Düzenlileştirme

(İktisatçılar İçin) Makine Öğrenmesi (TEK-ES-2020)

Hüseyin Taştan
Yıldız Teknik Üniversitesi

Plan

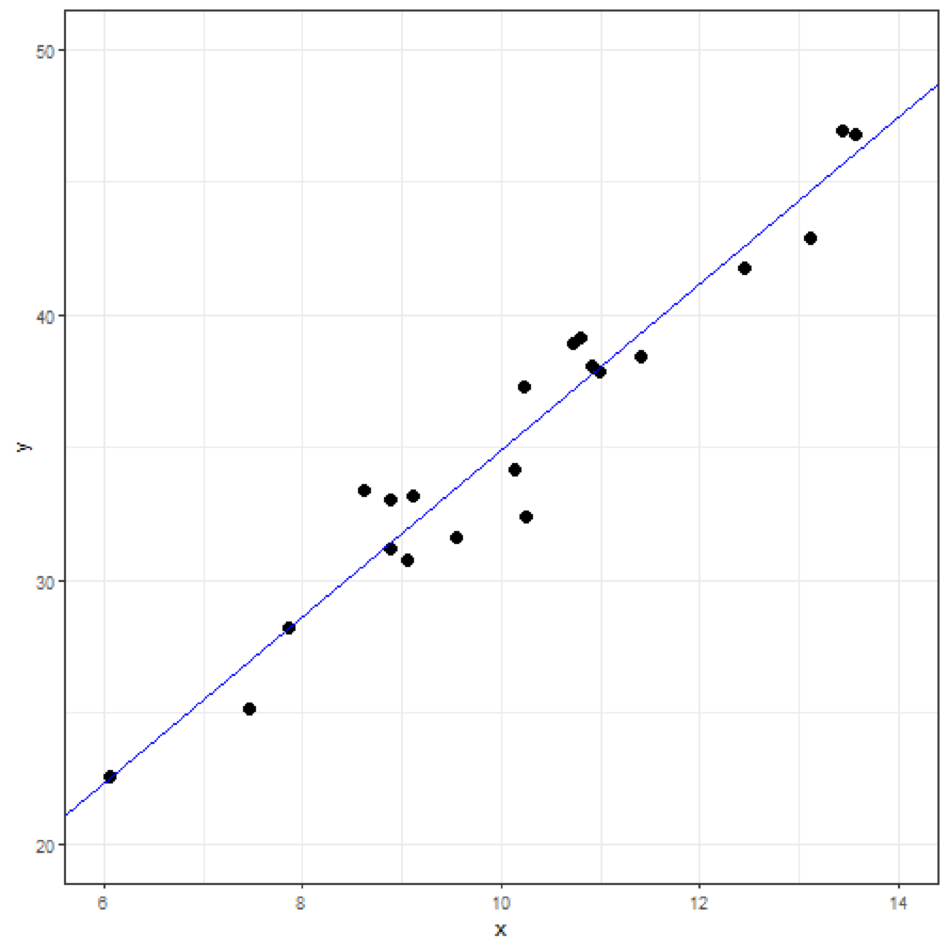
- Düzenlileştirme (Regularization)
- Çıkıntı regresyonu (Ridge regression)
- LASSO
- Elastik Net

Düzenlileştirme (Regularization)

