Düzenlileştirme

(İktisatçılar İçin) Makine Öğrenmesi (TEK-ES-2021)

Hüseyin Taştan
Yıldız Teknik Üniversitesi

Plan

- Doğrusal Model Seçimi
- Değişken altkümelerinin seçimi
- Düzenlileştirme (Regularization) kavramı
- Çıkıntı regresyonu (Ridge regression)
- LASSO
- Elastik Net
- Double-selection LASSO

Doğrusal model seçimi

- Pratikte genellikle bir başlangıç noktası olan doğrusal modeller çeşitli şekillerde geliştirilebilir.
- Bunlardan biri modelin doğrusal (additive) yapısını bozmadan doğrusal olmayan ilişkilerin eklenmesidir (örneğin polinom terimleri, splinelar, GAM). Modelin parametrelerde doğrusal yapısı aynı kaldığından en küçük kareler yöntemi uygulanabilir.
- Diğer bir yaklaşım ilgisiz kestirim değişkenlerin dışlanarak modelin performansının iyileştirilmesidir.
- Değişken sayısının, p, çok fazla olduğu durumlarda özellikle faydalı olabilir. Yaygın kullanılan yöntemler şunlardır:
    - Altküme seçimi
    - Düzenlileştirme yaklaşımı (Shrinkage or regularization): bu yaklaşımda tüm değişkenleri dikkate alınır ancak bazılarının katsayıları sıfıra doğru küçültülür.
    - Boyut küçültme.

Test hatasının dolaylı tahmini

- Bir doğrusal modelde değişken sayısın artarken SSR küçülür ve R^2 artar. Bu nedenle bu iki istatistik, SSR ve R^2 değerlendirme kıstası olamazlar.
- Dolaylı yaklaşımda eğitim verilerinden elde edilen hata baz alınır ve aşırı uyumdan kaynaklanan sapma nedeniyle bir ceza terimi eklenir. Elde edilen bu ölçüt model karşılaştırmasında kullanılabilir. Yaygın olarak kullanılan ölçütler şunlardır:
- Mellow's C_p

$$C_p = \frac{1}{n}\left(\text{RSS} + 2d\hat{\sigma}^2\right)$$

Burad d modeldeki toplam parametre sayısı ve $\hat{\sigma}^2$ hata varyansının bir tahminidir.

Test hatasının dolaylı tahmini

- Akaike's Information Criterion (AIC):

$$\text{AIC} = -2 \log L + 2 \cdot d$$

  Buarada L logolabilirlik fonksiyonunun maksimum değeridir.

- Bayesion Information Criterion (BIC):

$$\text{BIC} = \frac{1}{n}\left(\text{RSS} + \log(n) d \hat{\sigma}^2\right)$$

- C_p ve AIC ölçütleri doğrusal modellerde normal dağılım varsayımı altında aynıdır.

- Test hatasının küçük olduğu modellerde bu ölçütler de küçük değerler alma eğilimindedir. Bu nedenle farklı modelleri karşılaştırırken en küçük C_p, AIC, veya BIC değerine sahip olan seçilir.

- BIC daha büyük bir ceza terimi uyguladığından AIC'ye göre daha küçük modeller seçme eğilimindedir.

Test hatasının dolaylı tahmini

- Düzeltilmiş (Adjusted) R^2

$$\text{Adjusted } R^2 = 1 - \frac{\text{RSS}/(n-d-1)}{\text{TSS}/(n-1)}$$

  Burada RSS kalıntı kareleri toplamı ve TSS toplam kareler toplamıdır.

- Adjusted R^2 model daha karmaşık hale geldikçe azalabilir ya da artabilir. Gereksiz değişkenler eklendiğinde azalır.

- Büyük bir Düzeltilmiş R^2 değerine sahip bir modelin test hatası küçük olur. Bu nedenle büyük değerlere sahip modeller tercih edilir.

