

Introduction to Machine Learning CS182

Lu Sun

School of Information Science and Technology

ShanghaiTech University

Feb. 22, 2022

Today:

- Linear Methods for Regression
 - 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,

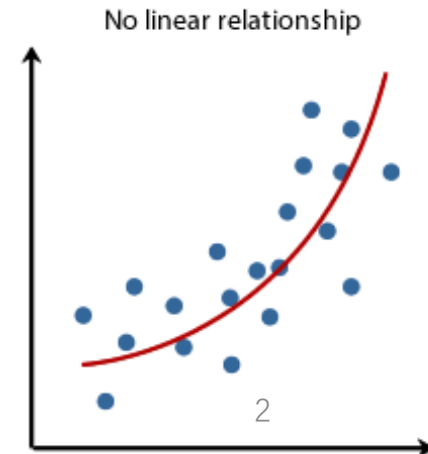
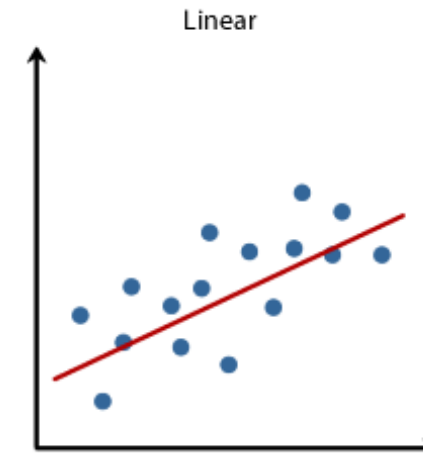
$$f(x) = E(Y|X = x)$$

Regression function

$$\min_f \text{EPE}(f)$$

- **linear** in the inputs X_1, X_2, \dots, X_p .
- Suitable for the situations:
 - small number of training samples
 - low signal-to-noise ratio
 - sparse data
- Generalize to many **nonlinear** techniques.

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



Linear Methods for Regression

--- Linear Regression Models

Simple Linear Regression

- **Training set:** $(x_1, y_1), \dots, (x_N, y_N)$
 - x_i : value of predictor X (covariate, independent variable, feature,...)
 - y_i : value of response Y (dependent variable, label,...)

- We denote the **regression function** by

$$f(x) = E(Y|X = x)$$

- 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

Simple Linear Regression

- Fitting the model by **least squares**

the values of β_0, β for which $\text{RSS}(\beta_0, \beta)$ attains its minimum.

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

- Solutions are

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

Q: How to get the solutions?

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

sample mean:

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

- $\hat{y}_i = \hat{\beta}_0 + \hat{\beta} x_i$ 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*

