# REGRESSION

Dr. Matthieu R Bloch

Wednesday October 06, 2021

# LOGISTICS

**Assignment 4** assigned Tuesday, October 5, 2021

- Includes a (small) programming component
- Due **October 14, 2021** (soft deadline, hard deadline on October 16)

# WHAT'S ON THE AGENDA FOR TODAY?

**Last time**: Least-square regression

**Today**

- Solving linear least-square regression
- Extension to infinite dimension

**Reading:** Romberg, lecture notes 8

# SOLVING THE LEAST-SQUARES PROBLEM

**Proposition.** Any solution $\boldsymbol{\theta}^*$ to the problem $\boxed{\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2}$ optimization must satisfy

$$\boxed{\mathbf{X}^\intercal\mathbf{X}\boldsymbol{\theta}^* = \mathbf{X}^\intercal\mathbf{y}} \text{ — algebraic}$$

This system is called *normal equations*

# SOLVING THE LEAST-SQUARES PROBLEM

**Proposition.** Any solution $\boldsymbol{\theta}^*$ to the problem $\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2$ must satisfy

$$\mathbf{X}^\intercal\mathbf{X}\boldsymbol{\theta}^* = \mathbf{X}^\intercal\mathbf{y}$$

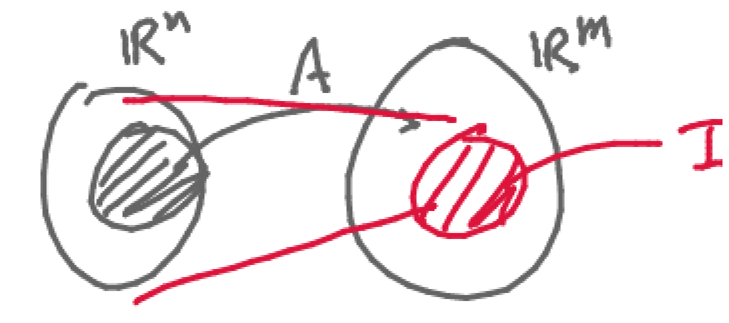
This system is called *normal equations*

**Facts:** for any matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$

- $\ker \mathbf{A}^\intercal\mathbf{A} = \ker \mathbf{A}$

$$\ker(A) \triangleq \mathrm{Null}(A) \triangleq \{x \in \mathbb{R}^n : Ax = 0\} \subset \mathbb{R}^n$$

$$\mathrm{Im}(A) \triangleq \mathrm{Col}(A) \triangleq \{Ax : x \in \mathbb{R}^n\} \subset \mathbb{R}^m$$



# SOLVING THE LEAST-SQUARES PROBLEM

> **Proposition.** Any solution $\boldsymbol{\theta}^*$ to the problem $\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2$ must satisfy
>
> $$\mathbf{X}^\intercal \mathbf{X}\boldsymbol{\theta}^* = \mathbf{X}^\intercal \mathbf{y}$$
>
> This system is called *normal equations*

**Facts:** for any matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$

- $\ker \mathbf{A}^\intercal\mathbf{A} = \ker \mathbf{A}$
- $\operatorname{col}(\underbrace{\mathbf{A}^\intercal\mathbf{A}}_{n\times n}) = \operatorname{row}(\mathbf{A})$ $\subset \mathbb{R}^n$

$$A = [a_{ij}]_{\substack{1\le i\le m\\ 1\le j\le n}} = \begin{pmatrix} | & & | \\ c_1 & \cdots & c_n \\ | & & | \end{pmatrix} = \begin{array}{c} \overset{n}{\longleftrightarrow} \\ \begin{pmatrix} - & r_1 & - \\ & \vdots & \\ - & r_m & - \end{pmatrix} \end{array} \Big\updownarrow m$$

