Introduction to Machine Learning CS182

Lu Sun

School of Information Science and Technology

ShanghaiTech University

October 15, 2023

Today:

- Linear Methods for Regression I
  - Linear regression models
  - The Gauss-Markov theorem
  - Subsets selection

Readings:

- The Elements of Statistical Learning (ESL), Chapters 3
- Pattern Recognition and Machine Learning (PRML), Chapter 3

Introduction

- A linear regression model assumes that,

Regression function
$\min_f \mathrm{EPE}(f)$

线性回归:假设f(x)是线性方程

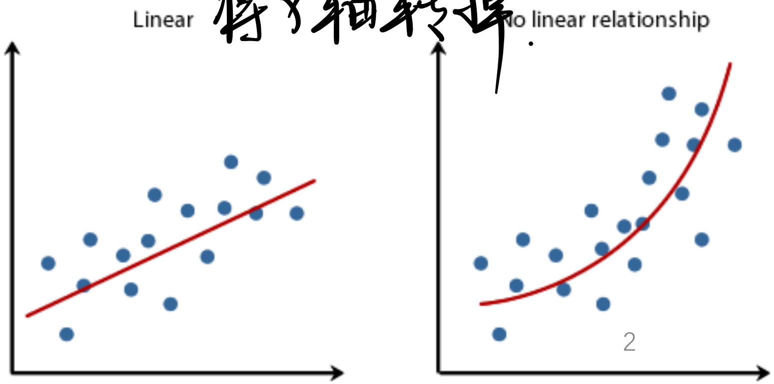
$$f(x) = \mathrm{E}(Y|X = x)$$

  - linear in the inputs $X_1, X_2, \ldots, X_p$.

- $p = 1 \rightarrow$ simple linear regression
- $p > 1 \rightarrow$ multiple linear regression

- Suitable for the situations:
  - small number of training samples
  - low signal-to-noise ratio
  - sparse data
- Generalize to many nonlinear techniques.

可以转换成线性.
将x轴转换.



Linear Methods for Regression

--- Linear Regression Models

Simple Linear Regression

- Training set: $(x_1, y_1), \dots, (x_N, y_N)$ 数据点
  - x_i: value of predictor X (covariate, independent variable, feature,…)
  - y_i: value of response Y (dependent variable, label,…) 标签.
- We denote the regression function by

$$f(x) = \mathrm{E}(Y|X = x)$$

  - conditional expectation of Y given x
- The linear regression model assumes a specific linear form

$$f(x) = \beta_0 + \beta x$$

  - usually thought of as an approximation to the truth

给底法，但β_0不涉及X，故能法采用.

去掉β_0，为正则化打基础.

Simple Linear Regression

单变量时尽量从 $\hat{\beta}_0$ 入手.
因为如果要正则化则不可含 $\hat{\beta}_0$.

- Fitting the model by least squares

the values of β_0, β for which RSS(β_0, β) attains it's minimum.

$$\hat{\beta}_0, \hat{\beta} = \boxed{\operatorname{argmin}_{\beta_0,\beta}} \sum_{i=1}^{N} (y_i - \beta_0 - \beta x_i)^2$$

求导为0

- Solutions are

$$\hat{\beta} = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \boxed{\bar{y}})}{\sum_{i=1}^{N} (x_i - \boxed{\bar{x}})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}\bar{x}$$

↳相当于把均值移除，过(0,0)点，然后可求导求 $\hat{\beta}$

sample mean:

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

$$\bar{y} = \frac{1}{N} \sum_{i=1}^{N} y_i$$

Q: How to get the solutions?

- $\hat{y}_i = \hat{\beta}_0 + \hat{\beta} x_i$ are called the *fitted* or *predicted* values
- $r_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta} x_i$ are called the *residuals*

假设：N个采样 相互独立，随机采样.

多元

Multiple Linear Regression

- Given $X=(X_1,X_2,\dots,X_p)^T$
- $\mathrm{E}(Y|X)$ is (approximately) linear:

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

- Sources of the variable X_j
  - quantitative inputs
  - transformation
  - basis expansions
  - dummy coding
  - interaction
- Linear in the parameters β

- Training data $(x_1,y_1),\dots,(x_N,y_N)$
- *Least squares*:

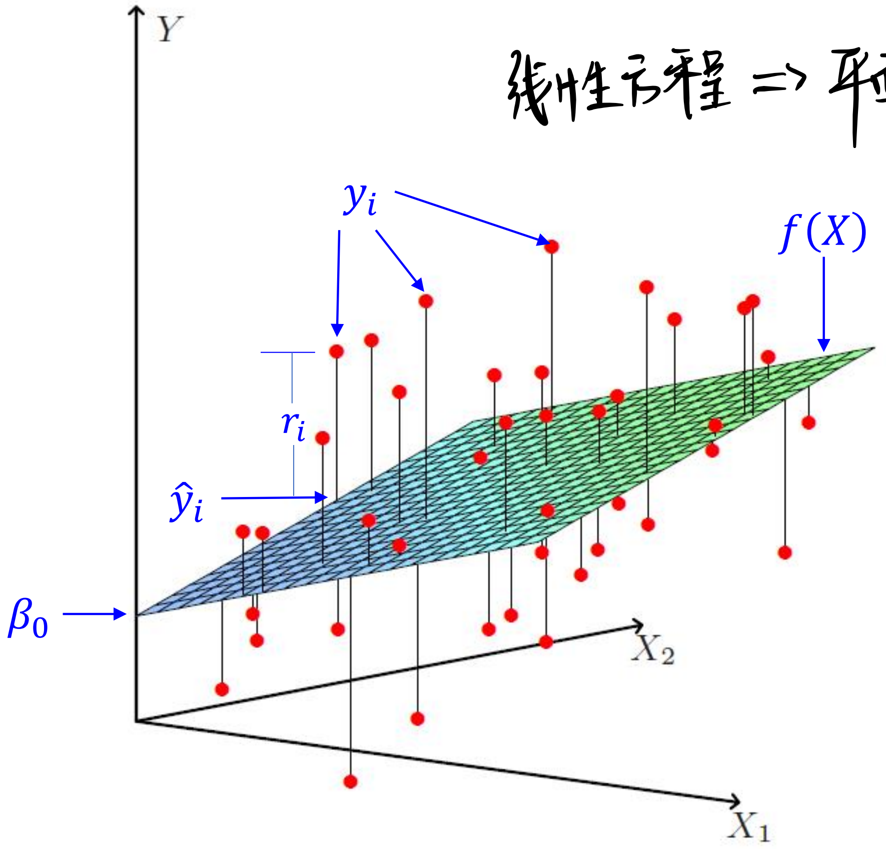
$$\mathrm{RSS}(\beta)=\sum_{i=1}^{N}(y_i-f(x_i))^2$$

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

- It is reasonable once
  - Observations (x_i,y_i) are randomly sampled from their population
  - Output y_i is conditionally independent w.r.t. the inputs x_i
- No guarantee on the validity of model

Multiple Linear Regression

线性方程 => 平面.



- Training data $(x_1, y_1), \ldots, (x_N, y_N)$
- *Least squares*:

$$\mathrm{RSS}(\beta) = \sum_{i=1}^{N} (y_i - \overbrace{f(x_i)}^{\hat{y}_i})^2$$

$$= \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\,\beta_j)^2$$

- It is reasonable once
  - Observations (x_i, y_i) are randomly sampled from their population
  - Output y_i is conditionally independent w.r.t. the inputs x_i
- No guarantee on the validity of model

Multiple Linear Regression

- Minimization of $\mathrm{RSS}(\beta)$
- Rewrite it by the vector form:

$$\mathrm{RSS}(\beta) = (\mathbf{y} - \mathbf{X}\beta)^T \boxed{(\mathbf{y} - \mathbf{X}\beta)}$$

- Differentiating w.r.t. β

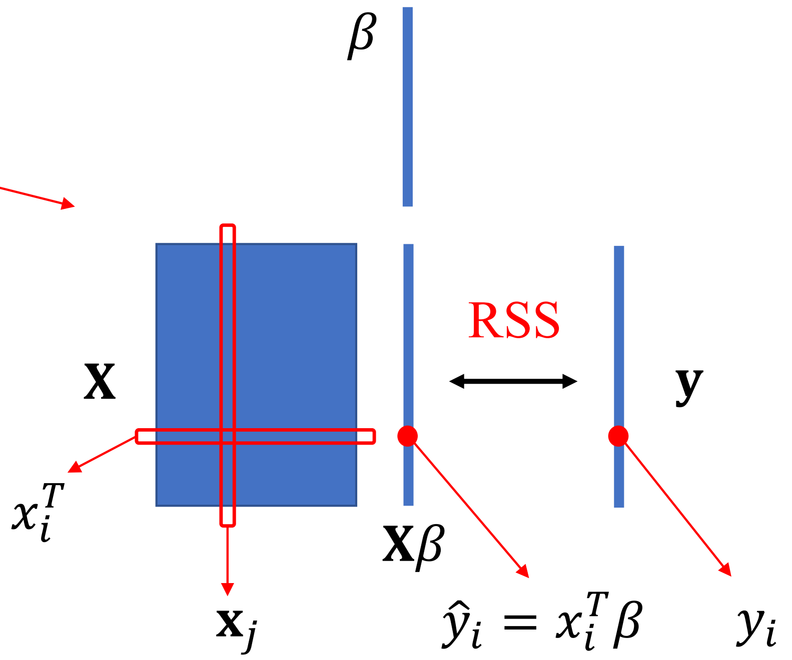
$$\frac{\partial \mathrm{RSS}}{\partial \beta} = -2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta)$$

- Set the first derivative to zero

$$\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta) = 0$$

- If $\mathbf{X}$ has full column rank,

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



Multiple Linear Regression

- Minimization of $\text{RSS}(\beta)$
- Rewrite it by the vector form:

$$\text{RSS}(\beta) = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta)$$

- Differentiating w.r.t. β

$$\frac{\partial \text{RSS}}{\partial \beta} = -2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta)$$

- Set the first derivative to zero

$$\boxed{\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta) = 0}$$

- If $\mathbf{X}$ has full column rank,

$$\hat{\beta} = \underbrace{(\mathbf{X}^T\mathbf{X})}_{}{}^{-1}\mathbf{X}^T\mathbf{y}$$

需要是满秩的(可逆)

从X空间向y的投影
结果为$\hat{y}$.