Multiple Linear Regression

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

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

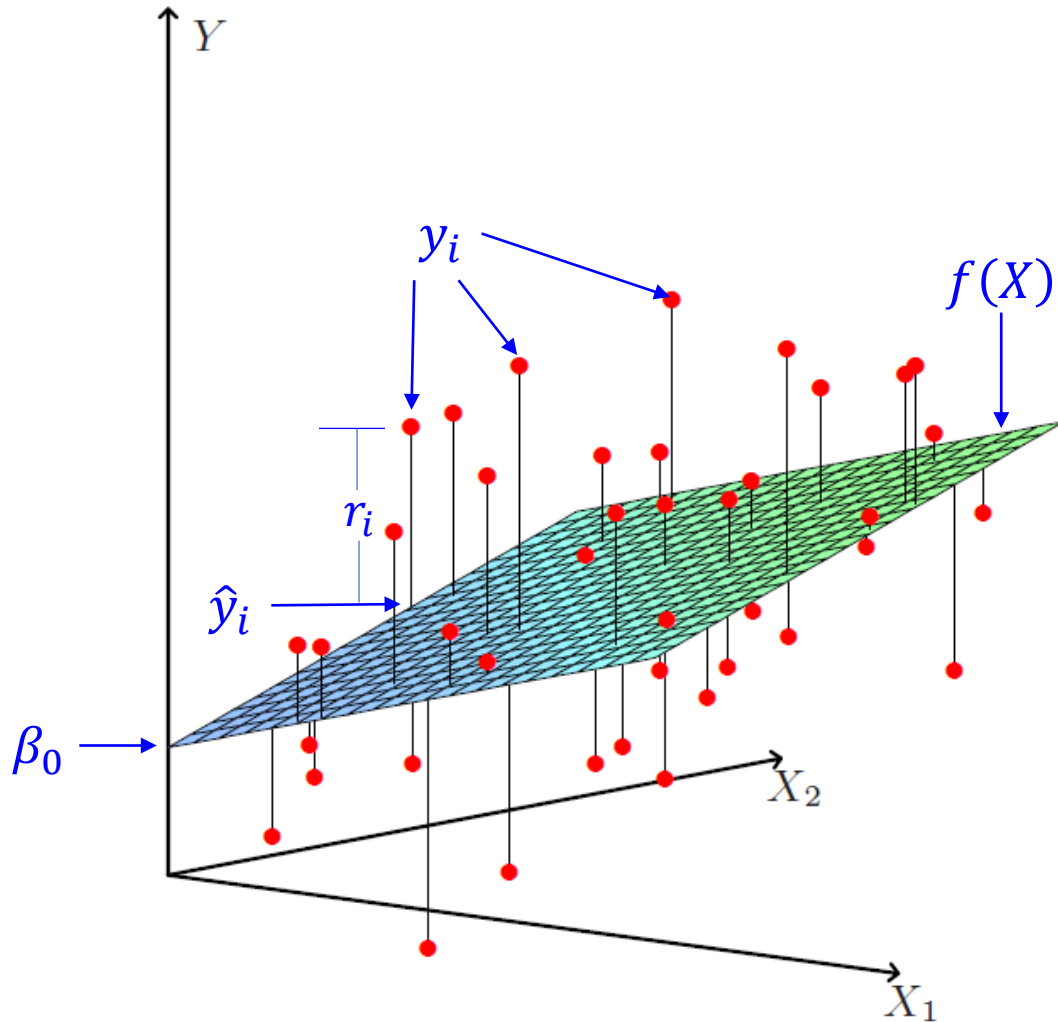
- **Sources** of the variable X_j
 - quantitative inputs
 - transformation
 - basis expansions
 - dummy coding
 - interaction
- **Linear** in the parameters β

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

$$\begin{aligned} \text{RSS}(\beta) &= \sum_{i=1}^N (y_i - f(x_i))^2 \\ &= \sum_{i=1}^N (y_i - \beta_0 - \sum_{j=1}^p x_{ij} \beta_j)^2 \end{aligned}$$

- 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



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

$$\begin{aligned} \text{RSS}(\beta) &= \sum_{i=1}^N (y_i - f(x_i))^2 \\ &= \sum_{i=1}^N (y_i - \beta_0 - \sum_{j=1}^p x_{ij} \beta_j)^2 \end{aligned}$$

- 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 $\text{RSS}(\beta)$
- 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

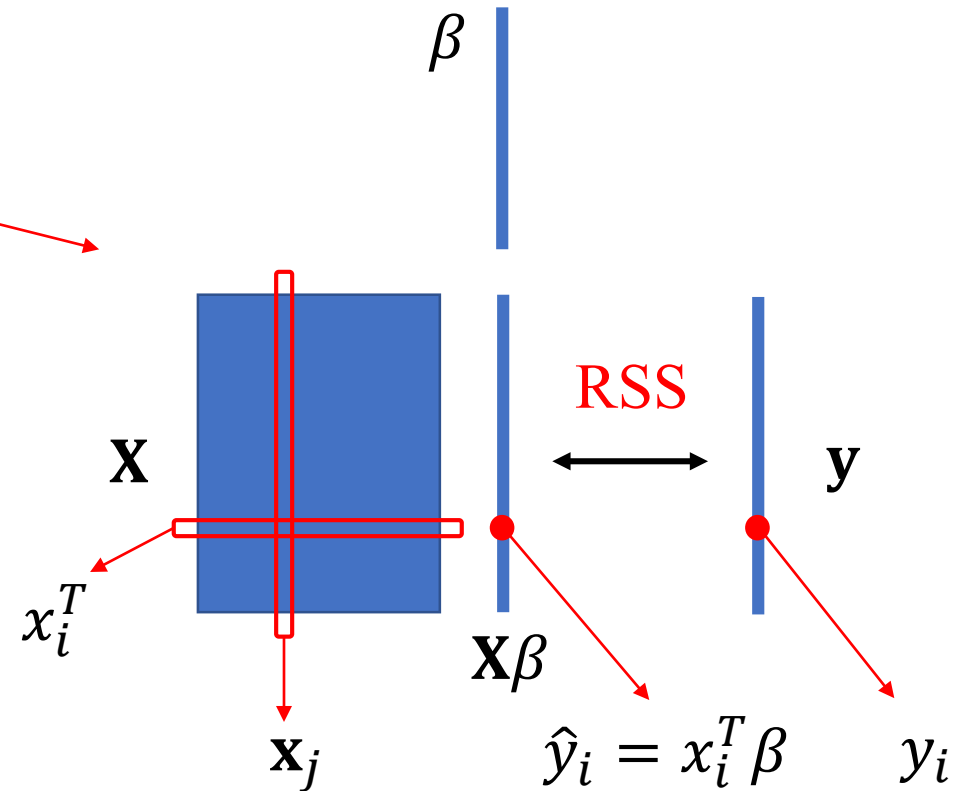
$$\frac{\partial \text{RSS}}{\partial \beta} = -2\mathbf{X}^T (\mathbf{y} - \mathbf{X}\beta)$$