# SOLVING THE LEAST-SQUARES PROBLEM

**Proposition.** Any solution $\boldsymbol{\theta}^*$ to the problem $\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2$ must satisfy

$$\mathbf{X}^\top\mathbf{X}\boldsymbol{\theta}^* = \mathbf{X}^\top\mathbf{y}$$

any element in col$(X^\top X)$

linear combination of columns of $X^\top$

rows of $X$

This system is called *normal equations*

**Facts:** for any matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$

- $\ker \mathbf{A}^\top\mathbf{A} = \ker \mathbf{A}$
- $\mathrm{col}(\mathbf{A}^\top\mathbf{A}) = \mathrm{row}(\mathbf{A})$
- $\mathrm{row}(\mathbf{A})$ and $\ker \mathbf{A}$ are orthogonal complements

We can say a lot more about the normal equations

1. There is always a solution

# SOLVING THE LEAST-SQUARES PROBLEM

**Proposition.** Any solution $\boldsymbol{\theta}^*$ to the problem $\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2$ must satisfy

$X \in \mathbb{R}^{n \times d}$

$$\underbrace{\mathbf{X}^\intercal\mathbf{X}}_{d \times d}\boldsymbol{\theta}^* = \mathbf{X}^\intercal\mathbf{y}$$

This system is called *normal equations*

**Facts:** for any matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$

- $\ker \mathbf{A}^\intercal\mathbf{A} = \ker \mathbf{A}$
- $\mathrm{col}(\mathbf{A}^\intercal\mathbf{A}) = \mathrm{row}(\mathbf{A})$
- $\mathrm{row}(\mathbf{A})$ and $\ker \mathbf{A}$ are orthogonal complements

We can say a lot more about the normal equations

1. There is always a solution
2. If **rank**$(\mathbf{X}) = d$, there is a unique solution: $(\cancel{\mathbf{A}}^\intercal\cancel{\mathbf{A}})^{-1}\cancel{\mathbf{A}}^\intercal\mathbf{y}$ (b/c $X^TX$ is invertible)

# SOLVING THE LEAST-SQUARES PROBLEM

**Proposition.** Any solution $\boldsymbol{\theta}^*$ to the problem $\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2$ must satisfy

$$\mathbf{X}^\intercal \mathbf{X} \boldsymbol{\theta}^* = \mathbf{X}^\intercal \mathbf{y}$$

This system is called *normal equations*

**Facts:** for any matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$

- $\ker \mathbf{A}^\intercal \mathbf{A} = \ker \mathbf{A}$
- $\operatorname{col}(\mathbf{A}^\intercal \mathbf{A}) = \operatorname{row}(\mathbf{A})$
- $\operatorname{row}(\mathbf{A})$ and $\ker \mathbf{A}$ are orthogonal complements

We can say a lot more about the normal equations

1. There is always a solution
2. If $\mathbf{rank}(\mathbf{X}) = d$, there is a unique solution: $(\mathbf{A}^\intercal \mathbf{A})^{-1} \mathbf{A}^\intercal \mathbf{y}$
3. if $\mathbf{rank}(\mathbf{X}) < d$ there are infinitely many non-trivial solution

If $\text{rank}(X) < d$ $\quad \text{Ker}(X) \neq \phi$

$\exists\, \theta_0$ st $\theta_0 \neq 0$ and $X\theta_0 = 0$

For any solution $\theta^*$ of the normal equa

$\theta^* + \theta_0$ is also a solution

$X^T X(\theta^* + \theta_0) = X^T X \theta^* = X^T y$

<u>Remark</u>: the pace of solutions is $\theta^* + \text{Ke}$

Remark: Assume $\tilde{\theta}$ a solution of $X^TX\tilde{\theta} = X^Ty$ (in addition $\theta^*$)

then $\tilde{\theta} \stackrel{a}{=} \boxed{\theta^* + \underbrace{\tilde{\theta} - \theta^*}}$

$X^TX(\hat{\theta} - \theta^*) = X^Ty - X^Ty = 0$ so that $\hat{\theta} - \theta^* \in \mathrm{Ker}(X^TX) = \mathrm{Ker}(X)$

Hence $\mathcal{S} \subset \theta^* + \mathrm{Ker}(X)$

space of solutions

Remark: if $\theta_0 \in \mathrm{Ker}(X)$ then $\forall \alpha \in \mathbb{R}$ $\alpha\theta_0 \in \mathrm{Ker}(X)$