- Prediction on a test sample x_0

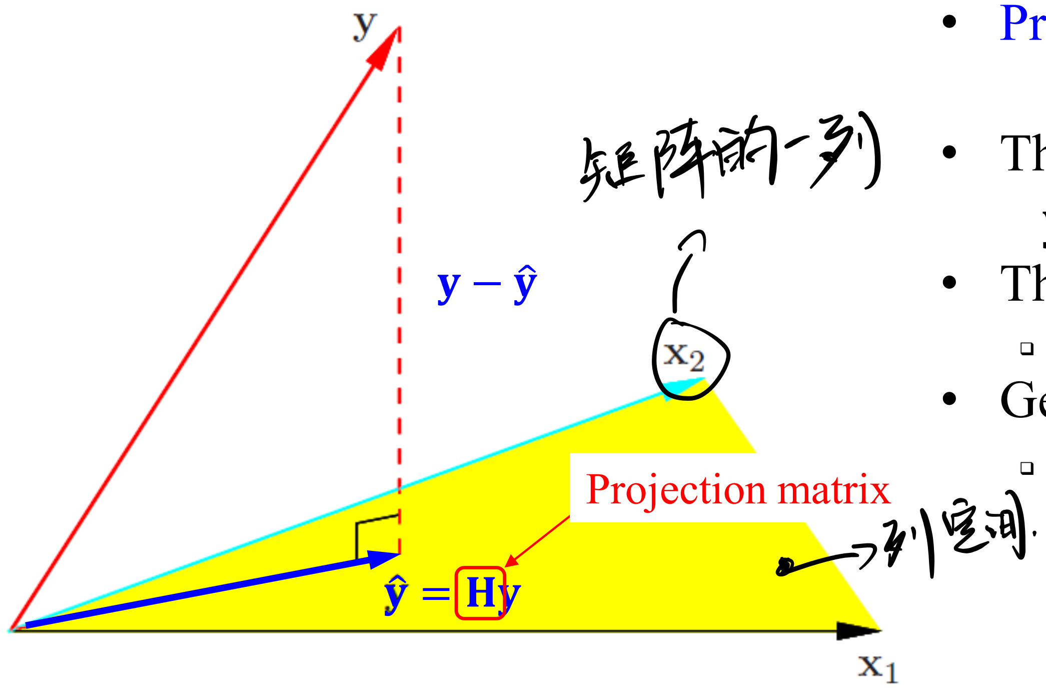
$$\hat{f}(x_0) = (1:x_0)^T\hat{\beta}$$

- The fitted values at the training inputs

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{H}\mathbf{y}$$

- The "hat" matrix $\mathbf{H}$
  - like a hat put on $\mathbf{y}$ 相当于是个投影.
- Geometrical interpretation
  - The optimal $\hat{\beta}$ makes the residual vector $\mathbf{y} - \hat{\mathbf{y}}$ orthogonal to the subspace spanned by the columns of $\mathbf{X}$

Multiple Linear Regression



$\mathbf{X} = (\mathbf{x_1}, \dots, \mathbf{x_p})$, where $\mathbf{x}_j = \left(x_{1j}, \dots, x_{Nj}\right)^T \in \mathbb{R}^N$

- Prediction on a test sample x_0

$$\hat{f}(x_0) = (1 \colon x_0)^T \hat{\beta}$$

- The fitted values at the training inputs

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{H}\mathbf{y}$$

- The "hat" matrix **H**
  - like a hat put on **y**
- Geometrical interpretation
  - The optimal $\hat{\beta}$ makes the residual vector $\mathbf{y} - \hat{\mathbf{y}}$ orthogonal to the subspace spanned by the columns of **X**

Multiple Linear Regression

$\boxed{B}$ $\boxed{A}$ $\boxed{AB}$

rank(col) = rank(row)