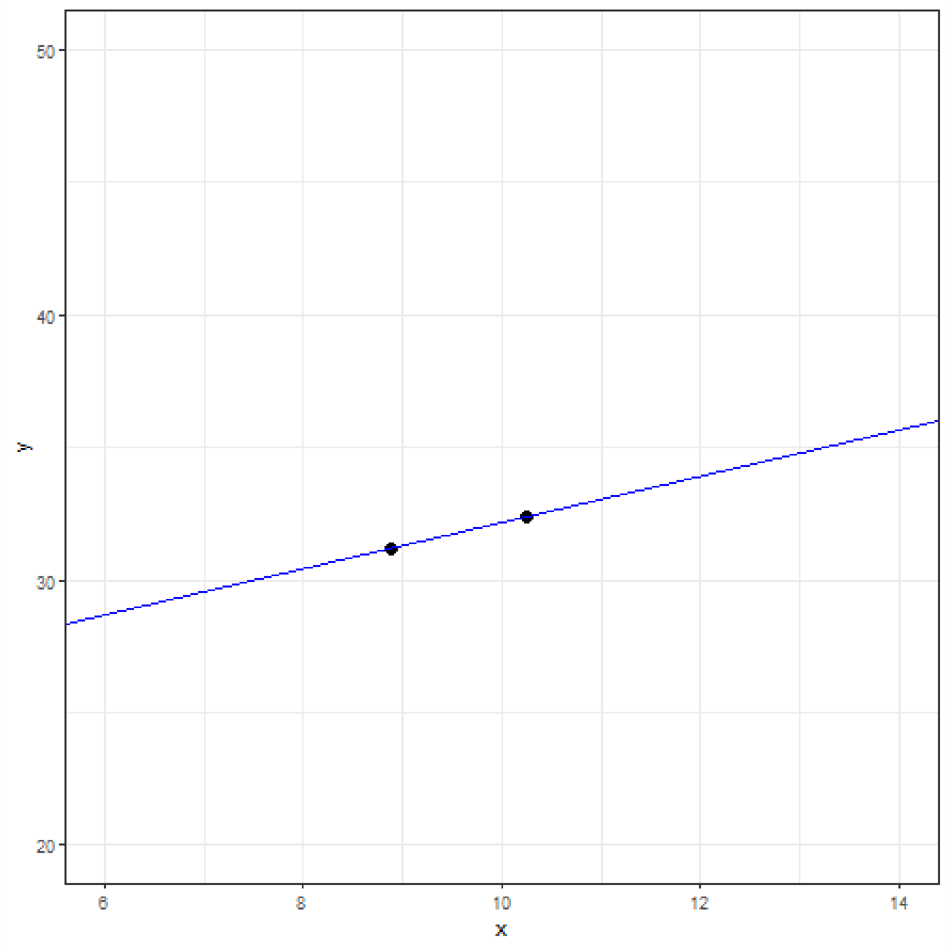
- Sıradan En Küçük Kareler (OLS) yöntemi Gauss-Markov varsayımları altında sapmasız ve en düşük varyanslı (etkin) tahminciler verir.
- Gözlem sayısının (n) değişken sayısından (p) çok daha büyük olduğu örtük olarak varsayılır: $n >> p$
- $n = p$ ise OLS tahmini **tam uyum** ile sonuçlanır.
- $p > n$ ise sonsuz sayıda OLS çözümü vardır (sonsuz varyans). OLS kullanamayız.
- Düzenlileştirme: model katsayılarını kısıtlayarak (shrinkage) varyansı düşürebilir miyiz?

Tam Uyum: Basit Regresyon

$n = 21, p = 1, R^2 = 0.94$



$n = 2, p = 1, R^2 = 1$



Çıkıntı (Ridge) Regresyonu

OLS amaç fonksiyonu

$$SSR = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_p x_{ip})^2$$

Ridge regresyonu OLS'ye çok benzer ancak amaç fonksiyonuna bir ceza terimi ekler:

$$SSR_R = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_p x_{ip})^2 + \lambda \sum_{j=1}^{p} \beta_j^2 = SSR + \lambda \sum_{j=1}^{p} \beta_j^2$$

$\lambda \geq 0$ ayarlama (tuning) parametresi

$\lambda \sum_{j=1}^{p} \beta_j^2$: küçültme cezası (shrinkage penalty). $\lambda = 0$ ise OLS=Ridge

$\lambda \to \infty$ ridge katsayıları, $\hat{\beta}_{\lambda}^{R}$, sıfıra yaklaşır. λ değiştikçe katsayı tahminleri değişir.