# SOLVING THE LEAST-SQUARES PROBLEM

**Proposition.** Any solution $\boldsymbol{\theta}^*$ to the problem $\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2$ must satisfy

$(X \in \mathbb{R}^{n \times d})$

$$\mathbf{X}^\intercal\mathbf{X}\boldsymbol{\theta}^* = \mathbf{X}^\intercal\mathbf{y}$$

This system is called *normal equations*

**Facts:** for any matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$

- $\ker \mathbf{A}^\intercal\mathbf{A} = \ker \mathbf{A}$
- $\mathrm{col}(\mathbf{A}^\intercal\mathbf{A}) = \mathrm{row}(\mathbf{A})$
- $\mathrm{row}(\mathbf{A})$ and $\ker \mathbf{A}$ are orthogonal complements

We can say a lot more about the normal equations

1. There is always a solution
2. If **rank**$(\mathbf{X}) = d$, there is a unique solution: $(\mathbf{A}^\intercal\mathbf{A})^{-1}\mathbf{A}^\intercal\mathbf{y}$
3. if **rank**$(\mathbf{X}) < d$ there are infinitely many non-trivial solution
4. if **rank**$(\mathbf{X}) = n$, there exists a solution $\boldsymbol{\theta}^*$ for which $\mathbf{y} = \mathbf{X}\boldsymbol{\theta}^*$

$X = \underbrace{\begin{bmatrix} & & \end{bmatrix}}_{d} \updownarrow n$

Note: rank$(X) = n$

then $XX^T$ is invertible *

so that $XX^TX\theta^* = XX^Ty$

and $X\theta^* = y$

$\|y - X\theta^*\|_2^2 = 0$

# SOLVING THE LEAST-SQUARES PROBLEM

> **Proposition.** Any solution $\boldsymbol{\theta}^*$ to the problem $\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2$ must satisfy
>
> $$\mathbf{X}^\intercal\mathbf{X}\boldsymbol{\theta}^* = \mathbf{X}^\intercal\mathbf{y}$$
>
> This system is called *normal equations*

**Facts:** for any matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$

- $\ker \mathbf{A}^\intercal\mathbf{A} = \ker \mathbf{A}$
- $\operatorname{col}(\mathbf{A}^\intercal\mathbf{A}) = \operatorname{row}(\mathbf{A})$
- $\operatorname{row}(\mathbf{A})$ and $\ker \mathbf{A}$ are orthogonal complements

We can say a lot more about the normal equations

1. There is always a solution
2. If **rank**$(\mathbf{X}) = d$, there is a unique solution: $(\mathbf{A}^\intercal\mathbf{A})^{-1}\mathbf{A}^\intercal\mathbf{y}$
3. if **rank**$(\mathbf{X}) < d$ there are infinitely many non-trivial solution
4. if **rank**$(\mathbf{X}) = n$, there exists a solution $\boldsymbol{\theta}^*$ for which $\mathbf{y} = \mathbf{X}\boldsymbol{\theta}^*$

In machine learning, there are often infinitely many solutions

# MINIMUM NORM 2 SOLUTIONS

One reasonable to choose a solution among infinitely many is the *minimum energy* principle

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \|\boldsymbol{\theta}\|_2^2 \text{ such that } \mathbf{X}^\intercal \mathbf{X} \boldsymbol{\theta} = \mathbf{X}^\intercal \mathbf{y}$$

- We will see the solution is always unique using the SVD

For now, assume that **rank**$(\mathbf{X}) = d$, so that the problem becomes

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \|\boldsymbol{\theta}\|_2^2 \text{ such that } \mathbf{X} \boldsymbol{\theta} = \mathbf{y}$$

# MINIMUM NORM 2 SOLUTIONS

One reasonable to choose a solution among infinitely many is the *minimum energy* principle

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \|\boldsymbol{\theta}\|_2^2 \text{ such that } \boxed{\mathbf{X}^\intercal\mathbf{X}\boldsymbol{\theta} = \mathbf{X}^\intercal\mathbf{y}}$$