奇异性

On the singularity of $\mathbf{X}^T\mathbf{X}$

$\mathbf{X}^T\mathbf{X}$是p×p的矩阵

非奇异需要满秩，即rank($\mathbf{X}^T\mathbf{X}$)=p

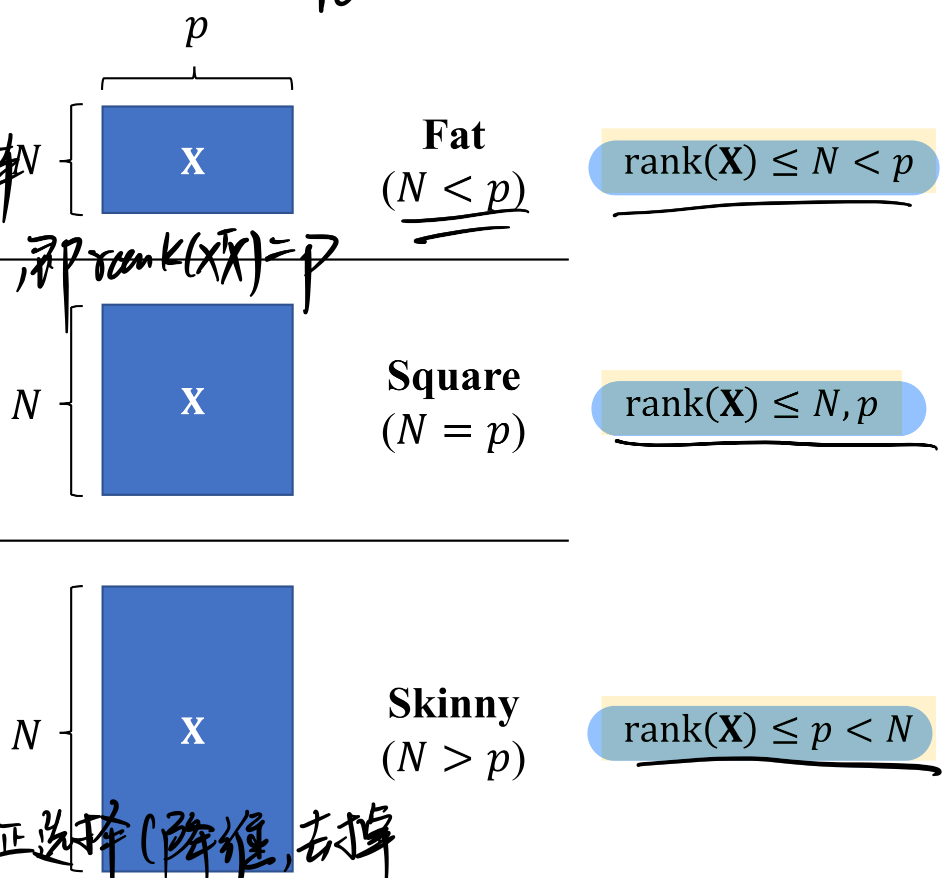
- *Fat* data matrix $\mathbf{X}$
  - singular

  X可能是奇异的

- *Square* data matrix $\mathbf{X}$
  - probably singular
  - nonsingular if $\text{rank}(\mathbf{X}) = p$

  需要有rank(X)=p

- *Skinny* data matrix $\mathbf{X}$
  - probably nonsingular
  - singular if $\text{rank}(\mathbf{X}) < p$



The solution $\hat{\beta}$ is unique once $\mathbf{X}^T\mathbf{X}$ is nonsingular ($\text{rank}(\mathbf{X}) = p$)

非满秩：有冗余信息，

信号：维度高，样本少

解决：特征选择（降维，去掉

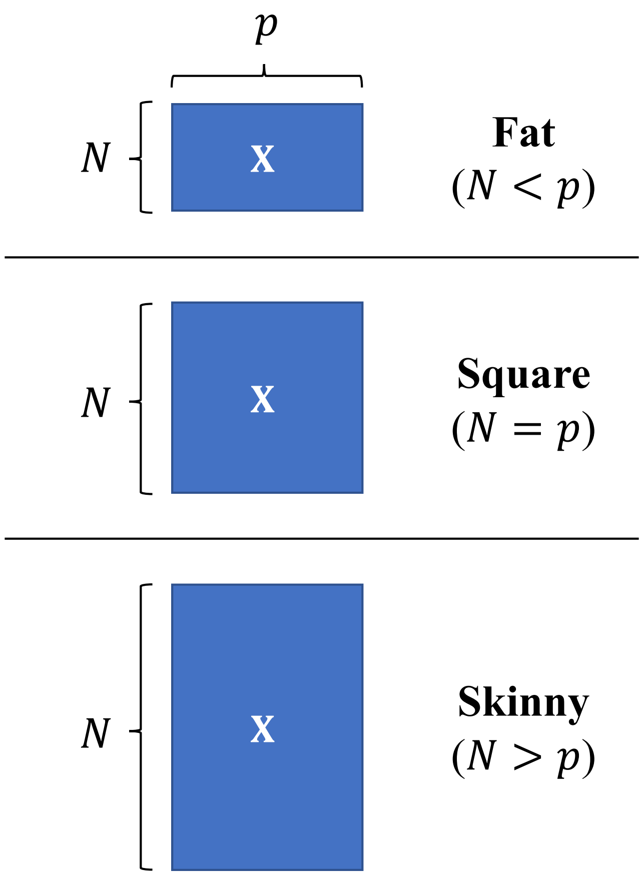
正则化：$\hat{\beta} = (X^TX + \lambda I)^{-1}X^Ty$

$X^TX+\lambda I$ 一定满秩. prove: $X^TX+\lambda I = \underbrace{Q\Lambda Q^{-1}}_{\text{奇异值分解}}+\lambda I = Q\Lambda Q^T+\lambda QQ^{-1}=Q(\Lambda+\lambda I)Q$

λ↑惩罚↑模型更简单. 所以要调整λ使模型最优

Multiple Linear Regression

- Rank deficient **X**
  - coding qualitative inputs
    - redundancy in columns of X
  - image and signal analysis
    - more features ($p > N$)
- Two ways to overcome it
  - feature selection (dimension reduction)
  - regularization



