

# Modern Statistical Modeling

Lectured by [Wei Lin](#)

L<sup>A</sup>T<sub>E</sub>Xed by [Chengxin Gong](#)

March 24, 2023

## Contents

|          |  |          |
|----------|--|----------|
| <b>1</b> | <b><a href="#">Review of Linear Algebra</a></b>        | <b>2</b> |
| <b>2</b> | <b><a href="#">Review of Probability Theory</a></b>    | <b>2</b> |
| <b>3</b> | <b><a href="#">Prediction and Nearest Neighbor</a></b> | <b>3</b> |
| <b>4</b> | <b><a href="#">Linear Regression</a></b>               | <b>4</b> |
| <b>5</b> | <b><a href="#">Generalized Linear Models</a></b>       | <b>6</b> |

# 1 Review of Linear Algebra

- Rank of  $A \in \mathbb{R}^{m \times n}$ : max # of linearly independent row/columns. Facts: (i)  $0 \leq \text{rank}(A) \leq \min(m, n)$ ; (ii)  $\text{rank}(A) = \text{rank}(A^T) = \text{rank}(AA^T) = \text{rank}(A^T A)$ ; (iii)  $\text{rank}(BAC) = \text{rank}(A)$  for nonsingular compatible  $B, C$ .
- Range(column space):  $\mathcal{C}(A) = \{Ax : x \in \mathbb{R}^n\} \subset \mathbb{R}^m$ . Null space:  $\mathcal{N}(A) = \{x \in \mathbb{R}^n : Ax = 0\}$ . Facts: (i)  $\text{rank}(A) = \dim \mathcal{C}(A)$ ; (ii)  $\dim \mathcal{C}(A) + \dim \mathcal{N}(A) = n$ ; (iii)  $\mathcal{N}(A) = \mathcal{C}(A^T)^\perp$ ; (iv)  $\mathcal{C}(AA^T) = \mathcal{C}(A)$ .
- Trace of  $A \in \mathbb{R}^{m \times n}$ :  $\text{tr}(A) = \sum_{i=1}^n a_{ii}$ . Facts: (i) linearity:  $\text{tr}(A+B) = \text{tr}(A) + \text{tr}(B)$ ,  $\text{tr}(cA) = c\text{tr}(A)$ ; (ii) cyclic property:  $\text{tr}(AB) = \text{tr}(BA)$ ,  $\text{tr}(ABC) = \text{tr}(BCA) = \text{tr}(CAB)$ ; (iii)  $\text{tr}(A) = \sum_{i=1}^n \lambda_i a_{ij} b_{ij}$ .
- Trace product:  $\langle A, B \rangle = \text{tr}(A^T B) = \text{tr}(AB^T) = \sum_i \sum_j a_{ij} b_{ij}$ . It induces Frobenius norm:  $\|A\|_F = \sqrt{\langle A, A \rangle} = (\sum_{i,j} a_{ij}^2)^{1/2}$ .
- Determinant:  $\det(A)$  or  $|A|$ . Facts: (i)  $\det(cA) = c^n \det(A)$ ; (ii)  $\det(AB) = \det A \det B$ ; (iii)  $\det(A^{-1}) = \det(A)^{-1}$ ; (iv)  $\det(A) = \prod_{i=1}^n \lambda_i$ .
- Three decomposition. (1) For symmetric  $A$ , spectrum(eigen) decomposition:  $A = V \Lambda V^T = \sum_{i=1}^r \lambda_i v_i v_i^T$  where  $V$  is orthogonal ( $V^T V = V V^T = I$ ) and  $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_n)$ . (2) SVD for  $A \in \mathbb{R}^{n \times p}$  of rank  $r$ :  $A = U \Sigma V^T = \sum_{i=1}^r \sigma_i u_i v_i^T$  where  $\Sigma = \text{diag}(\sigma_1, \dots, \sigma_r, 0, \dots, 0)$ ,  $\sigma_1 \geq \dots \geq \sigma_r \geq 0$  and  $\{u_i\}, \{v_i\}$  orthonormal.  $\arg \min_{Y \in \mathbb{R}^{n \times p}, \text{rank}(Y) \leq r} \|X - Y\|_F = \sum_{i=1}^r \sigma_i u_i v_i^T$  (low rank- $r$  approximation). (3) QR decomposition:  $A = QR$  where  $Q$  is orthonormal and  $R$  is upper-triangular. It corresponds to Gram-Schmidt orthogonalization process.
- Idempotent:  $P^T = P$ . Facts: (i) If  $P$  is symmetric, then  $P$  is idempotent of rank  $r$  iff it has  $r$  eigenvalues 1 and  $n - r$  0; (ii) If  $P$  is a projection matrix, then  $\text{tr}(P) = \text{rank}(P)$ .
- Generalized inverses: For  $A \in \mathbb{R}^{m \times n}$ ,  $A^- \in \mathbb{R}^{n \times m}$  is called a generalized inverse of  $A$  if  $AA^-A = A$ . Moore-Penrose inverse  $A^+$  if (i)  $AA^+A = A$ ; (ii)  $A^+AA^+ = A^+$ ; (iii)  $(A^+A)^T = A^+A$ ; (iv)  $(AA^+)^T = AA^+$ . Such  $A^+$  is unique, and  $A^+ = V \Sigma^+ U^T = \sum_{i=1}^r \sigma_i^{-1} v_i u_i^T$ .
- **Theorem 1.1**  $P_X = X(X^T X)^- X^T$  is the orthogonal projection onto  $\mathcal{C}(X)$ . [ $P_X$  does not depend on the choice of  $(X^T X)^-$ ]