Test hatasının dolaysız tahmini

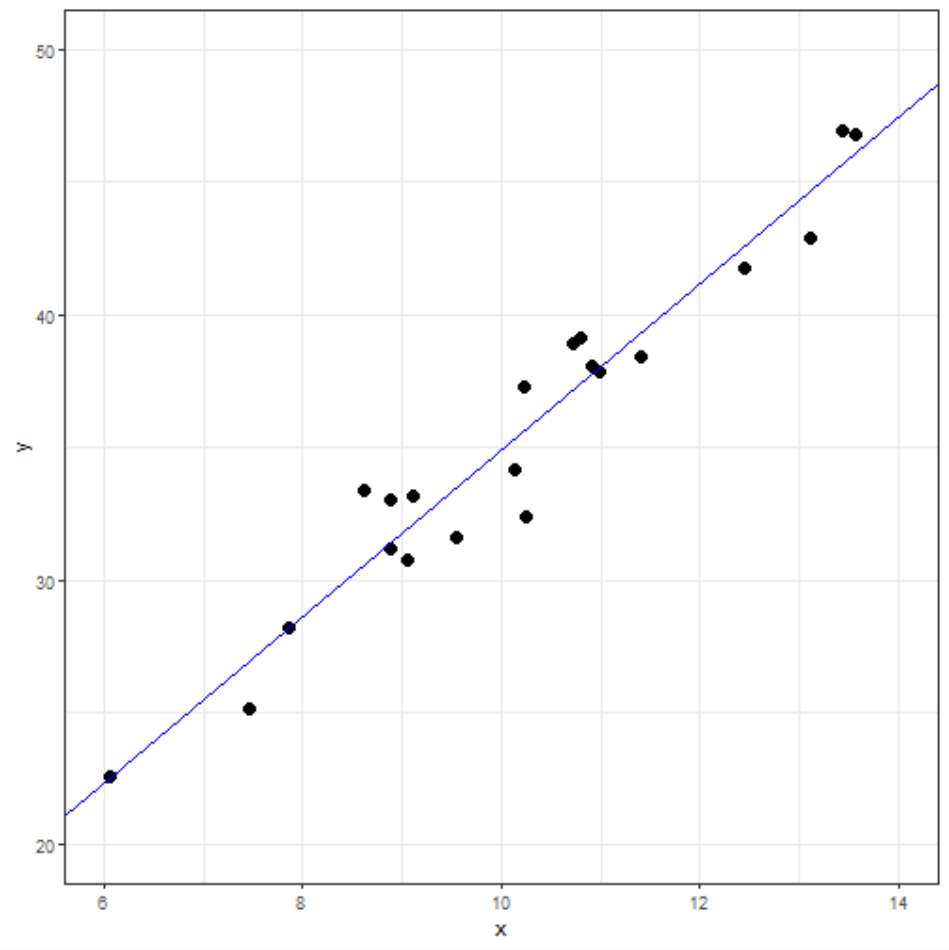
- Test hatası geçerleme ve çapraz-geçerleme yaklaşımlarıyla doğrudan tahmin edilebilir.
- Bunun için veriler eğitim ve geçerleme olmak üzere iki parçaya ayrılır. Eğitim verileriyle model tahmin edilirken sadece geçerleme verileriyle test hatası tahmin edilir.
- $\mathcal{M}_k$ ile gösterilen model büyüklüğü $k = 0, 1, 2, \ldots$, ile endekslenmiş bir model kümesi için geçerleme ve çapraz geçerleme hatası tahmin edilir.
- Bunların arasından en küçük test hatasına sahip olan tercih edilir.
- Bu yöntemde test hatasının, σ^2, tahmin edilmesi gerekmez.

Düzenlileştirme (Regularization)