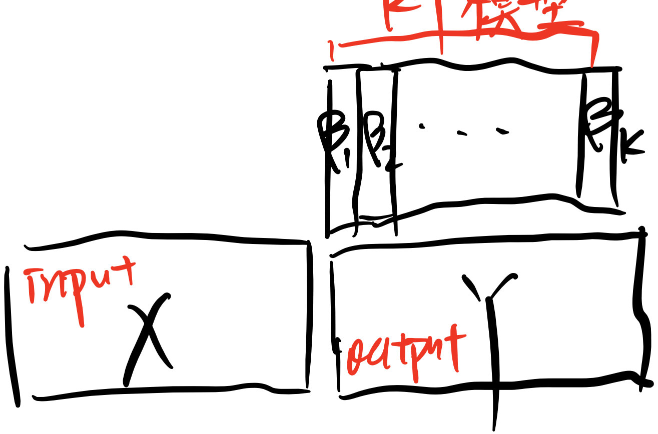
$\text{rank}(\mathbf{X}) \le N < p$

$\text{rank}(\mathbf{X}) \le N, p$

$\text{rank}(\mathbf{X}) \le p < N$

多输出.

Multiple Output Regression*



- Multiple outputs $Y_1, Y_2, \dots, Y_K$
- Assume a linear model for each output

$$Y_k = \beta_{0k} + \sum_{j=1}^{p} X_j \beta_{jk} + \varepsilon_k = f_k(X) + \varepsilon_k$$

现在变成矩阵:

- In matrix notation

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E}$$

有K个标签,K种模型
独立处理K个模型
存入Y中

where $\mathbf{X} \in \mathbb{R}^{N\times(p+1)}$, $\mathbf{B} \in \mathbb{R}^{(p+1)\times K}$ and $\mathbf{E} \in \mathbb{R}^{N\times K}$.

- A generalization of the univariate loss function

$$\mathrm{RSS}(\mathbf{B}) = \sum_{k=1}^{K}\sum_{i=1}^{N}\left(y_{ik} - f_k(x_i)\right)^2 = \|\mathbf{Y} - \mathbf{XB}\|_F^2$$

F-范数(矩阵的长度)

对角线之和(迹)

For an arbitrary matrix $\mathbf{A}$, the Frobenius-norm is defined by $\|\mathbf{A}\|_F^2 = \mathrm{Tr}(\mathbf{A}^T\mathbf{A}) = \sum_{ij} a_{ij}^2$.

Multiple Output Regression*

- Our problem:

$$\hat{\mathbf{B}} = \text{argmin}_{\mathbf{B}} \text{RSS}(\mathbf{B}) = \text{argmin}_{\mathbf{B}} \|\mathbf{Y} - \mathbf{XB}\|_F^2$$

- A quadratic function with global minimum 是一个点
- Rewrite $\text{RSS}(\mathbf{B})$ as follows

Matrix trace

$\mathbb{R}^{K \times K}$

$$\begin{aligned} \text{RSS}(\mathbf{B}) &= \boxed{\text{Tr}}\big((\mathbf{Y} - \mathbf{XB})^T(\mathbf{Y} - \mathbf{XB})\big) \\ &= \text{Tr}(\mathbf{Y}^T\mathbf{Y} - \boxed{\mathbf{Y}^T\mathbf{XB} - \mathbf{B}^T\mathbf{X}^T\mathbf{Y}} + \mathbf{B}^T\mathbf{X}^T\mathbf{XB}) \\ &= \text{Tr}(\mathbf{Y}^T\mathbf{Y}) - 2\text{Tr}(\mathbf{B}^T\mathbf{X}^T\mathbf{Y}) + \text{Tr}(\mathbf{B}^T\mathbf{X}^T\mathbf{XB}) \end{aligned}$$

- Differentiating w.r.t. $\mathbf{B}$

$$\frac{\partial \text{RSS}(\mathbf{B})}{\partial \mathbf{B}} = -2\mathbf{X}^T\mathbf{Y} + 2\mathbf{X}^T\mathbf{XB}$$

K次

- If $\mathbf{X}^T\mathbf{X}$ is nonsingular, $\hat{\mathbf{B}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \longrightarrow \hat{\beta}_k = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}_k, \forall k$

Multiple outputs do not affect one another's least squares estimates.

Linear Methods for Regression

--- The Gauss-Markov Theorem

The Gauss-Markov Theorem

但实际上，现实中无偏估计不多

对线性无偏估计中，最小二乘法有最小方差

> *The least squares estimator has the lowest sampling variance within the class of linear unbiased estimators.*

Proof: suppose $\tilde{\beta} = \mathbf{C}\mathbf{y}$ is a linear estimator of $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$,

where $\mathbf{C} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T + \mathbf{D}$, and $\mathbf{D} \in \mathbb{R}^{p\times N}$ is a non-zero matrix

噪声

⇒ 噪声的方差

$$
\begin{aligned}
\mathrm{E}[\tilde{\beta}] &= \mathrm{E}[Cy] \\
&= \mathrm{E}\left[\left((X'X)^{-1}X' + D\right)(X\beta + \varepsilon)\right] \\
&= \left((X'X)^{-1}X' + D\right)X\beta + \left((X'X)^{-1}X' + D\right)\boxed{\mathrm{E}[\varepsilon]} \\
&= \left((X'X)^{-1}X' + D\right)X\beta \\
&= (X'X)^{-1}X'X\beta + DX\beta \\
&= (I_p + \boxed{DX})\beta.
\end{aligned}
$$