**Proof**  $\forall v \in \mathbb{R}^n$ , write  $v = x + w$  where  $x \in \mathcal{C}(X), w \in \mathcal{C}(X)^T$ . By definition,  $P_X v = P_X x + P_X w = P_X x + X(X^T X)^- X^T w = P_X x$ . We need to show  $u^T X(X^T X)^- X^T X = u^T X, \forall u \in \mathbb{R}^n$ .

**Lemma 1.1**  $\mathcal{C}(X^T) = \mathcal{C}(X^T X)$ .

**Proof** Use  $\mathcal{C}(X^T X) \subset \mathcal{C}(X^T)$  and  $\text{rank}(X^T X) = \text{rank}(X)$ . □

By the lemma,  $u^T X(X^T X)^- X^T X = z^T X^T X(X^T X)^- X^T X = z^T X^T X = u^T X$ . □

# 2 Review of Probability Theory

- Distribution related to multivariate normal:  $X \sim \mathcal{N}_p(\mu, \Sigma)$ . Moment generating function:  $M_X(t) = \mathbb{E}e^{t^T X} = \exp(t^T \mu + \frac{1}{2} t^T \Sigma t)$ . Characteristic function:  $\phi_X(t) = \mathbb{E}e^{it^T X} = \exp(it^T \mu - \frac{1}{2} t^T \Sigma t)$ . Facts: (i)  $A_{g \times p} X + b_{g \times 1} \sim \mathcal{N}_g(A\mu + b, A\Sigma A^T)$ ; (ii)  $X \sim \mathcal{N}_p(\mu, \Sigma) \Leftrightarrow a^T X \sim \mathcal{N}(a^T \mu, a^T \Sigma a), \forall a \in \mathbb{R}^p$ ; (iii)  $Y_1 = A_1 X + b_1 \perp\!\!\!\perp Y_2 = A_2 X + b_2 \Leftrightarrow \text{Cov}(Y_1, Y_2) = A_1 \Sigma A_2^T = 0$ .
- Noncentral  $\chi^2$ :  $X \sim \mathcal{N}_p(\mu, I_p)$ . Then  $X^T X \sim \chi_p^2(\lambda)$  with noncentral parameter  $\lambda = \mu^T \mu$ . Pdf of  $\chi_p^2(\lambda)$ :  $f(x; p, \lambda) = \sum_{k=0}^{\infty} \frac{e^{-\lambda/2} (\lambda/2)^k}{k!} f(x; p + 2k, 0)$  where  $f_q(x) = f(x; q, 0) = \frac{x^{q/2} e^{-x/2}}{2^{q/2} \Gamma(q/2)} I(x > 0)$ , a  $\text{Poisson}(\frac{\lambda}{2})$ -weighted mixture of  $\chi_{p+2k}^2$ . M.g.f.:  $M_X(t; p, \lambda) = \frac{1}{(1-2it)^{p/2}} \exp(\frac{\lambda t}{1-2it})$ . Ch.f.:  $\Phi_X(t; p, \lambda) = \frac{1}{(1-2it)^{p/2}} \exp(\frac{i\lambda t}{1-2it})$ . Facts: (i)