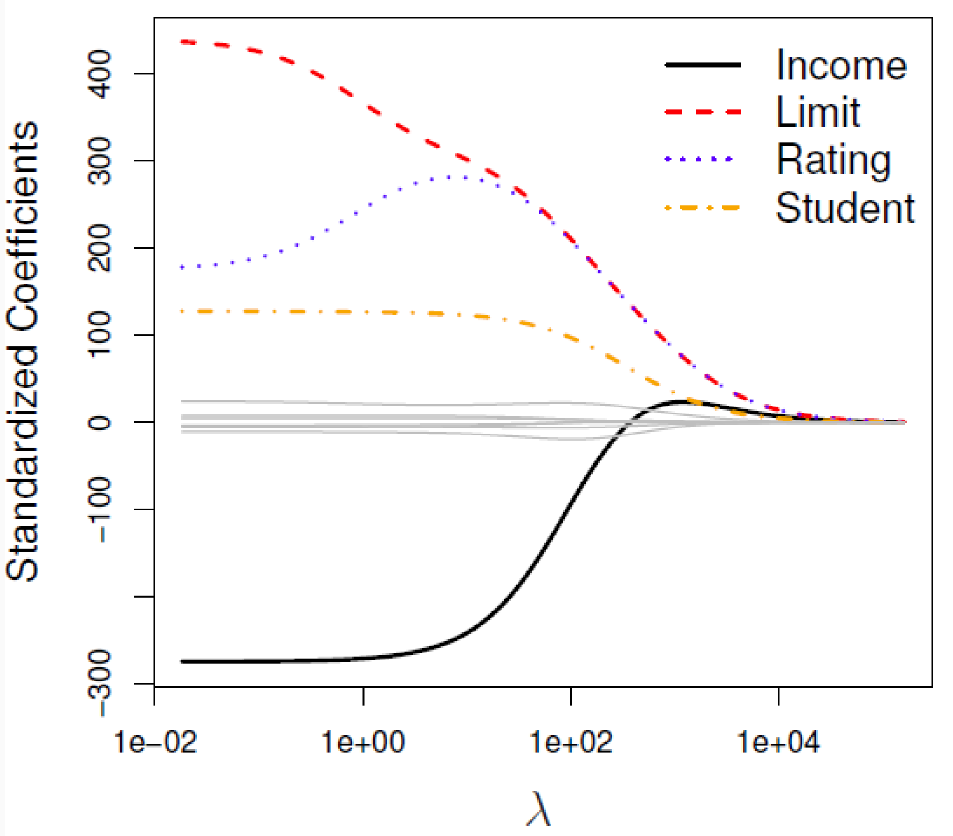
Örnek

ID	Income	Limit	Rating	Cards	Age	Education	Gender	Student	Married	Ethnicity	Balance
1	14.891	3606	283	2	34	11	Male	No	Yes	Caucasian	333
2	106.025	6645	483	3	82	15	Female	Yes	Yes	Asian	903
3	104.593	7075	514	4	71	11	Male	No	No	Asian	580
4	148.924	9504	681	3	36	11	Female	No	No	Asian	964

- $p = 10$, Çıktı değişkeni = Balance
- Amaç çıktı değişkenini en iyi kestiren doğrusal modeli kurmak.
- OLS katsayıları X'lerin ölçü birimlerine bağlı olarak değişir. Örneğin $X = Gelir$ TL olarak ölçülmüş olsun. Eğer $Gelir2 = Gelir/1000$ dönüştürmesi ile 1000TL cinsinden yeni bir değişken yaratırsak bunun katsayısı $1000 \times \hat{\beta}$ olarak değişir ve sonuçta $X \times \hat{\beta}$ aynı kalır.
- Ridge regresyonu için ise bu özellik geçerli değildir. Bu nedenle tüm değişkenleri standardize etmek gerekir (Paydada x_j'nin örneklem standart sapması yer almaktadır):

$$\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_{ij} - \bar{x}_j\right)^2}}$$