$E[\varepsilon] = 0$

假设无偏

If and only if $\mathbf{DX} = 0$, $\tilde{\beta}$ is unbiased.

$$
\begin{aligned}
\mathrm{Var}(\tilde{\beta}) &= \mathrm{Var}(Cy) \\
&= C\,\boxed{\mathrm{Var}(y)}\,C' \\
&= \sigma^2 CC' \\
&= \sigma^2\left((X'X)^{-1}X' + D\right)\left(X(X'X)^{-1} + D'\right) \\
&= \sigma^2\left((X'X)^{-1}X'X(X'X)^{-1} + (X'X)^{-1}X'D' + DX(X'X)^{-1} + DD'\right) \\
&= \sigma^2(X'X)^{-1} + \sigma^2(X'X)^{-1}\boxed{(DX)}' + \sigma^2\boxed{DX}(X'X)^{-1} + \sigma^2 DD' \\
&= \sigma^2(X'X)^{-1} + \sigma^2 DD' \\
&= \boxed{\mathrm{Var}(\hat{\beta})} + \boxed{\sigma^2 DD'}
\end{aligned}
$$

$\mathrm{Var}(\mathbf{y}) = E[\mathbf{y} - E[\mathbf{y}]]^2 = \mathrm{Var}(\varepsilon)$

$\mathbf{DX} = 0$

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

Positive semidefinite

The Gauss-Markov Theorem

The least squares estimator has the lowest sampling variance within the class of linear unbiased estimators.

Proof: suppose $\tilde{\beta} = \mathbf{C}\mathbf{y}$ is a linear estimator of $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$,

where $\mathbf{C} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T + \mathbf{D}$, and $\mathbf{D} \in \mathbb{R}^{p \times N}$ is a non-zero matrix

$$
\begin{aligned}
\operatorname{Var}(\tilde{\beta}) &= \operatorname{Var}(Cy) \\
&= C \operatorname{Var}(y) C' \\
&= \sigma^2 CC' \\
&= \sigma^2 \left((X'X)^{-1}X' + D\right)\left(X(X'X)^{-1} + D'\right) \\
&= \sigma^2 \left((X'X)^{-1}X'X(X'X)^{-1} + (X'X)^{-1}X'D' + DX(X'X)^{-1} + DD'\right) \\
&= \sigma^2 (X'X)^{-1} + \sigma^2 (X'X)^{-1}(DX)' + \sigma^2 DX(X'X)^{-1} + \sigma^2 DD' \\
&= \sigma^2 (X'X)^{-1} + \sigma^2 DD' \\
&= \operatorname{Var}\left(\hat{\beta}\right) + \sigma^2 DD'
\end{aligned}
$$

Given an arbitrary test point x_0, we have

$$
\begin{aligned}
\operatorname{Var}(\tilde{y}_0) &= \operatorname{Var}\left(x_0^T \tilde{\beta}\right) \\
&= x_0^T \operatorname{Var}\left(\tilde{\beta}\right) x_0 \\
&= x_0^T \operatorname{Var}\left(\hat{\beta}\right) x_0 + \sigma^2 x_0^T \mathbf{D}\mathbf{D}^T x_0 \\
&= \operatorname{Var}(\hat{y}_0) + \sigma^2 \left(\underline{x_0^T \mathbf{D}\mathbf{D}^T x_0} > 0\right)
\end{aligned}
$$

The Gauss-Markov Theorem

The least squares estimator has the lowest sampling variance within the class of linear unbiased estimators.

Remarks

- Among the unbiased linear methods, least squares has the lowest MSE
  - MSE = Var + Bias^2
- A biased methods probably has lower MSE
  - Var-Bias trade-off
  - A small increase in Bias might gives rise to a large reduction in Var ← Model selection

Linear Methods for Regression

--- Subset Selection

Introduction

局限性:

Two limitations of least squares

- prediction accuracy 准确相对不高（相对线性回归）
  - low bias and high variance
    - → sacrifice a little bias to reduce the variance
- interpretation 解释性不足（无法解释哪种特征影响大）
  - hard to interpret a large number of input features
    - → find a subset of features exhibiting strong effects

模型选择.

We use model selection to overcome the limitations

- variable subset selection, shrinkage, dimension reduction.

  子集选择　正则化　维度降低
- not restricted to linear models

Subset Selection

- Best-subset selection，复杂度极高.
  - For each $s \in \{0,1,\dots,p\}$, find the subset in size of s that gives lowest $\mathrm{RSS}(\beta) = \|\mathbf{y} - \mathbf{X}^{(s)}\beta\|_2^2$

p>40时不可使用这种方法

$\binom{4}{2} = 6$

只是选了2个特征
还可选1个/3个/4个特征去训练.