If  $X \sim \mathcal{N}(\mu, \Sigma)$  then  $(X - \mu)^T \Sigma^{-1} (X - \mu) \sim \chi_p^2$  and  $X^T \Sigma^{-1} X \sim \chi_p^2(\mu^T \Sigma^{-1} \mu)$ ; (ii) Additivity: If  $X \sim \chi_{p_i}^2(\lambda_i)$  independent for  $i = 1, \dots, k$ , then  $\sum_{i=1}^n X_i \sim \chi_{\sum_i p_i}^2(\sum_i \lambda_i)$ ; (iii) Rank deficient: If  $X \sim \mathcal{N}_p(\mu, I_p)$ ,  $A \in \mathbb{R}^{p \times p}$  symmetric, then  $X^T A X \sim \chi_p^2(\lambda)$  with  $\lambda = \mu^T A \mu \Leftrightarrow A$  is idempotent of rank  $r$ ; (iv) If  $X \sim \mathcal{N}_p(\mu, \Sigma)$ ,  $A \in \mathbb{R}^{p \times p}$  symmetric,  $B \in \mathbb{R}^{q \times p}$ , then  $X^T A X \perp\!\!\!\perp B X \Leftrightarrow B \Sigma A = 0_{q \times p}$ ; (v)  $X^T A X \perp\!\!\!\perp X^T B X \Leftrightarrow A \Sigma B = 0_{p \times p}$ .

- **Theorem 2.1** (Cochran)  $X \sim \mathcal{N}_p(\mu, I_p)$ ,  $X^T X = X^T A_1 X + \dots + X^T A_k X \equiv Q_1 + \dots + Q_k$ ,  $A_i \in \mathbb{R}^{p \times p}$  symmetric of rank  $r_i$ . Then  $Q_i \sim \chi_{r_i}^2(\lambda_i)$  independent for  $i = 1, \dots, k \Leftrightarrow p = r_1 + \dots + r_k$ . In this case,  $\lambda_i = \mu^T A_i \mu$  and  $\lambda_1 + \dots + \lambda_k = \mu^T \mu$ .

**Proof** “ $\Leftarrow$ ”: Note that  $\forall i, \exists c_{ij} \in \mathbb{R}^p, j = 1, \dots, r_i$  s.t.  $Q_i = X^T A_i X = \pm (c_{i1}^T X)^2 \pm \dots \pm (c_{ir_i}^T X)^2$ . Let  $C_i = (c_{i1}, \dots, c_{ir_i})$  and  $C_{p \times r} = (C_1, \dots, C_k)^T$ , then  $X^T X = X^T C \Delta C X$ , where  $\Delta$  is  $p \times p$  diagonal with diagonal entries  $\pm 1 \Rightarrow C^T \Delta C = I_p$ . Thus  $C$  is of full rank and hence  $\Delta = (C^T)^{-1} C^{-1} = (C^{-1})^T C^{-1} = (C^{-1})^T C^{-1}$  is positive definite  $\Rightarrow \Delta = I_p$  and  $C^T C = I_p$ .

“ $\Rightarrow$ ”:  $X^T A_i \sim \chi_{r_i}^2(\lambda_i)$  independent  $\Rightarrow X^T X = \sum_i X^T A_i X \sim \chi_{\sum_i r_i}^2(\sum_i \lambda_i) \Rightarrow \sum_i r_i = p$ .  $\square$

- Noncentral  $F$ : If  $Q_1 \sim \chi_p^2(\lambda)$  and  $Q_2 \sim \chi_q^2$  are independent, then  $\frac{Q_1/p}{Q_2/q} \sim F_{p,q}(\lambda)$ .
- Noncentral  $t$ : If  $U_1 \sim \mathcal{N}(\lambda, 1)$  and  $U_2 \sim \chi_q^2$  are independent, then  $T = \frac{U_1}{\sqrt{U_2/q}} \sim t_q(\lambda)$ .