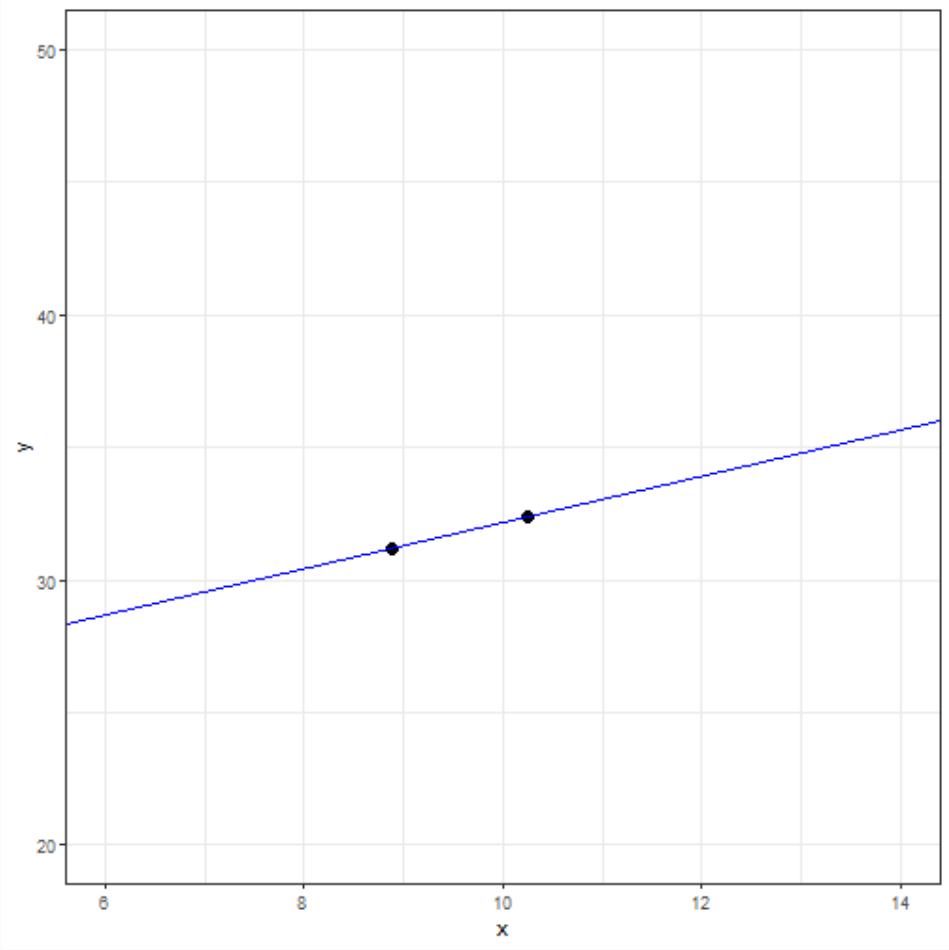
- Sıradan En Küçük Kareler (OLS) yöntemi Gauss-Markov varsayımları altında sapmasız/tutarlı ve en düşük varyanslı (etkin) tahminciler verir.
- Gözlem sayısının (n) değişken sayısından (p) çok daha büyük olduğu örtük olarak varsayılır: $n >> p$
- $n = p$ ise OLS tahmini **tam uyum** ile sonuçlanır.
- $p > n$ 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.