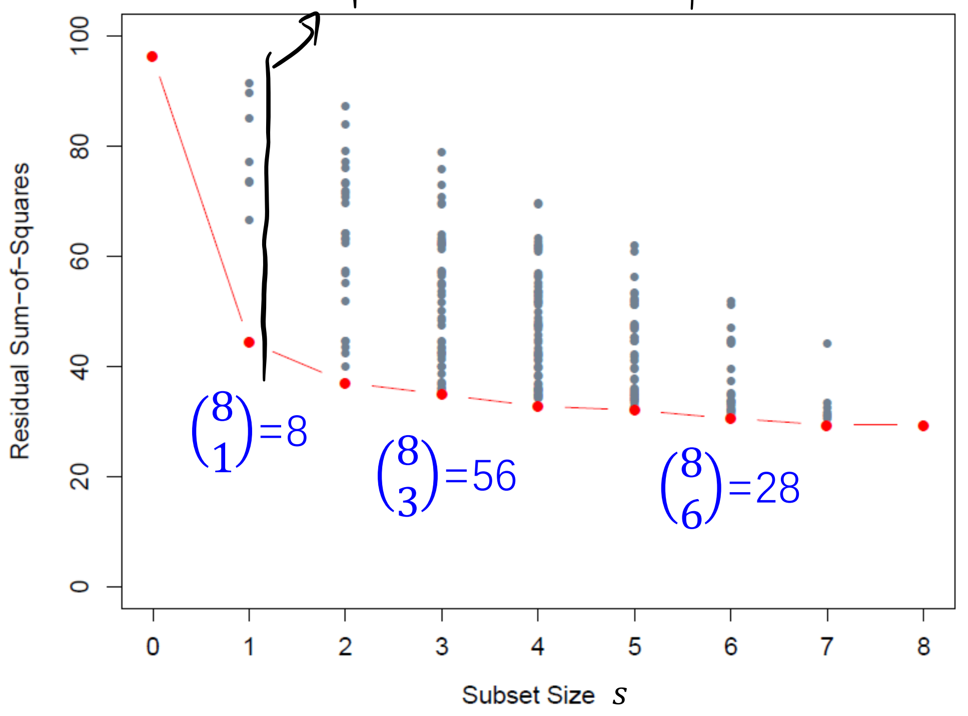
$p = 4$ $s = 2$	X_1	X_2	X_3	X_4	$\mathbf{X}^{(s)}$
Model 1	√	√	×	×	$(\mathbf{x}_1, \mathbf{x}_2)$
Model 2	√	×	√	×	$(\mathbf{x}_1, \mathbf{x}_3)$
Model 3	√	×	×	√	$(\mathbf{x}_1, \mathbf{x}_4)$
Model 4	×	√	√	×	$(\mathbf{x}_2, \mathbf{x}_3)$
Model 5	×	√	×	√	$(\mathbf{x}_2, \mathbf{x}_4)$
Model 6	×	×	√	√	$(\mathbf{x}_3, \mathbf{x}_4)$

Subset Selection

越小越好

8个点，标示每种的RSS值

- Best-subset selection
  - For each $s \in \{0,1,\ldots,p\}$, find the subset in size of s that gives lowest $\text{RSS}(\beta) = \|\mathbf{y} - \mathbf{X}^{(s)}\beta\|_2^2$
- Example
  - prostate cancer example ($p = 8$) <40
  - the red lower bound denotes the models eligible for selection
  - the red lower bound keeps decreasing (s = 8?)
  - *cross-validation* to estimate prediction error and select s
- Typically intractable for $p > 40$

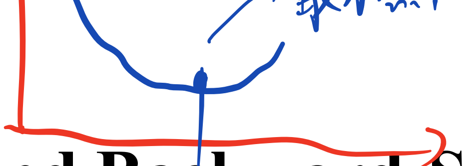


All the subset models for the prostate cancer example.

复杂度：$\hat\beta = (X^TX)^{-1}X^Ty$

$O(N^3)$ $O(p^3)$ $(N\times p)\times(p\times 1) = N\times 1$ 可整体看.

$O(p^3+Np^2)$



Forward- and Backward-Stepwise Selection

贪心：子集最优

- Forward-stepwise
  - starts with intercept
  - sequentially adds the best predictor
- Greedy algorithm
  - sub-optimal
- Advantages
  - Computational
    - even $p \gg N$
  - Statistical
    - constrained search 有约束的搜索
    - lower variance, more bias

Forward- and Backward-Stepwise Selection

- Forward-stepwise
  - starts with intercept
  - sequentially adds the best predictor
- Greedy algorithm
  - sub-optimal
- Advantages
  - Computational
    - even $p \gg N$
  - Statistical
    - constrained search
    - lower variance, more bias

扔掉

- Backward-stepwise
  - starts with the full model
  - sequentially deletes the worst predictor
- Greedy algorithm