### 3 Prediction and Nearest Neighbor

- Goal: (1) predict  $y$  from  $x$  (“black box”); (2) which variable(s) in  $x$  contributes to the prediction of  $y$  (“ $x^T \beta$ ”), estimation, testing, variable selection.
- Why are prediction and estimation different: (1) model parameters; (2) identifiability ( $f_{\theta_1} \neq f_{\theta_2} \Rightarrow \theta_1 \neq \theta_2$ ).
- Find prediction function  $f: \mathcal{X} \rightarrow \mathcal{Y}$  that minimizes  $\mathbb{E}_{X,Y} \mathcal{L}(f(X), Y) = \mathbb{E}\{\mathbb{E}(\mathcal{L}(f(X), Y) | X)\}$  where loss function  $\mathcal{L}: \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$ .
- Optimal predictor conditioned on  $x$ :  $f^*(x) = \arg \min_{f(x) \in \mathcal{Y}} \mathbb{E}\{\mathcal{L}(f(X), Y) | X = x\}$ .
- Regression:  $y$  numerical, squared error ( $L_2$ -loss)  $\mathcal{L}(\hat{y}, y) = (\hat{y} - y)^2$ ,  $\mathbb{E}\{(Y - f(X))^2 | X\} = \{\mathbb{E}(Y | X) - f(X)\}^2 + \mathbb{E}\{(Y - \mathbb{E}(Y | X))^2 | X\} = \text{bias}^2 + \text{variance}$ . Optimal  $f^*(X) = \mathbb{E}(Y | X)$ .
- To model  $f^*$ ,  $\begin{cases} \text{parametric: linear, } f^*(x) = x^T \beta, \beta \in \mathbb{R}^2 \\ \text{nonparametric: infinite dimension, } f^*(x) = m(x), m \text{ satisfying certain smoothness} \end{cases}$ .
- Classification: 0-1 loss  $\mathcal{L}(\hat{y}, y) = I(\hat{y} \neq y)$ ,  $\mathbb{E}\{\mathcal{L}(h(X), Y) | X = x\} = \sum_{j \neq h(x)} P(Y = j | X = x) = 1 - P(Y = h(X) | X = x)$ . Optimal classification (Bayes classifier):  $h^*(x) = \arg \max_{h(x) \in \mathcal{Y}} P(Y = h(X) | X = x)$ .
- A fully nonparametric approach:  $k$  nearest neighbor ( $k$ -NN). Given training data  $\{(x_i, y_i)\}_{i=1}^m$ , use data “around”  $x$  to estimate  $m(x) = \mathbb{E}(Y | X = x)$ . Rationale: “Things that look alike must be alike”. Classification:  $h_{k\text{-NN}}(x) = \text{majority label among } \{y_i, i \in N_k(x)\}$ . Regression:  $m_{k\text{-NN}}(x) = \frac{1}{k} \sum_{i \in N_k(x)} y_i$ .  $k$  controls size of neighbor set.  $k \uparrow$ : effective sample size  $\uparrow$ , variance  $\downarrow$ , heterogeneity  $\uparrow$ , bias  $\uparrow$ .
- Theory for 1-NN: Consider binary classification:  $\mathcal{Y} = \{0, 1\}$ ,  $\mathcal{L}(h(x), y) = I(h(x) \neq y)$ . Assume  $\mathcal{X} \subset [0, 1]^d$ ,  $\rho$  Euclidean distance,  $S = \{(x_i, y_i)\}_{i=1}^n$ .  $\forall x \in \mathcal{X}$ , let  $\pi_1(x), \dots, \pi_n(x)$  be an ordering of  $\{1, \dots, n\}$  with increasing distance to  $x$ .  $\eta(x) = \mathbb{E}(Y = 1 | X = x)$ . Bayes classifier:  $h^*(x) = I(\eta(x) > \frac{1}{2})$ . Assumption on  $\eta$ :  $\eta$  is  $c$ -Lipschitz for some  $c > 0$ . Goal: Derive an upper bound on  $\mathbb{E}_{S \sim \mathcal{D}^n} \mathcal{L}(\hat{h}_S) = \mathbb{E}_{S \sim \mathcal{D}^n} \mathbb{E}_{(x,y) \sim \mathcal{D}} I(\hat{h}_S(x) \neq y)$ .
- **Lemma 3.1** The 1-NN rule  $\hat{h}_S$  satisfies  $\mathbb{E}_{S \sim \mathcal{D}^n} \mathcal{L}(\hat{h}_S) \leq 2\mathcal{L}(h^*) + c \mathbb{E}_{S \sim \mathcal{D}^n, x \sim \mathcal{D}} \|x - x_{\pi_1}(x)\|$ .