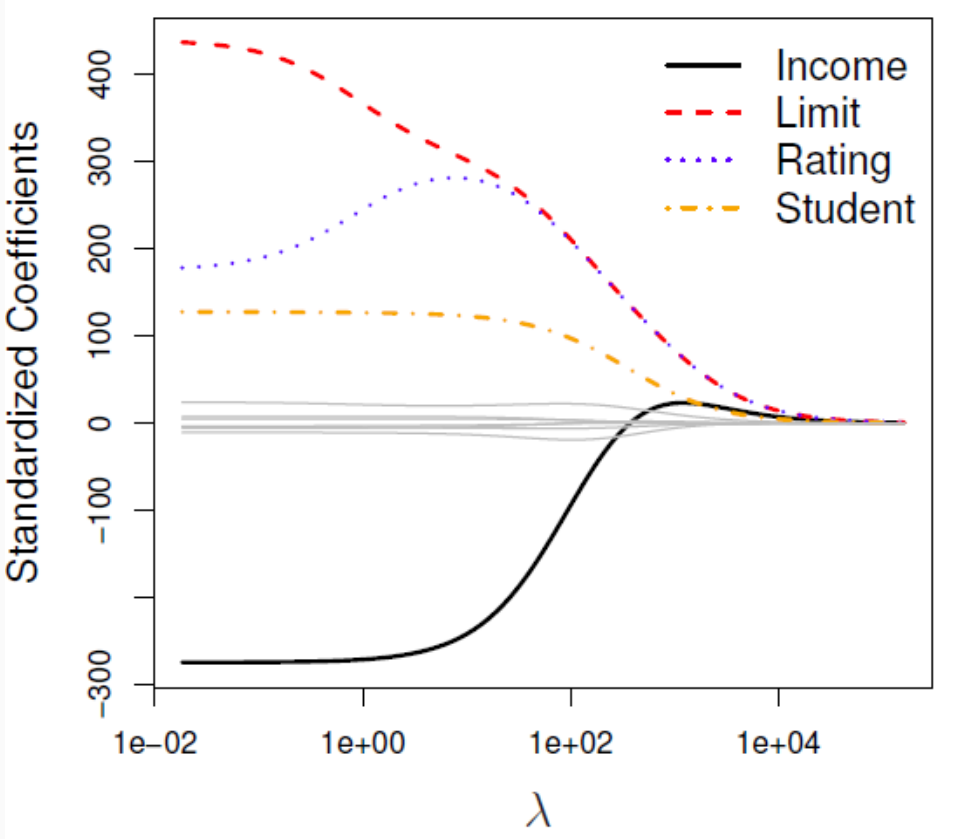
Örnek

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):

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

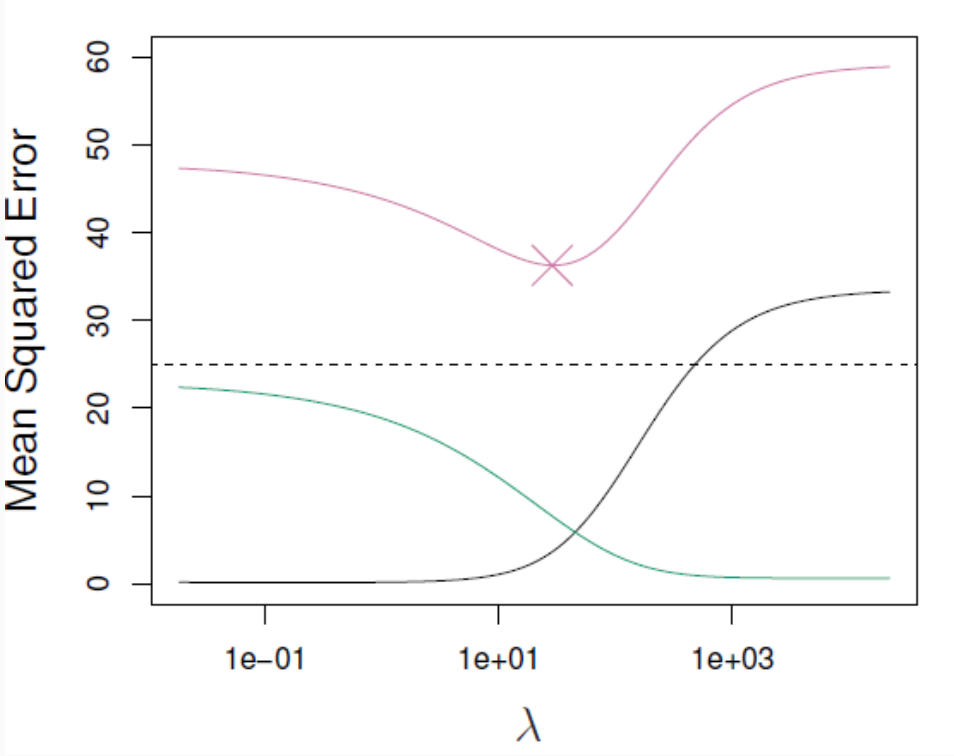
Örnek: Credit data



(ISLR Fig-6.4, p.216)

- 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.

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:

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

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

Ridge ve LASSO'nun geometrik yorumu

- LASSO optimizasyon problemi aşağıdaki gibi yazılabilir:

$$\underset{\beta}{\text{minimize}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \quad \text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \leq s$$