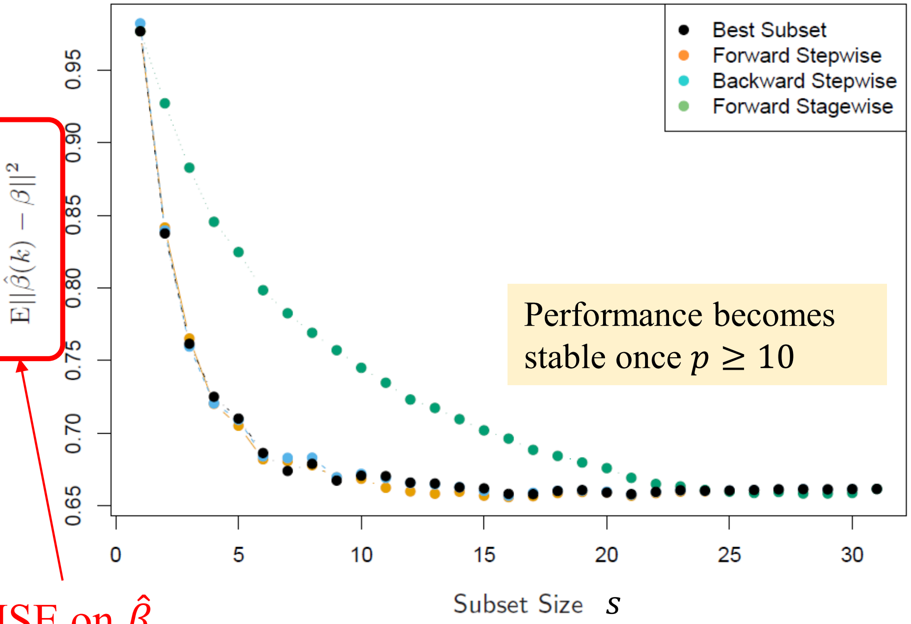
使用限制

- Only useful when $N > p$
  - linear regression

$N<P$时是奇异的.

- Smart stepwise
  - group of variables
  - add or drop whole groups at a time

Forward- and Backward-Stepwise Selection



MSE on $\hat{\beta}$

- Example
  - $Y = X^T\beta + \varepsilon$
  - $N = 300, p = 31$
  - only 10 variables are effective
  - similar performance

?

用1个做为验证

若极少，可以使用 1 ...
(极端情况)

从训练集中抽一部分做为 validation
用其它部分训练，用这个部分做 validation.
为防止 train 太少 使得 $N<P$ (Fat) K-Fold Cross-Validation

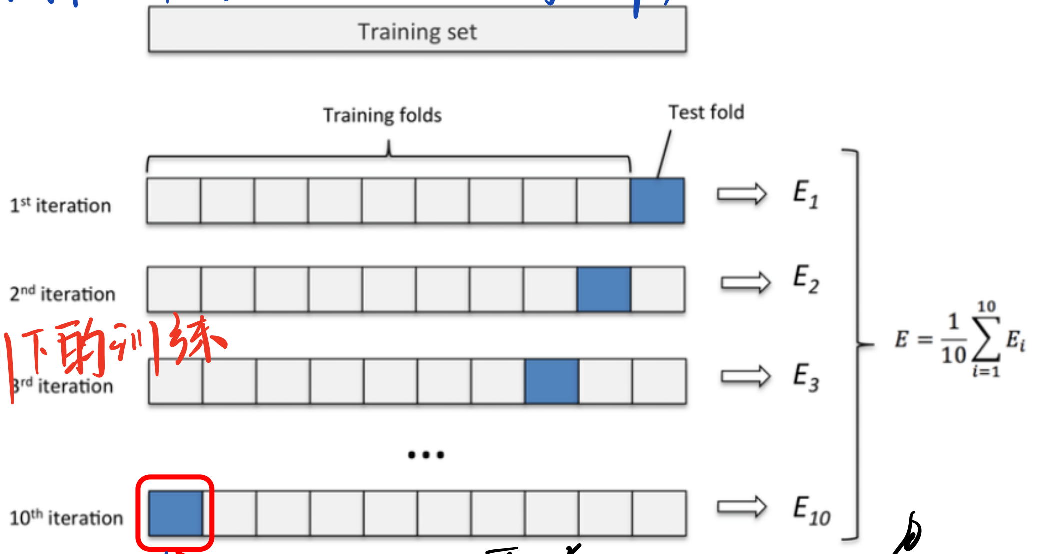
K-Fold Cross-Validation

成本：训练太多次.(时间成本)

- Each has a complexity parameter λ
  - the subset size in subset selection
  - the neighborhood size in k-NN
  - The coefficient of regularization
- K-fold cross validation 分成K等份
  - divide the training data into K roughly equal parts ($K = 5$ or 10)
  - for $k = 1, \ldots, K$, 用一份做验证，剩下的训练
    - fit the model with $K-1$ parts
    - compute the error E_k on the rest part
  - The K-fold cross validation error

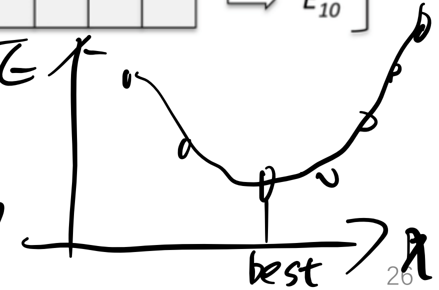
一共K等份
所以跑K次训练

$$E(\lambda) = \frac{1}{K}\sum_{k=1}^{K} E_k(\lambda)$$



validation error ← E

validation set



Repeat this for many values of λ, and choose the best value that makes $E(\lambda)$ lowest.