**Proof**  $\mathbb{E}_S \mathcal{L}(\hat{h}_S) = \mathbb{E}_{S_x \sim \mathcal{D}_x^n, x \sim \mathcal{D}_x, y \sim \eta(x), y' \sim \eta(\pi_1(x))} P(y \neq y')$ . Note that  $P(y \neq y') = \eta(x')(1 - \eta(x)) + (1 - \eta(x'))\eta(x) = (\eta - \eta + \eta')(1 - \eta) + (1 - \eta + \eta - \eta')\eta = 2\eta(1 - \eta) + (\eta - \eta')(2\eta - 1)$ . Since  $\eta$  is  $c$ -Lipschitz and  $|2\eta - 1| \leq 1$ ,  $P(y \neq y') \leq 2\eta(1 - \eta) + c\|x - x'\|$ . Substituting back,  $\mathbb{E}_S \mathcal{L}(\hat{h}_S) \leq 2\mathbb{E}_x \eta(x)(1 - \eta(x)) + c\mathbb{E}_{S,x} \|x - x_{\pi_1(x)}\|$ . The Bayes error  $\mathcal{L}(h^*) = \mathbb{E}_x \{\eta(x) \wedge (1 - \eta(x))\} \geq \mathbb{E}_x (\eta(x)(1 - \eta(x)))$ .  $\square$

- **Lemma 3.2** Let  $C_1, \dots, C_r$  be a collection of subsets of  $\mathcal{X}$ . Then  $\mathbb{E}_{S \sim \mathcal{D}^n} \{\sum_{i: C_i \cap S = \emptyset} P(C_i)\} \leq \frac{r}{ne}$  (“probability of subsets that not hit by  $S$ ”).

**Proof** By linearity,  $\mathbb{E}_S \{\sum_{i: C_i \cap S = \emptyset} P(C_i)\} = \sum_{i=1}^r P(C_i) \mathbb{E}_S I(C_i \cap S = \emptyset) = \sum_{i=1}^r P(C_i) P(C_i \cap S = \emptyset)$ . Note that  $P(C_i \cap S = \emptyset) = (1 - P(C_i))^n \leq e^{-nP(C_i)}$ . Thus, LHS  $\leq \sum_{i=1}^r P(C_i) e^{-nP(C_i)} \leq r \max P(C_i) e^{-nP(C_i)} \leq \frac{r}{ne}$ .  $\square$

- **Theorem 3.1** (Generalization upper bound for 1-NN)  $\mathbb{E}_S \mathcal{L}(\hat{h}_S) \leq 2\mathcal{L}(h^*) + 2c\sqrt{dn}^{-\frac{1}{d+1}}$ .

**Proof** Take  $C_i$  of the form  $\{x : x_j \in [(\alpha_j - 1)/T, \alpha_j/T], \forall j\}$ , where  $\alpha_1, \dots, \alpha_d \in \{1, \dots, T\}^d$ .

Case 1: If  $x, x' \in C_i$  for some  $i$ , then  $\|x - x'\| \leq \sqrt{d}\epsilon$ .