- Set the first derivative to zero

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

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

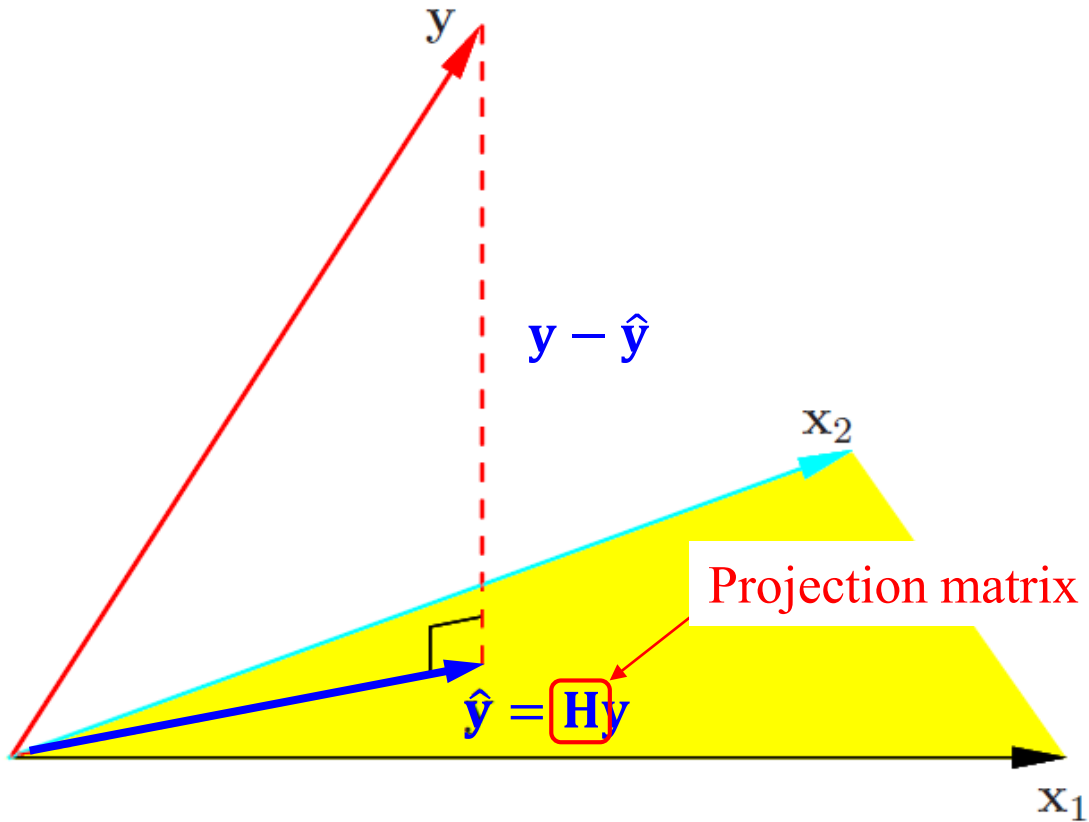
$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$