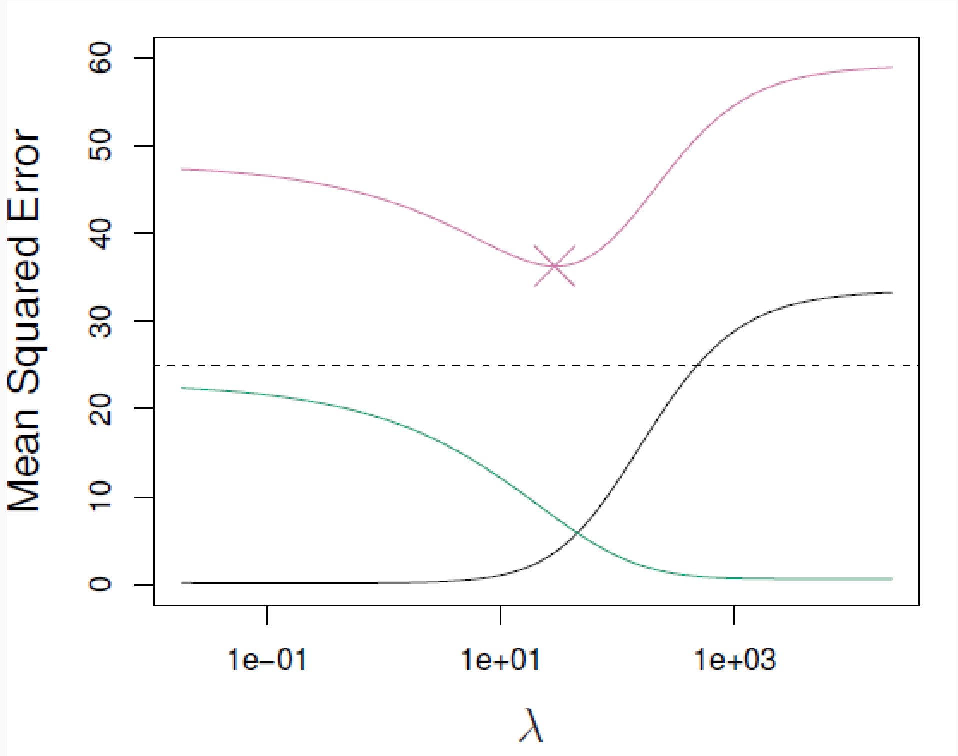
Örnek: Credit data



- Bu grafik λ değiştikçe katsayı tahminlerinin nasıl değiştiğini göstermektedir
- Dikey eksen: standardize edilmiş ridge katsayı tahminleri
- Yatay eksen: λ ayarlama parametresi
- $\lambda = 0$: OLS katsayıları
- λ büyüdükçe katsayılar küçülmektedir; limitte tüm katsayılar 0 olur.

(ISLR Fig-6.4, p.216)

Ridge regresyonunda sapma-varyans ilişkisi



(ISLR Fig-6.5, p.218)

- Simülasyon verileri ile edilen grafikte λ ile ortalama hata karesi arasındaki ilişki gösteriliyor.
- MSE (mor) = Sapmakare (siyah) + Varyans (yeşil) + İndirgenemez hata varyansı (kesikli yatay)
- $\lambda = 0$ iken sapma çok küçük ancak varyans yüksek.
- $\lambda \approx 10$ değerine kadar MSE hızlı bir şekilde azalıyor, sapmada da bir artış var ancak çok fazla değil.
- $\lambda = 30$ için MSE en küçük.

