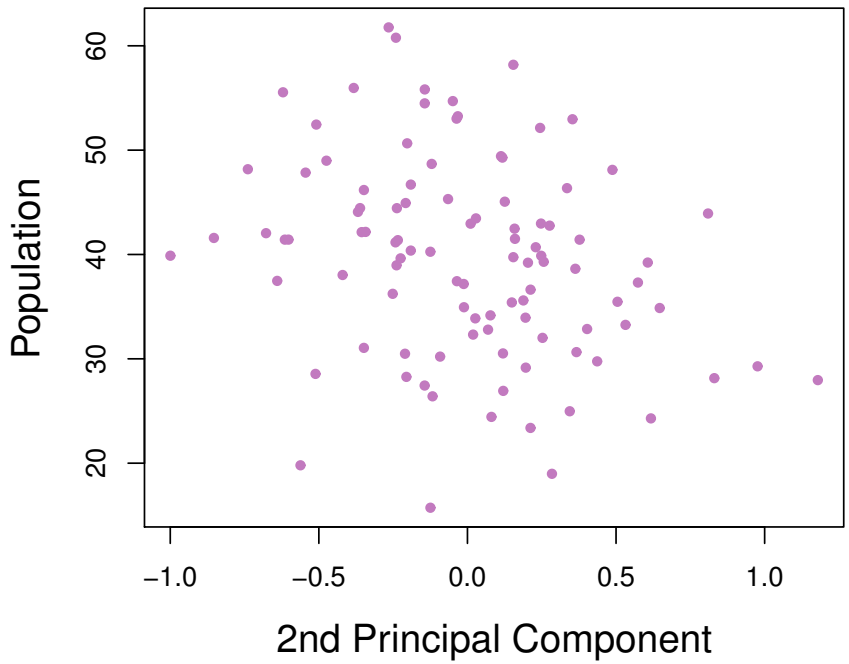
Kierunek o największej wariancji w danych, niezależny od Z_1 (**prostopadły** do Z_1; przerywana niebieska linia na rysunku).

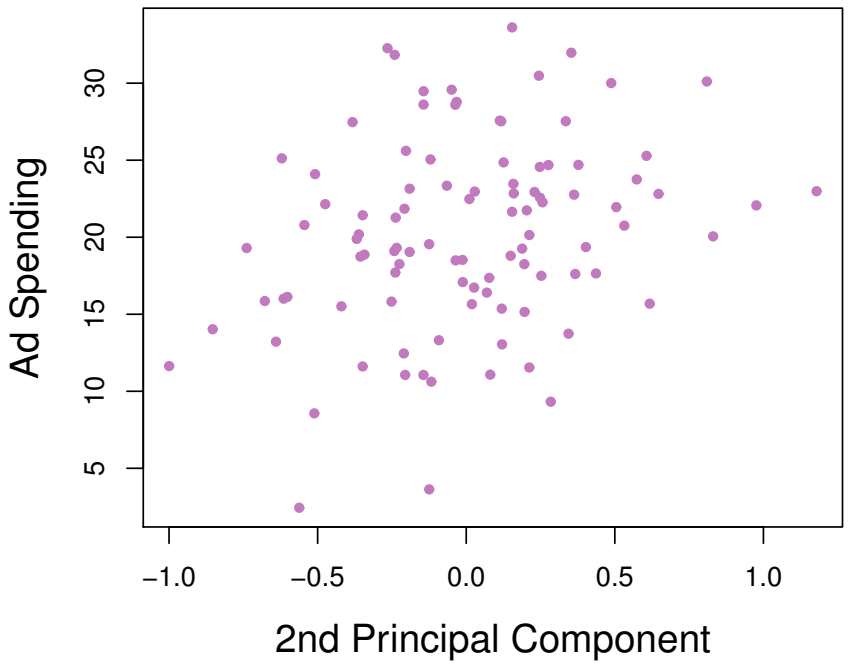


$$Z_2 = 0.544 \times (pop - \overline{pop}\,\mathbb{1}) - 0.839 \times (ad - \overline{ad}\,\mathbb{1})$$

Ponieważ mamy tu tylko dwie zmienne objaśniające, to obie składowe główne zawierają dokładnie tyle informacji co całe dane.