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
$$\mathbf{X}^T (\mathbf{y} - \mathbf{X}\beta) = 0$$
- If \mathbf{X} has **full column rank**,
$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$
- **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}
- 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



$$\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_p), \text{ where } \mathbf{x}_j = (x_{1j}, \dots, x_{Nj})^T \in \mathbb{R}^N$$

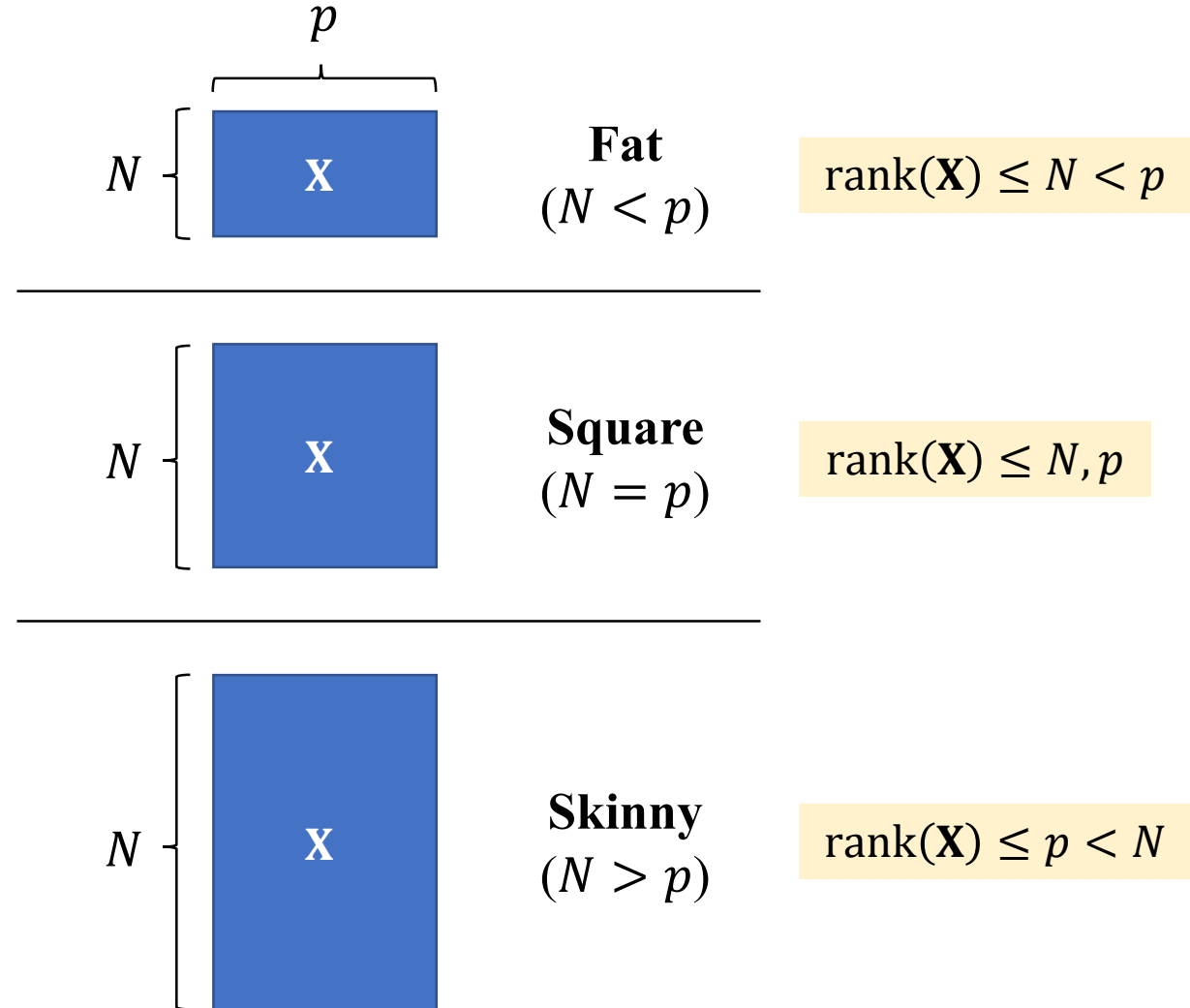
- **Prediction** on a test sample x_0
$$\hat{f}(x_0) = (1: x_0)^T \hat{\beta}$$
- The fitted values at the training inputs
$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} = \mathbf{H}\mathbf{y}$$
- The “**hat**” matrix \mathbf{H}
 - like a hat put on \mathbf{y}
- 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