Case 2: Otherwise,  $\|x - x'\| \leq \sqrt{d}$ .

Hence,  $\mathbb{E}_{S,x} \|x - x_{\pi_1(x)}\| \leq \mathbb{E}_S \{P(\cup_{i: C_i \cap S \neq \emptyset} C_i) \sqrt{d}\epsilon + P(\cup_{i: C_i \cap S = \emptyset} C_i) \sqrt{d}\} \leq \sqrt{d}(\epsilon + \frac{r}{ne})$ . Since  $r = (\frac{1}{\epsilon})^d, \dots \leq \sqrt{d}(\epsilon + \frac{1}{\epsilon^d ne})$ . Matching the two terms gives  $\epsilon = (\frac{1}{ne})^{\frac{1}{d+1}}$  and the optimal bound  $2\sqrt{d}(ne)^{-\frac{1}{d+1}} \leq 2\sqrt{dn}^{-\frac{1}{d+1}}$ .  $\square$

- **Theorem 3.2** (Generalization upper bound for  $k$ -NN)  $\mathbb{E}_S \mathcal{L}(\hat{h}_S) \leq (1 + \sqrt{\frac{8}{k}})\mathcal{L}(h^*) + (6c\sqrt{d} + k)n^{-\frac{1}{d+1}}$ .

**Remark 3.1**  $k$  is called regularization parameter/hyperparameter and the optimal  $k \sim n^d$ .

**Remark 3.2** Exponential dependence on  $d$ : “curse of dimensionality”.

- **Theorem 3.3** (Lower bound)  $\forall c > 1$  and any learning rule  $h$ ,  $\exists$  a distribution over  $[0, 1]^d \times \{0, 1\}$  s.t.  $\eta(x)$  is  $c$ -Lipschitz, the Bayes error is 0, but for  $n < (c+1)^d/2$ ,  $\mathbb{E} \mathcal{L}(h) > \frac{1}{4}$  (i.e. minimax bound  $\inf_h \sup_y \mathbb{E} \mathcal{L}(h) \geq Cn^{-\frac{1}{d+1}}$ ).

**Hint** Let  $G_c^d$  be the regular grid on  $[0, 1]^d$  with distance  $1/c$  between points. Then any  $\eta : G_c^d \rightarrow \{0, 1\}$  is  $c$ -Lipschitz. Then use the following theorem.  $\square$

- **Theorem 3.4** (No free-lunch theorem) Let  $A$  be any learning rule for binary classification with 0-1 loss over  $\mathcal{X}^d$  and  $n < |\mathcal{X}|/2$ . Then  $\exists$  distribution  $D$  over  $\mathcal{X} \times \{0, 1\}$  s.t.  $\mathbb{E} \mathcal{L}(A) \geq \frac{1}{4}$ . Furthermore, with prob  $\geq \frac{1}{7}$ ,  $\mathcal{L}(A_S) \geq \frac{1}{8}$ .

## 4 Linear Regression

- $Y_{n \times 1} = X_{n \times p} \beta_{p \times 1} + \epsilon_{n \times 1}$ ,  $\mathbb{E}(\epsilon|X) = 0$ ,  $\text{Var}(\epsilon) = \sigma^2 I_n$  and  $X$  fixed.
- Least squares estimator (LSE) solves the normal equation  $X^T X \hat{\beta} = X^T Y$ ,  $\hat{\beta} = (X^T X)^{-1} X^T Y$ .
- ANOVA:  $y_{ij} = \mu + \alpha_j + \epsilon_{ij}$ ,  $i = 1, \dots, n_j$ ,  $j = 1, \dots, J$ .  $\sum_j n_j = n$ ,  $\sum_j \alpha_j = 0$ .
- **Definition 4.1**  $\theta$  is estimable if  $\exists$  an unbiased estimator of  $\theta$ .  $c^T \beta$  is linearly estimable if  $\exists l \in \mathbb{R}^n$  s.t.  $\mathbb{E}(l^T Y) = c^T \beta$ ,  $\forall \beta \in \mathbb{R}^p \Leftrightarrow c = X^T l \in \mathcal{C}(X^T)$ .
- **Theorem 4.1** (1) If  $c^T \hat{\beta}$  is unique, then  $c \in \mathcal{C}(X^T X) = \mathcal{C}(X^T)$ .  
 (2) If  $c \in \mathcal{C}(X^T)$ , then  $c^T \hat{\beta}$  is unique and unbiased for  $c^T \beta$ .  
 (3) If  $c^T \beta$  is estimable and  $\epsilon \sim \mathcal{N}_n(0, \sigma^2 I_n)$ , then  $c \in \mathcal{C}(X^T)$ .