Przykład: słaba zależność pomiędzy drugą składową główną a obydwoma predyktorami





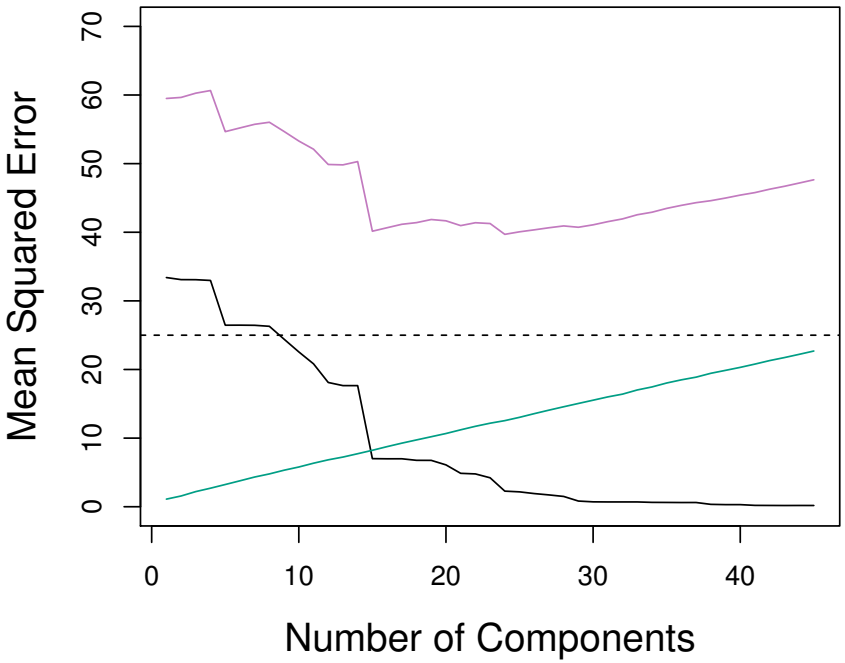
Regresja składowych głównych

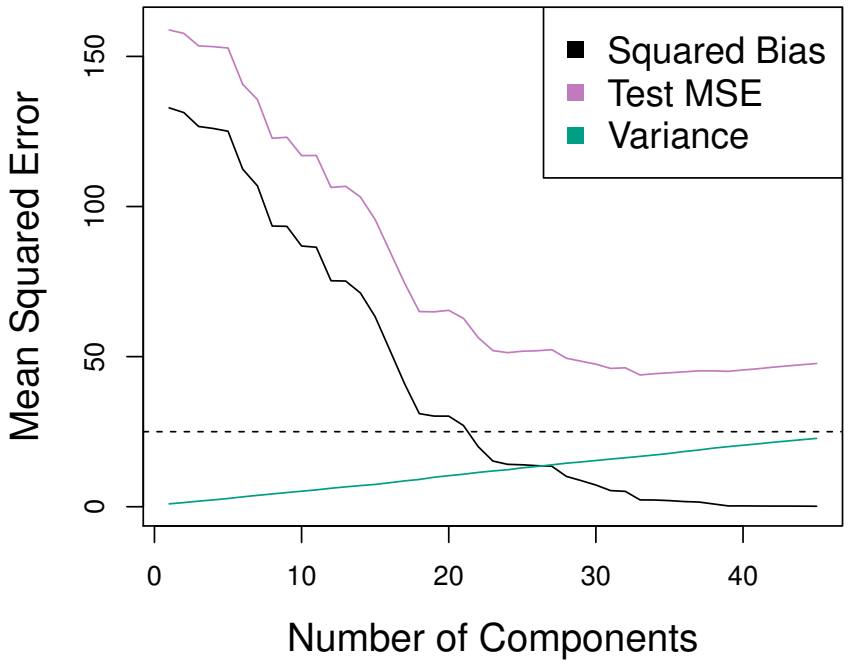
Regresja składowych głównych

- Główne założenie: stosunkowo mała liczba składowych głównych, wyjaśniająca zmienność w danych X, dobrze przewiduje wartości Y
- Potencjalnie pomaga uniknąć przeuczenia (korzystamy z $M < p$ predyktorów)