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

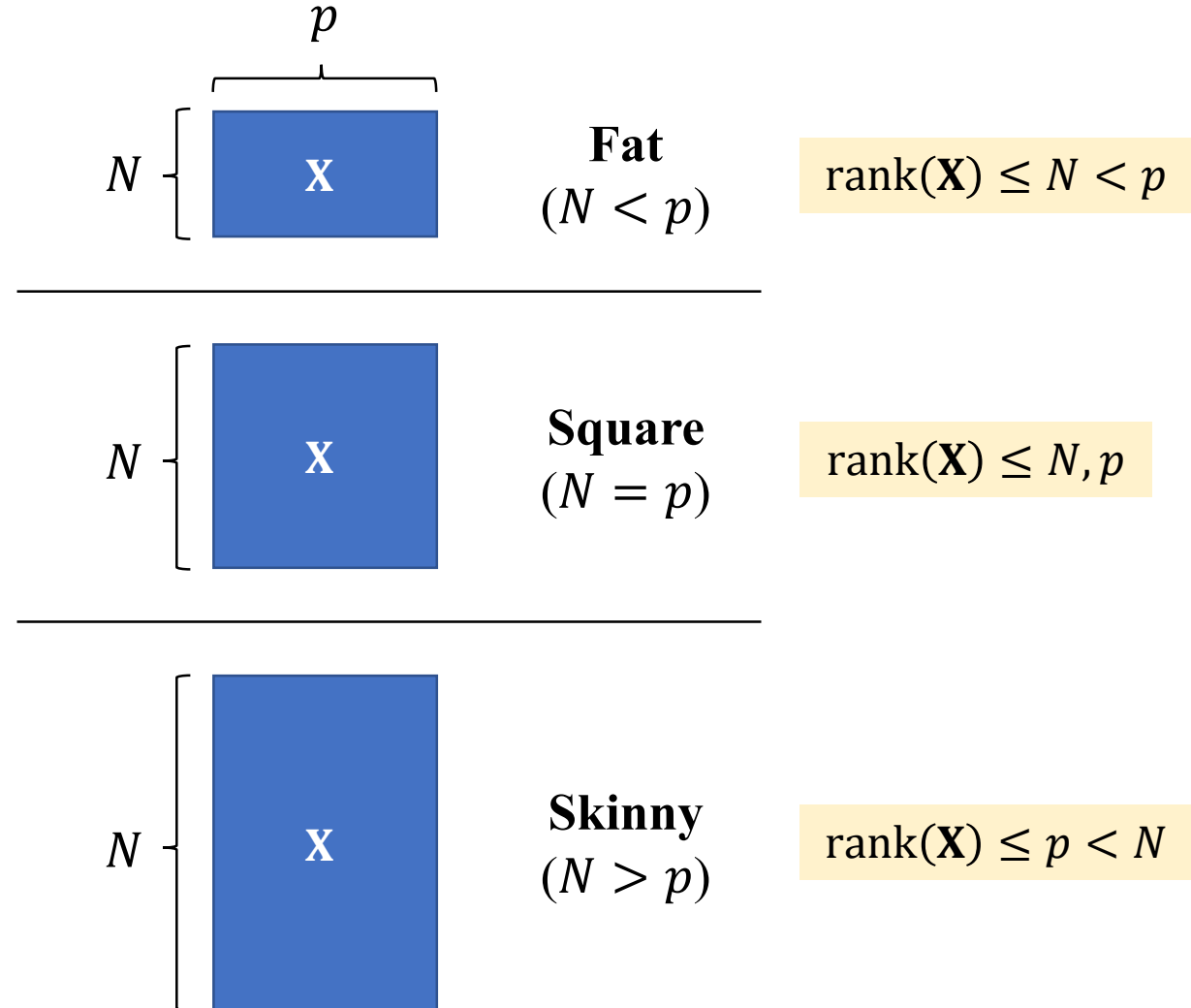
- *Fat* data matrix \mathbf{X}
 - singular
- *Square* data matrix \mathbf{X}
 - probably singular
 - nonsingular if $\text{rank}(\mathbf{X}) = p$
- *Skinny* data matrix \mathbf{X}
 - probably nonsingular
 - singular if $\text{rank}(\mathbf{X}) < p$