**Proof** (1) Let  $b \in \mathcal{C}(X^T X)^\perp$  be arbitrary, then  $X^T Y = X^T X \hat{\beta} = X^T X(\hat{\beta} + b) \Rightarrow c^T \hat{\beta} = c^T(\hat{\beta} + b) \Rightarrow c^T b = 0$ .  
 (2)  $c = X^T l$  for some  $l \in \mathbb{R}^n$ , then  $c^T \hat{\beta} = l^T X^T \hat{\beta} = l^T X^T (X^T X)^{-1} X^T Y = l^T P_X Y$  is unique.  $\mathbb{E}(c^T \hat{\beta}) = l^T P_X \mathbb{E} Y = l^T P_X X \beta = l^T X \beta = c^T \beta$ .

(3) If  $\exists$  an estimator  $T(X, Y)$  unbiased for  $c^T \beta$ , then  $c^T \beta = \int T(X, y) \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\{-\frac{1}{2\sigma^2} \|y - X\beta\|^2\} dy$ . Differentiate with  $\beta$ ,  $c = X^T \int \frac{y - X\beta}{(2\pi\sigma^2)^{\frac{n}{2}} \sigma^2} T(X, y) \exp\{-\frac{1}{2\sigma^2} \|y - X\beta\|^2\} dy$ .  $\square$

**Remark 4.1**  $A\beta$  with  $A \in \mathbb{R}^{q \times p}$  is estimable iff  $\mathcal{C}(A^T) \subset \mathcal{C}(X^T) \Leftrightarrow A = A_* X$  for some  $A_* \in \mathbb{R}^{q \times n}$ . In particular,  $\beta$  is estimable iff  $X$  has full column.

- Ordinary least squares:  $\hat{\beta} = (X^T X)^{-1} X^T Y$ .
- **Proposition 4.1** For any estimable  $A\beta$  and  $B\beta$ ,  $\text{Cov}(A\hat{\beta}, B\hat{\beta}) = \sigma^2 A(X^T X)^{-1} B^T$ ,  $\text{Var}(A\hat{\beta}) = \sigma^2 A(X^T X)^{-1} A^T$ .

**Proof**  $\exists A_*$  and  $B_*$  s.t.  $A = A_* X, B = B_* X$ . Since  $\hat{Y} = X\hat{\beta} = X(X^T X)^{-1} X^T Y = P_X Y$ , we have  $\text{Var}(\hat{Y}) = P_X \text{Var}(Y) P_X^T = \sigma^2 P_X$ . Hence  $\text{Cov}(A\hat{\beta}, B\hat{\beta}) = \text{Cov}(A_* \hat{Y}, B_* \hat{Y}) = A_* \text{Var}(\hat{Y}) B_*^T = \sigma^2 A_* P_X B_*^T = A(X^T X)^{-1} B^T$ .  $\square$

- **Theorem 4.2** (Gauss-Markov) If  $c^T \beta$  is estimable, then  $c^T \hat{\beta}$  has the minimum variance among all linear unbiased estimates. (Best Linear Unbiased Estimator, BLUE)

**Proof** Let  $l^T Y$  be an unbiased estimator of  $c^T \beta$ . Hence,  $c = X^T l$ , so that  $c^T \hat{\beta} = l^T X \hat{\beta} = l^T \hat{Y}$ . Thus,  $\text{Var}(l^T Y) - \text{Var}(c^T \hat{\beta}) = l^T [\text{Var}(Y) - \text{Var}(\hat{Y})] l = \sigma^2 l^T (I - P_X) l \geq 0$ .  $\square$