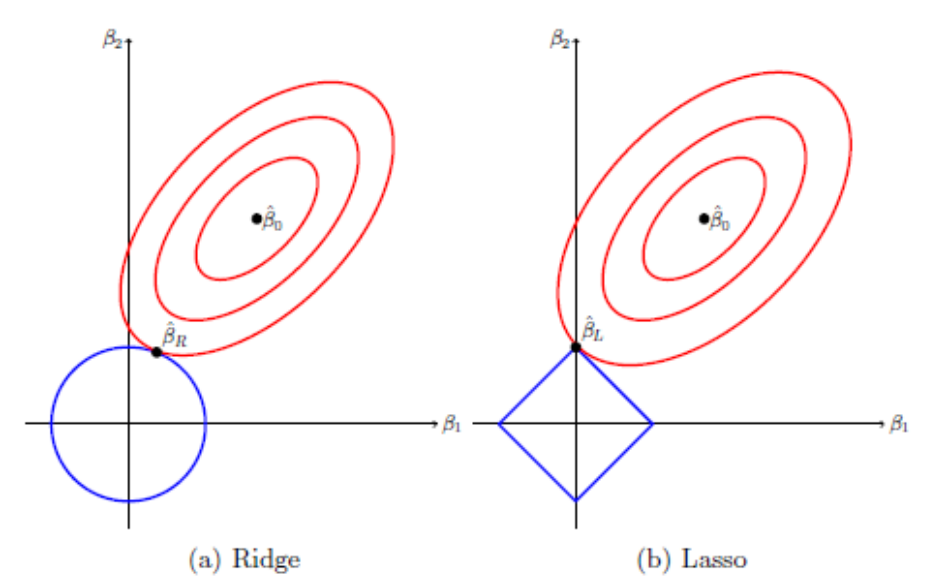
- Ridge regresyonu için optimizasyon problemi

$$\underset{\beta}{\text{minimize}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \quad \text{subject to} \quad \sum_{j=1}^{p} \beta_j^2 \leq s$$

  burada s kullanıcı tarafından tanımlanan ayarlanma parametresidir.

- Bu kısıtlanmış optimizasyon problemi izleyen grafiklerdeki gibi görselleştirilebilir. Basitlik amacıyla iki parametreli bir model varsayılmıştır.

Ridge ve LASSO'nun geometrik yorumu

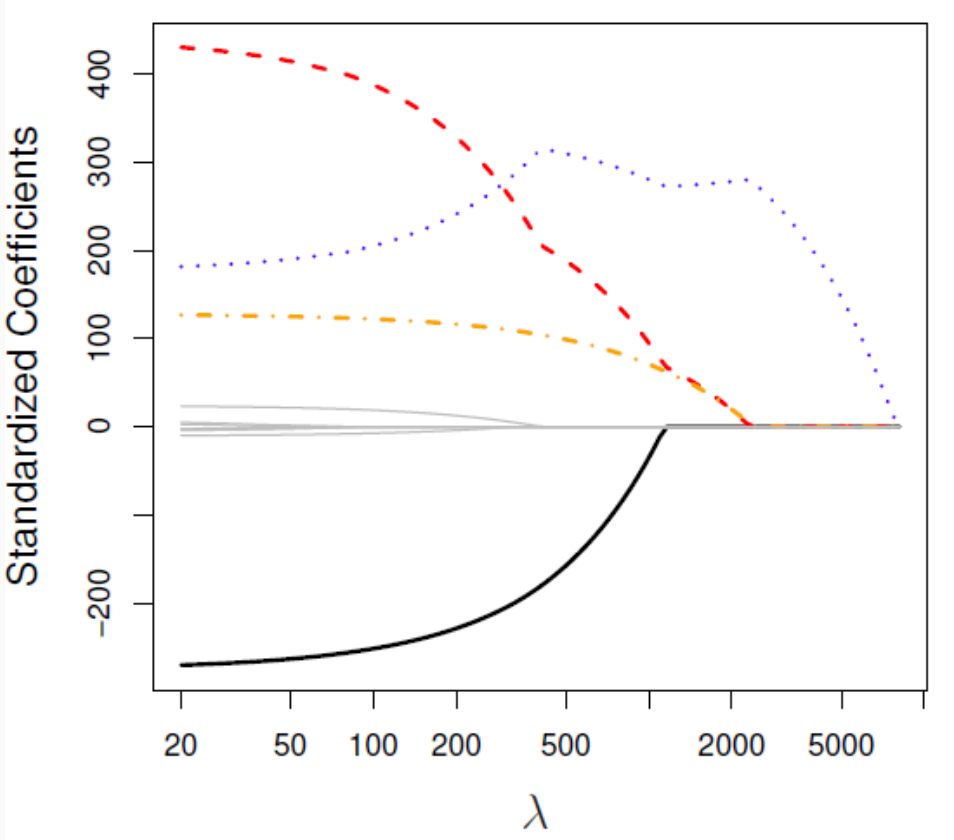


(a) Ridge (b) Lasso

- Ridge kısıt kümesi: $\beta_1^2 + \beta_2^2 \leq s$
- LASSO kısıt kümesi: $|\beta_1| + |\beta_2| \leq s$

- Eliptik çizgiler (kırmızı) kalıntı kareleri toplamı kontür çizgileridir
- Mavi çizgiler Ridge ve LASSO kısıtlarıdır.
- Dikkat edilirse Ridge regresyonunun kısıt kümesi parametreleri küçültürken hiçbirini sıfır yapmaz.
- Diğer taraftan,LASSO kısıt seti bazı parametrelerin tam olarak sıfıra eşitlenmesiyle sonuçlanabilir.