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



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**



Multiple Output Regression

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

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

- In matrix notation

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{E}$$

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

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

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

Multiple Output Regression

- Our problem:

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

- A quadratic function with global minimum

- Rewrite $\operatorname{RSS}(\mathbf{B})$ as follows

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

Matrix trace (pointing to Tr)
 $\mathbb{R}^{K \times K}$ (pointing to $\mathbf{B}^T \mathbf{X}^T \mathbf{XB}$)

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

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

- 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

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

$$\begin{aligned} \mathbb{E}[\tilde{\beta}] &= \mathbb{E}[\mathbf{C}\mathbf{y}] \\ &= \mathbb{E}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' + \mathbf{D})(\mathbf{X}\beta + \varepsilon)] \\ &= ((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' + \mathbf{D})\mathbf{X}\beta + ((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' + \mathbf{D})\mathbb{E}[\varepsilon] \\ &= ((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' + \mathbf{D})\mathbf{X}\beta \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\beta + \mathbf{D}\mathbf{X}\beta \\ &= (\mathbf{I}_p + \mathbf{D}\mathbf{X})\beta. \end{aligned}$$

$\mathbb{E}[\varepsilon] = 0$

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

$$\begin{aligned} \text{Var}(\tilde{\beta}) &= \text{Var}(\mathbf{C}\mathbf{y}) \\ &= \mathbf{C} \text{Var}(\mathbf{y}) \mathbf{C}' \\ &= \sigma^2 \mathbf{C} \mathbf{C}' \\ &= \sigma^2 ((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' + \mathbf{D})(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} + \mathbf{D}') \\ &= \sigma^2 ((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}' + \mathbf{D}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} + \mathbf{D}\mathbf{D}') \\ &= \sigma^2 (\mathbf{X}'\mathbf{X})^{-1} + \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{D}\mathbf{X}) + \sigma^2 \mathbf{D}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} + \sigma^2 \mathbf{D}\mathbf{D}' \\ &= \sigma^2 (\mathbf{X}'\mathbf{X})^{-1} + \sigma^2 \mathbf{D}\mathbf{D}' \\ &= \text{Var}(\hat{\beta}) + \sigma^2 \mathbf{D}\mathbf{D}' \end{aligned}$$

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

$\mathbf{D}\mathbf{X} = 0$

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

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

Given an arbitrary test point x_0 , we have

$$\begin{aligned} \text{Var}(\tilde{y}_0) &= \text{Var}(x_0^T \tilde{\beta}) \\ &= x_0^T \text{Var}(\tilde{\beta}) x_0 \\ &= x_0^T \text{Var}(\hat{\beta}) x_0 + \sigma^2 x_0^T \mathbf{D} \mathbf{D}^T x_0 \\ &= \text{Var}(\hat{y}_0) + \sigma^2 x_0^T \mathbf{D} \mathbf{D}^T x_0 \end{aligned}$$

$$\begin{aligned} \text{Var}(\tilde{\beta}) &= \text{Var}(\mathbf{C}\mathbf{y}) \\ &= \mathbf{C} \text{Var}(\mathbf{y}) \mathbf{C}' \\ &= \sigma^2 \mathbf{C} \mathbf{C}' \\ &= \sigma^2 ((\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' + \mathbf{D}) (\mathbf{X} (\mathbf{X}' \mathbf{X})^{-1} + \mathbf{D}') \\ &= \sigma^2 ((\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{X} (\mathbf{X}' \mathbf{X})^{-1} + (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{D}' + \mathbf{D} \mathbf{X} (\mathbf{X}' \mathbf{X})^{-1} + \mathbf{D} \mathbf{D}') \\ &= \sigma^2 (\mathbf{X}' \mathbf{X})^{-1} + \sigma^2 (\mathbf{X}' \mathbf{X})^{-1} (\mathbf{D} \mathbf{X})' + \sigma^2 \mathbf{D} \mathbf{X} (\mathbf{X}' \mathbf{X})^{-1} + \sigma^2 \mathbf{D} \mathbf{D}' \\ &= \sigma^2 (\mathbf{X}' \mathbf{X})^{-1} + \sigma^2 \mathbf{D} \mathbf{D}' \\ &= \text{Var}(\hat{\beta}) + \sigma^2 \mathbf{D} \mathbf{D}' \end{aligned}$$