- We will see the solution is always unique using the SVD

For now, assume that **rank**$(\mathbf{X}) = \cancel{d}$, so that the problem becomes

$n$

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \|\boldsymbol{\theta}\|_2^2 \text{ such that } \mathbf{X}\boldsymbol{\theta} = \mathbf{y}$$

**Proposition.** The solution is $\boldsymbol{\theta}^* = \cancel{\mathbf{A}}^\intercal(\cancel{\mathbf{A}}\cancel{\mathbf{A}}^\intercal)^{-1}\mathbf{y} \in \mathbb{R}^d$

$X^T \underbrace{(XX^T)^{-1} y}_{\alpha \in \mathbb{R}^n}$ is in $\mathrm{col}(X^T) = \mathrm{row}(X) = \mathrm{span}(\{x_i\}_{i=1}^n)$

Proof: A solution is in $\mathbb{R}^d = \mathrm{Ker}(X) \oplus \mathrm{row}(X)$ so that it can be written $\underbrace{\theta_1}_{\in \mathrm{row}(X)} + \underbrace{\theta_2}_{\in \mathrm{Ker}(X)}$

$\left( X = \begin{pmatrix} - & x_1^T & - \\ & \vdots & \\ - & x_n^T & - \end{pmatrix} \in \mathbb{R}^{n \times d} \right)$

Hence we are looking for $\min_{\substack{\theta_1 \in \mathrm{row}(X) \\ \theta_2 \in \mathrm{Ker}(X)}} \overbrace{\|\theta_1 + \theta_2\|_2^2}^{\|\theta_1\|_2^2 + \|\theta_2\|_2^2}$ s.t. $X(\theta_1 + \theta_2) = y = X\theta_1$

Since $\mathrm{Ker}(X)^\perp = \mathrm{row}(X)$ then $\|\theta_1 + \theta_2\|_2^2 = \|\theta_1\|^2 + \|\theta_2\|^2$

Hence any solution of the problem must be such that $\theta_2 = 0$ and must be in $\mathrm{row}(X)$

Therefore we can parametrize any solution as $\theta = X^T \alpha$

Our equivalent problem is then to find $\min_{\alpha \in \mathbb{R}^{n \times 1}} \|X^T \alpha\|_2^2$ s.t. $y = \underbrace{XX^T}_{\in \mathbb{R}^{n \times n}} \alpha$ is full rank b/c rank(X

The only solution is actually w/ $\alpha = (XX^T)^{-1} y$ so that $\boxed{\theta = X^T (XX^T)^{-1} y}$

# REGULARIZATION

Recall the problem

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \|\boldsymbol{\theta}\|_2^2 \text{ such that } \mathbf{X}^\top\mathbf{X}\boldsymbol{\theta} = \mathbf{X}^\top\mathbf{y}$$

- There are infinitely many solution if $\ker \mathbf{X}$ is non trivial
- The space of solution is unbounded! ( $\theta^k + \theta_0$ : $\|\theta^k + \theta_0\|_2$ is unbounded )

# REGULARIZATION

Recall the problem

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \|\boldsymbol{\theta}\|_2^2 \text{ such that } \mathbf{X}^\top\mathbf{X}\boldsymbol{\theta} = \mathbf{X}^\top\mathbf{y} \qquad (*)$$

$y_i = x_i^\top \theta$

- There are infinitely many solution if $\ker \mathbf{X}$ is non trivial
- The space of solution is unbounded! $(\dagger)$
- Even if $\ker \mathbf{X} = \{0\}$, the system can be poorly conditioned

**Regularization** with $\lambda > 0$ consists in solving

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \underbrace{\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2}_{\text{solution}} + \overbrace{\lambda\underbrace{\|\boldsymbol{\theta}\|_2^2}_{\text{good solution}}}^{\text{regularization}}$$

hyperparameter (chosen by crossvalidation)