LASSO Örnek: Credit data

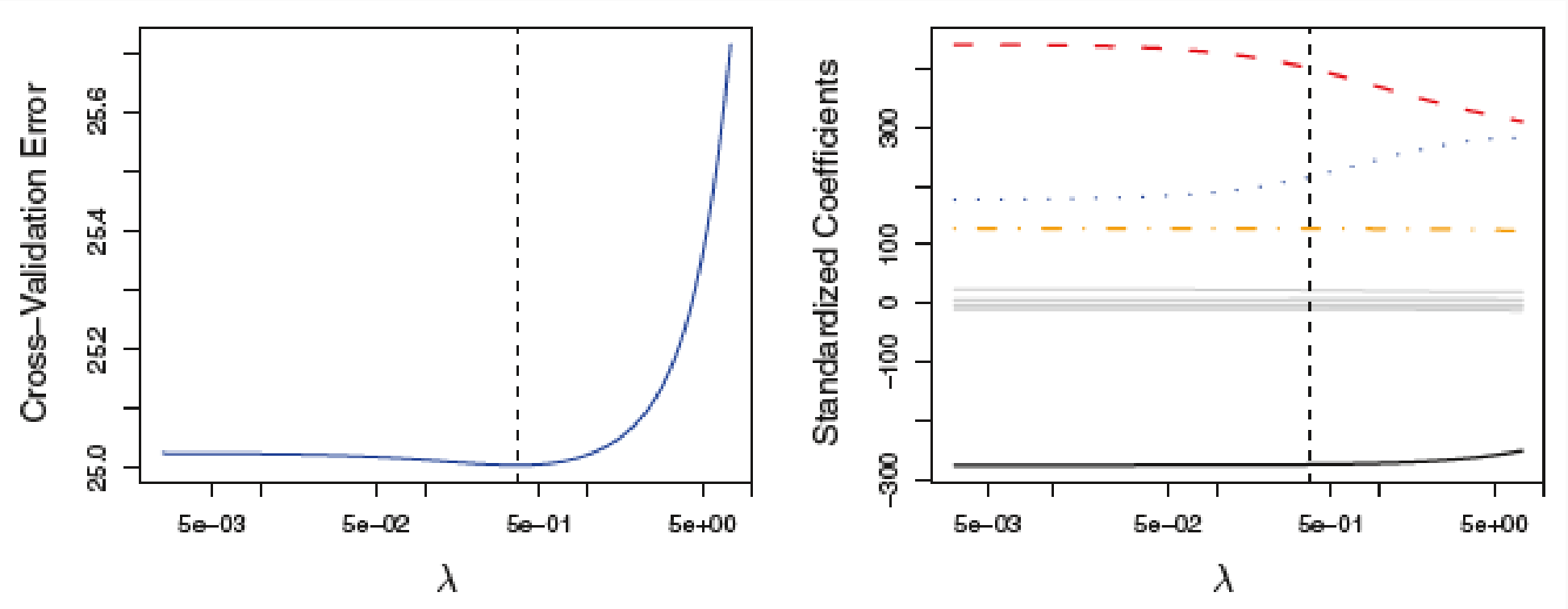


- $\lambda = 0 \rightarrow$ OLS
- $\lambda \rightarrow \infty$ tüm katsayılar 0 (null model)
- Ara değerler için bazı katsayılar 0.
- Bazı değişkenler modelden dışlanıyor.

(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:

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

- 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.

Elastic Net

- Amaç fonksiyonu aşağıdaki gibi yazılabilir:

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

  $0 \leq \alpha \leq 1$.

- $\alpha = 0$ olduğunda ridge regresyonu, $\alpha = 1$ olduğunda ise LASSO elde edilir.