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
 - $\text{MSE} = \text{Var} + \text{Bias}^2$
- 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

Two **limitations** of least squares

- prediction accuracy
 - **low bias and high variance**
 - sacrifice a little bias to reduce the variance
- interpretation
 - hard to interpret **a large number** of input features
 - find a subset of features exhibiting strong effects

We use **model selection** to overcome the limitations

- variable subset selection, shrinkage, dimension reduction.
- not restricted to linear models

Subset Selection

- Best-subset selection

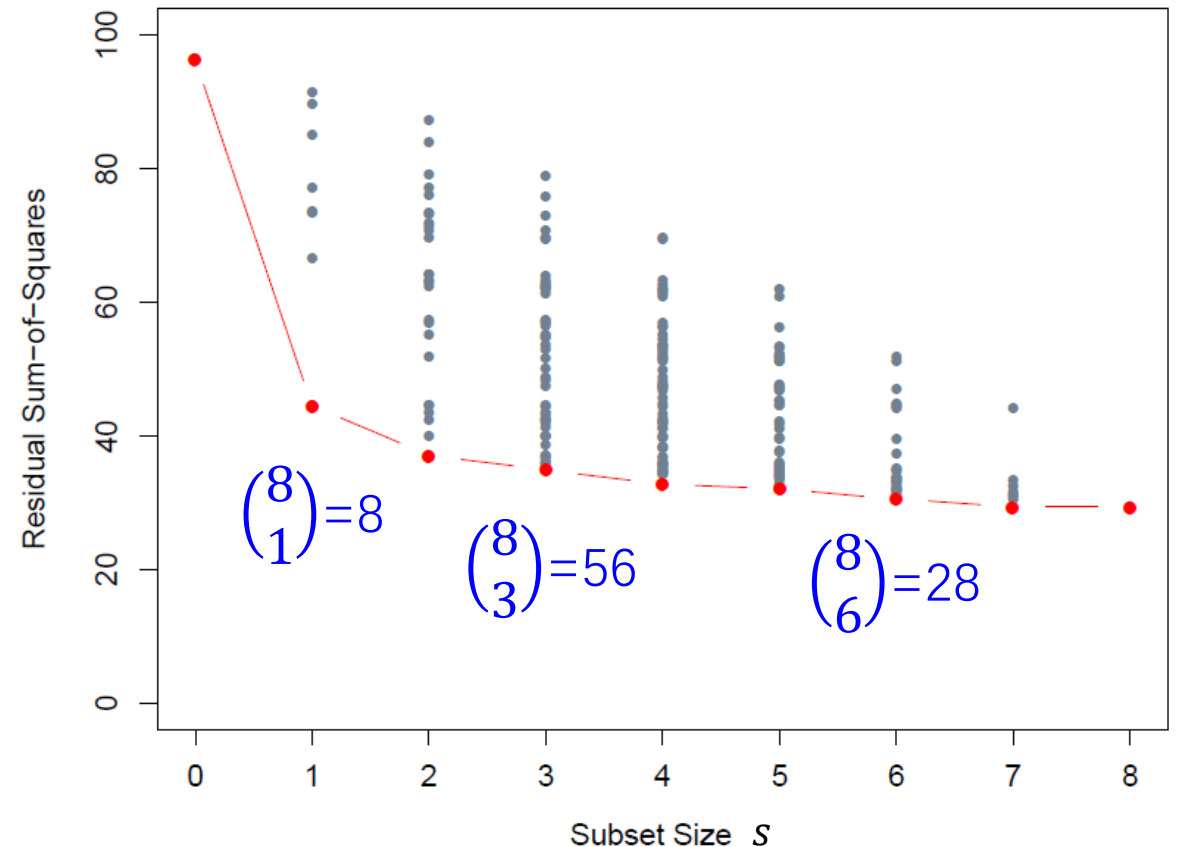
- For each $s \in \{0, 1, \dots, p\}$, find the subset in size of s that gives **lowest** $\text{RSS}(\beta) = \|\mathbf{y} - \mathbf{X}^{(s)}\beta\|_2^2$