# REGULARIZATION

Recall the problem

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \|\boldsymbol{\theta}\|_2^2 \text{ such that } \mathbf{X}^\intercal\mathbf{X}\boldsymbol{\theta} = \mathbf{X}^\intercal\mathbf{y}$$

- There are infinitely many solution if $\ker \mathbf{X}$ is non trivial
- The space of solution is unbounded!
- Even if $\ker \mathbf{X} = \{0\}$, the system can be poorly conditioned

**Regularization** with $\lambda > 0$ consists in solving

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_2^2$$

- This problem *always* has a unique solution

**Proposition.** The solution is $\boldsymbol{\theta}^* = (\underbrace{\mathbf{X}^\intercal\mathbf{X}}_{\in\mathbb{R}^{d\times d}} + \lambda\mathbf{I})^{-1}\mathbf{X}^\intercal\mathbf{y} = \mathbf{X}^\intercal(\underbrace{\mathbf{X}\mathbf{X}^\intercal}_{\in\mathbb{R}^{n\times n}} + \lambda\mathbf{I})^{-1}\mathbf{y}$

Proof. $\|y - X\theta\|_2^2 + \lambda \|\theta\|_2^2 = (y - X\theta)^T (y - X\theta) + \lambda \theta^T \theta$

$$= y^T y - 2 y^T X\theta + \underbrace{\theta^T X^T X \theta + \lambda \theta^T \theta}_{\theta^T (X^T X + \lambda I)\theta}$$

Taking the gradient and setting it to zero we obtain $\underbrace{(X^T X + \lambda I)}_{\text{is always invertible!}} \theta = X^T y$

Facts: ① $X^T X + \lambda I$ is symmetric

② $X^T X + \lambda I$ is positive definite : $w^T (X^T X + \lambda I) w = w^T X^T X w + \lambda w^T w = \|Xw\|_2^2 + \lambda \|w\|_2^2 \geq 0$

w/ equality iff $w = 0$

③ $X^T X + \lambda I$ is invertible b/c $\lambda > 0$

Hence $\theta = (X^T X + \lambda I)^{-1} X^T y$

We want to show that $\theta = X^T(XX^T + \lambda \underbrace{I}_{n\times n})^{-1} y$

Since $\theta$ is unique, all we have to check is that $\theta$ is a solution of $(X^TX + \lambda \overbrace{I}^{d\times d})\theta = X^T y$

Note that

$$(\underbrace{X^TX + \lambda I}_{d\times d})X^T(\underbrace{XX^T + \lambda I}_{n\times n})^{-1} y = (X^TXX^T + \lambda \overbrace{X^T}^{d\times n})(XX^T + \lambda I)^{-1} y$$

$$= X^T\underbrace{(\underbrace{XX^T + \lambda I}_{n\times n})(\underbrace{XX^T + \lambda I}_{n\times n})^{-1}}_{I} y$$

$$= X^T y$$

# REGULARIZATION

Recall the problem

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \|\boldsymbol{\theta}\|_2^2 \text{ such that } \mathbf{X}^\intercal\mathbf{X}\boldsymbol{\theta} = \mathbf{X}^\intercal\mathbf{y}$$

- There are infinitely many solution if $\ker \mathbf{X}$ is non trivial
- The space of solution is unbounded!
- Even if $\ker \mathbf{X} = \{0\}$, the system can be poorly conditioned

**Regularization** with $\lambda > 0$ consists in solving

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_2^2$$

- This problem *always* has a unique solution

**Proposition.** The solution is $\boldsymbol{\theta}^* = (\mathbf{X}^\intercal\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\intercal\mathbf{y} = \mathbf{X}^\intercal(\mathbf{X}\mathbf{X}^\intercal + \lambda\mathbf{I})^{-1}\mathbf{y}$

Note that $\boldsymbol{\theta}^*$ is the row space of $\mathbf{X}$

$$\boldsymbol{\theta}^* = \mathbf{X}\boldsymbol{\alpha} \text{ with } \boldsymbol{\alpha} = (\mathbf{X}\mathbf{X}^\intercal + \lambda\mathbf{I})^{-1}\mathbf{y}$$