Liczba składowych głównych a jakość predykcji

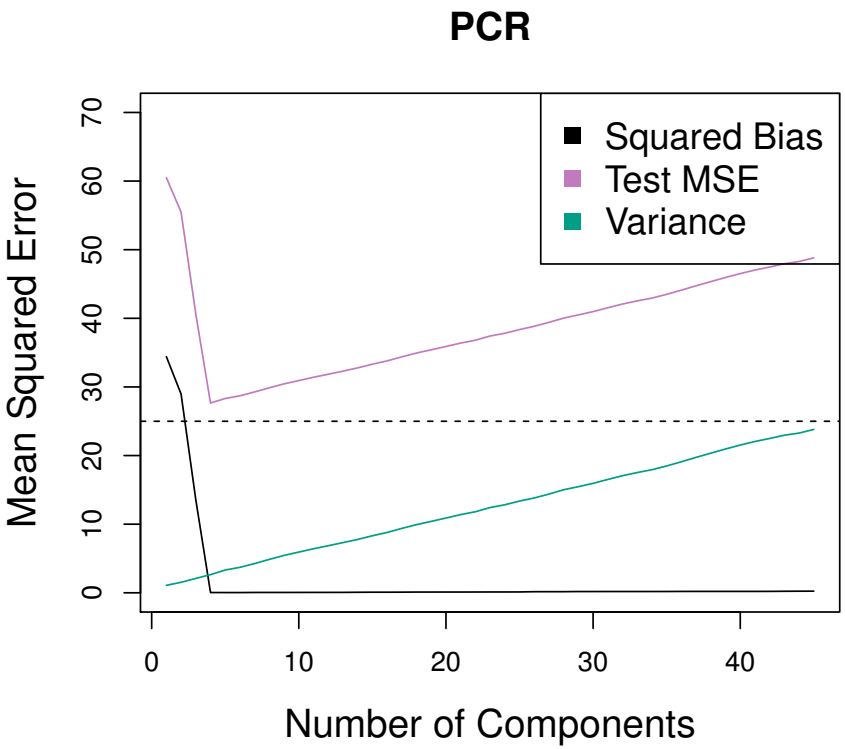
- Dwa syntetycznie generowane zbiory danych ($p = 45$, $n = 50$)
- Pozioma linia przerywana: błąd nieredukowalny $Var(\epsilon)$
- Przy wzroście liczby składowych, obciążenie maleje, a wariancja rośnie
- Potrzebujemy sporo składowych aby osiągnąć mały błąd testowy

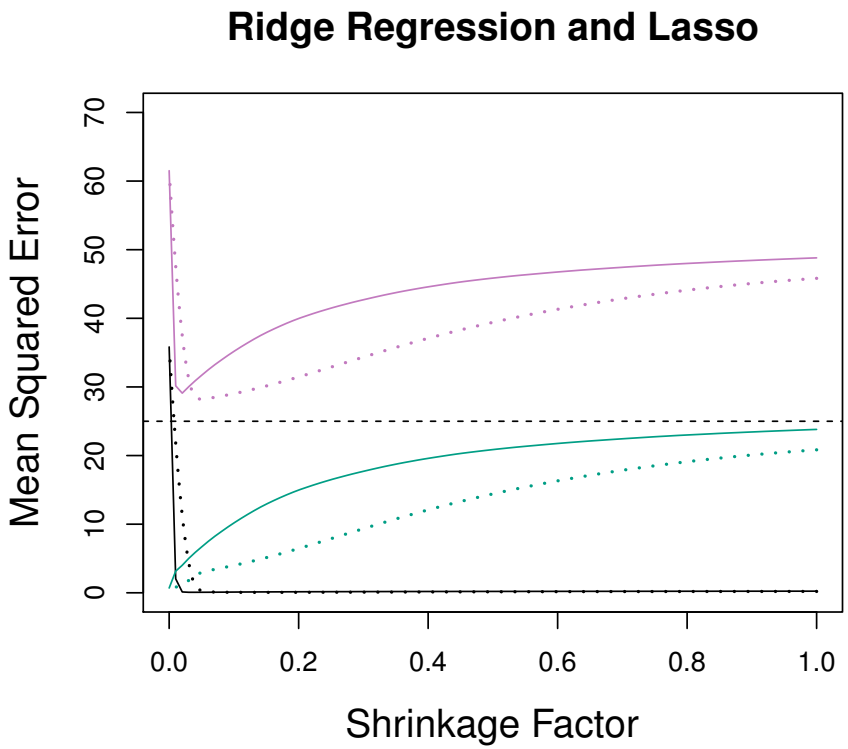




Porównanie

Dane syntetyczne ($p = 45$, $n = 50$) wygenerowane tak, aby zmienna objaśniana zależała od tylko 5 pierwszych składowych głównych





ciągła linia: lasso, przerywana linia: regresja grzbietowa .

Shrinkage factor: $||\hat{\beta}^R_\lambda||_2/||\hat{\beta}||_2$.

Ważne uwagi dotyczące regresji składowych głównych

- Dobór liczby M składowych głównych opiera się na walidacji krzyżowej.
- Przed rozpoczęciem budowy składowych głównych należy dokonać standaryzacji predyktorów
- Bez standardyzacji, zmienne objaśniające o większej wariancji miałyby największe wagi w składowych głównych
- Standardyzacji nie trzeba robić jedynie wtedy, gdy zmienne są wyrażone w tych samych jednostkach, np. kilogramach czy centymetrach.
- PCR nie jest metodą wyboru modelu. Składowe główne są liniową kombinacją wszystkich p zmiennych objaśniających.
- Ponieważ Z_m są ortogonalne, PCR to suma regresji prostych
- Dla $M = p$, PCR odpowiada regresji dla wszystkich p predyktorów

Podsumowanie

- Regularyzacja modeli liniowych
- Redukcja wymiaru, regresja składowych głównych