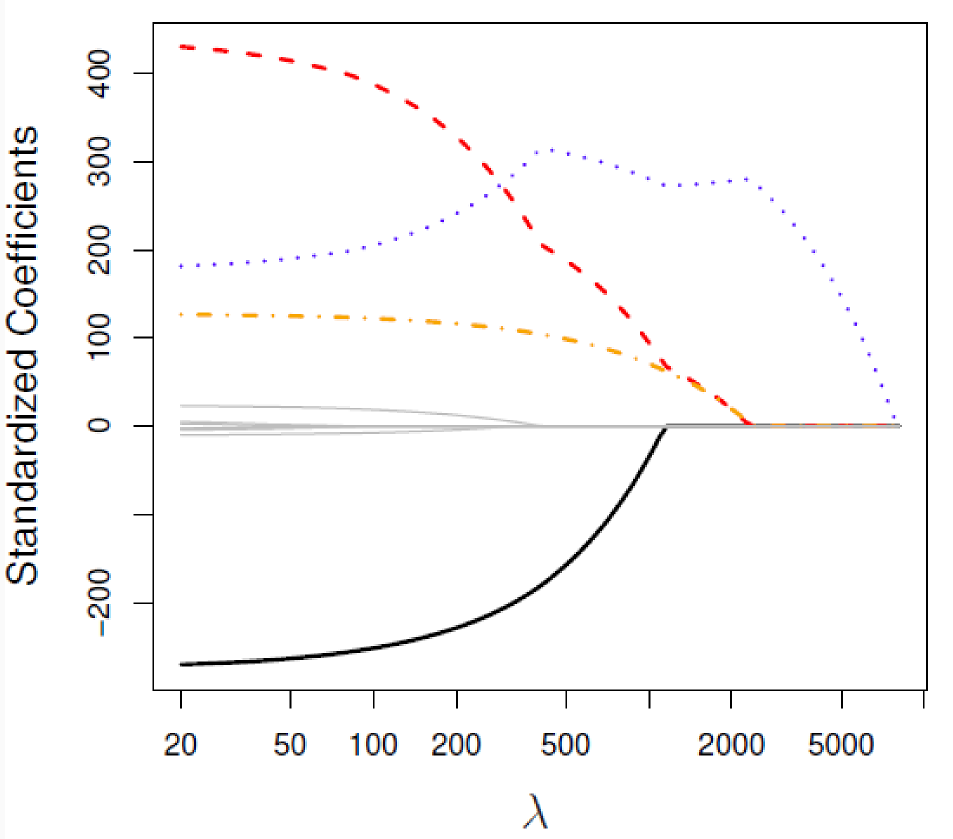
LASSO

- Çıkıntı regresyonunun en önemli zaafı tüm değişkenlerin modelde yer almasıdır (katsayıları küçük de olsa). Model katsayıları tam olarak $\beta = 0$ olmaz ($\lambda = \infty$ değilse).
- Eğer amacımız değişkenlerin seçimi ise ridge regresyonu uygun olmayabilir.
- Örneğin Credit veri setinde Balance için kurduğumuz model 10 değişkenin hepsini içerecektir. Ancak bunların içinde bazıları diğerlerinden daha önemli olabilir (income, limit, rating, student).
- Alternatif: LASSO (Least Absolute Shrinkage and Selection Operator)
- Tıpkı Ridge regresyonu gibi LASSO regresyonu da OLS amaç fonksiyonuna bir ceza terimi ekler:

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

- LASSO'nun en önemli farkı bazı değişkenlerin katsayılarını sıfıra eşitleyerek **değişken seçimi** yapabilmesidir.