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

$p = 4$ $s = 2$	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

- Best-subset selection
 - For each $s \in \{0, 1, \dots, p\}$, 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 ($p = 8$)
 - 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.

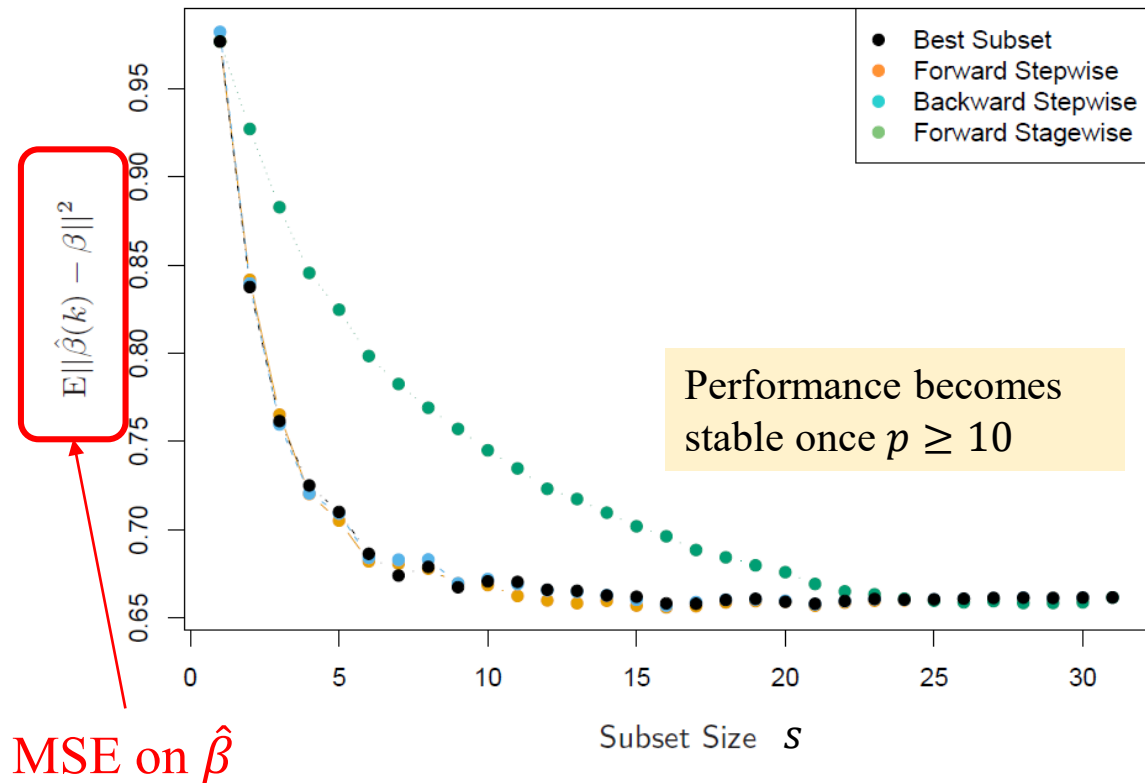
Forward- and Backward-Stepwise Selection

- 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

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
- Backward-stepwise
 - starts with the full model
 - sequentially deletes the worst predictor
- Greedy algorithm
- Only useful when $N > p$
 - linear regression
- Smart stepwise
 - group of variables
 - add or drop whole groups at a time

Forward- and Backward-Stepwise Selection



- Example

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

K -Fold Cross-Validation

- 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
 - divide the training data into K roughly **equal** parts ($K = 5$ or 10)
 - for $k = 1, \dots, K$,
 - fit the model with $K - 1$ parts
 - compute the error E_k on the rest part
 - The K -fold cross validation error

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

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