- Aslında elastik net bu iki arasında bir yerdedir. Bire yakın α değerleri için elastik net LASSO'ya benzer davranış gösterir. Ancak özellikle değişkenler arasında çok yüksek korelasyonun olduğu durumlarda LASSO'ya göre daha iyi bir performans gösterebilir.

- `R` kütüphanesi `glmnet` kullanıcıların α parametresini seçmesine izin verir. Bkz. An Introduction to glmnet

Post-double selection LASSO

- LASSO'nun model seçiminde kullanılabileceğini öğrendik. Ancak model seçiminde iktisat teorisi de önemlidir. Örneğin, teorik modelden hareketle oluşturulan aşağıdaki modeli düşünelim:

$$y = \alpha\, d + X\beta + \epsilon$$

  burada y tepki değişkeni, d potansiyel olarak içsel (endogenous) değişken (treatment variable) ve X ise p kontrol değişkenini gösteren matristir. ($p>>n$ olabilir).

- Kontrol değişkenlerini, x, nasıl seçmeliyiz? Tek adımlı seçim prosedürü tipik olarak şöyle çalışır: herhangi bir x_j değişkeni LASSO ve t-testi sonuçlarına göre anlamlıysa modele dahil edilir (post single-selection procedure).

- Belloni, Chernozukov ve Hansen (2014), parametrelerin sıfıra yakın (ancak tam olarak sıfır değil) olduğu durumda bu yaklaşımın başarısız olacağını göstermiştir.

Değişken seçimi

- Örnek olarak Acemoglu, Johnson, ve Robinson (2001) tarafından incelenen kurum kalitesi ve ekonomik büyüme performansı ilişkisine bakalım:

$$log(GDPpc) = \alpha\, Q + X\beta + \epsilon$$

burada Q kurum kalitesini gösteren bir değişkendir. Kontrol değişkenleri arasında coğrafi ve ülkeye özgü bir çok faktör yer almaktadır.

- Bu modelde Q içseldir. Acemoglu et al. (2001) erken dönem yerleşimcilerinin yaşam süresini araç değişken olarak kullanmıştır.

- Ancak hangi X değişkenlerinin modele eklenmesi gerektiği açık değildir. İki adımlı LASSO seçimi bu konuda yardımcı olabilir.

Double-Selection LASSO

- Belloni, Chernozhukov, ve Hansen tarafından önerilmiştir.
- Algoritma adımları: (daha detaylı bilgi için bkz. How to do model selection with inference in mind)

Adım 1: LASSO (veya başka bir düzenlileştirme yöntemini) kullanarak y'yi en iyi kestiren x_j değişkenlerini bul.

Adım 2: LASSO (veya başka bir düzenlileştirme yöntemini) kullanarak d'yi en iyi kestiren x_j değişkenlerini bul.

Adım 3: Her iki adımda seçilen değişkenleri kullanarak OLS ile modeli yeniden tahmin et

- Dikkat edilirse değişken seçimi iki kere uygulanır: ilk adımda y tepki değişkenidir, ikinci adımda ise d tepki değişkenidir. Burada amaç dışlanmış değişken sapmasını en küçük yapmaktır.

Boyut küçültme yöntemleri

- Bu yöntemler kestirim değişkenlerinin bir dönüştürmesini bularak modelin boyutu küçültür.
- $Z_1, Z_2, \ldots, Z_M$, p kestirim değişkeninin $M < p$ doğrusal kombinasyonu olsun:

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

  Burada $\phi_{m1}, \ldots, \phi_{mp}$ doğrusal kombinasyon katsayılarıdır.

- $M < p$ olduğundan model sıradan en küçük kareler yöntemi ile tahmin edilebilir:

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

Temel bileşenler regresyonu

- Temel Bileşenler Regresyonu (Principal Components Regression (PCR)) yüksek boyutlu kestirim değişkenleri kümesinin bir lineer kombinasyonunu baz alır.
- Temel Bileşenler Analizi (PCA) birbiriyle ilişkisiz doğrusal kombinasyonların bulunmasının bir yöntemini sunar. Bir veri kümesinde değişken sayısı kadar temel bileşen bulunur.
- Bunlardan birinci temel bileşen verilerdeki değişkenliği en fazla açıklayan bileşendir. Bu bileşen bulunduktan sonra doğrusal regresyonda kullanılabilir.