LASSO Örnek: Credit data

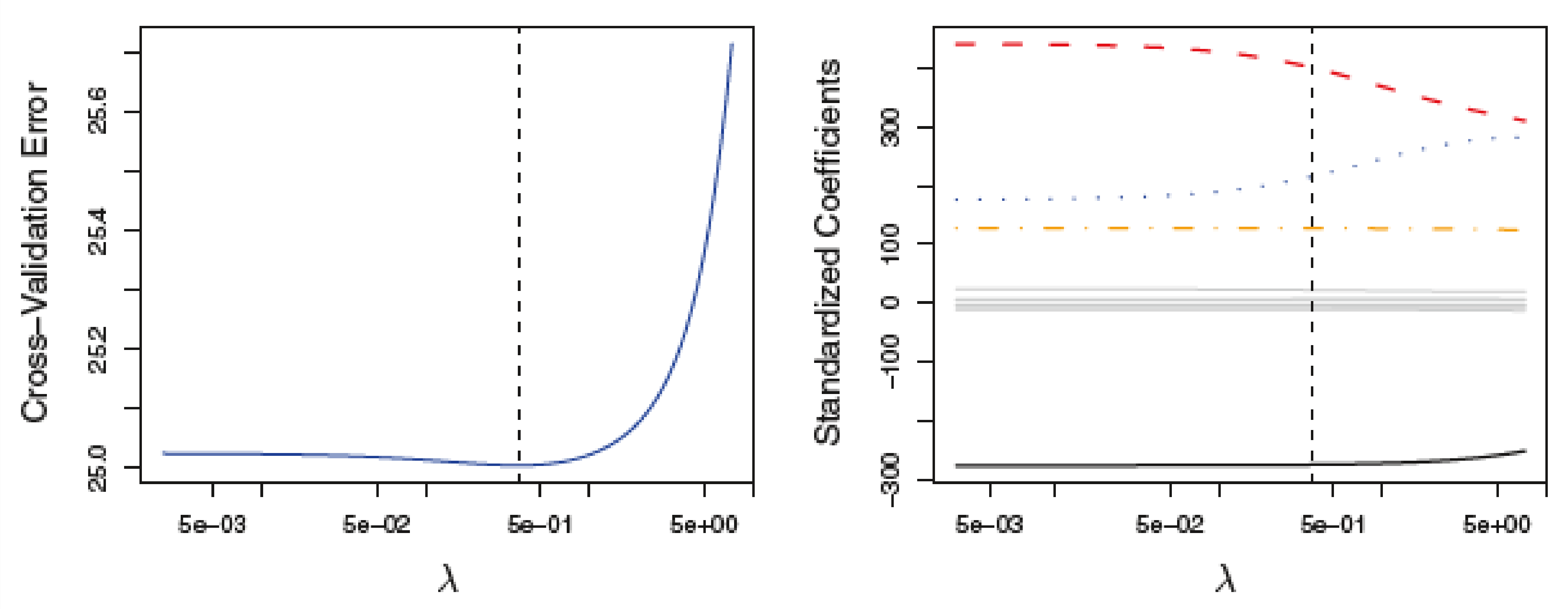


- $\lambda = 0 \rightarrow$ OLS
- $\lambda \rightarrow \infty$ tüm katsayılar 0 (null model)
- Ara değerler için bazı katsayılar 0.
- Bazı değişkenler modelden dışlanıyor.

(ISLR Fig-6.6, p.220)

Ayarlama parametresinin seçimi

- λ ayarlama parametresi çapraz geçerleme (cross validation) ile seçilebilir
- Önce λ için bir kesikli değerler kümesi (grid) belirlenir.
- Daha sonra her bir λ_j değeri için çapraz geçerleme hatası hesaplanır.
- En küçük çapraz geçerleme hatasını veren λ değeri seçilir.
- Son olarak, seçilen λ parametresi ile model tahmin edilir.



Elastik Net

- Zou ve Hastie (2005) ridge ve LASSO regresyonlarını özel durum olarak barındıran bir model önermiştir.
- Naif elastik net aşağıdaki fonksiyonu en küçük yapacak şekilde katsayıları seçer:

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda_1 \sum_{j=1}^{p}\beta_j^2 + \lambda_2 \sum_{j=1}^{p}|\beta_j| = \mathrm{SSR} + \lambda_1 \sum_{j=1}^{p}\beta_j^2 + \lambda_2 \sum_{j=1}^{p}|\beta_j|$$

- Naif yaklaşım: iki adımlı tahmin, önce Verilmiş bir λ_2 değeri için ridge regresyonunu tahmin et; ikinci adıma LASSO uygula.
- Ancak bu yöntem iki kere küçültme yaptığı için kestirim performansı başarılı değildir.
- Zou ve Hastie naif yaklaşım yerine alternatif bir tahmin çerçevesi önermiştir.