- Residual  $\hat{\epsilon} = Y - \hat{Y} = (I - P_X)Y \in \mathcal{C}(X)^\perp$ ,  $\mathbb{E}(\hat{\epsilon} | (I - P_X)\mathbb{E}Y) = (I - P_X)X\beta = 0$ ,  $\text{Var}(\hat{\epsilon}) = \sigma^2 (I - P_X)^2 = \sigma^2 (I - P_X)$ ,  $\text{Cov}(\hat{\epsilon}, \hat{Y}) = \text{Cov}((I - P_X)Y, P_X Y) = (I - P_X)(\sigma^2 I)P_X = 0$ .
- Residual sum of squares (RSS):  $\|\hat{\epsilon}\|^2 = \hat{\epsilon}^T \hat{\epsilon} = Y^T (I - P_X) Y$ .  $\mathbb{E}(\text{RSS}) = \mathbb{E} \text{tr}(\hat{\epsilon} \hat{\epsilon}^T) = \text{tr}(\mathbb{E}(\hat{\epsilon} \hat{\epsilon}^T)) = \text{tr}\{(I - P_X)\sigma^2\} = \sigma^2(n - \text{rank}(X))$ .  $\hat{\sigma}^2 = \frac{\text{RSS}}{n-r}$  is an unbiased estimator of  $\sigma^2$ .
- Restricted LSE:  $Y = X\beta + \epsilon$ ,  $\mathbb{E}\epsilon = 0$ ,  $\text{Var}(\epsilon) = \sigma^2 I$ ,  $\text{rank}(X) = r$ ,  $X = \begin{pmatrix} X_1 & X_2 \end{pmatrix}$ ,  $\beta = \begin{pmatrix} \beta_1^T & \beta_2^T \end{pmatrix}^T$ .  $H_0 : \beta_2 = \beta_2^*$  vs  $\beta_2 \neq \beta_2^*$ .  $\beta_2$  is estimable  $\Rightarrow \text{rank}(X_2) = s$ ,  $\text{rank}(X_1) = r - s$  and  $\mathcal{C}(X_1) \cap \mathcal{C}(X_2) = \{0\}$ .

**Proof**  $\exists C \in \mathbb{R}^{q \times n}$  s.t.  $(0_{s \times (p-s)}, I_s) = CX = (CX_1, CX_2)$ . Hence  $\text{rank}(X_2) = s$  and  $\text{rank}(X_1) = r - s$ . If  $X_1 b_1 = X_2 b_2$  then  $b_2 = CX_1 b_1 = 0$ .  $\square$

- Under  $H_0 : \beta_2 = \beta_2^*$ ,  $Y = X_1 \beta_1 + X_2 \beta_2 + \epsilon$  becomes  $Y - X_2 \beta_2^* = X_1 \beta_1 + \epsilon$ . Restricted normal equation:  $X_1^T X_1 \tilde{\beta}_1 = X_1^T (Y - X_2 \beta_2^*)$ .  $\mathcal{C}(X_1) \subset \mathcal{C}(X) \Rightarrow P_{X_1} P_X = P_{X_1}$ . Since  $P_X Y = \hat{Y} = X \hat{\beta} = X_1 \hat{\beta}_1 + X_2 \hat{\beta}_2$ , we have  $X_1 \tilde{\beta}_1 = P_{X_1} (Y - X_2 \beta_2^*) = P_{X_1} (P_X Y - X_2 \beta_2^*) = P_{X_1} (X_1 \hat{\beta}_1 + X_2 (\hat{\beta}_2 - \beta_2^*)) = X_1 \hat{\beta}_1 + P_{X_1} X_2 (\hat{\beta}_2 - \beta_2^*)$ . Let  $\tilde{Y} = X_1 \tilde{\beta}_1 + X_2 \beta_2^*$  the fitted value of the restricted model.  $\hat{Y} - \tilde{Y} = X_1 \hat{\beta}_1 + X_2 \hat{\beta}_2 - [X_1 \hat{\beta}_1 + P_{X_1} X_2 (\hat{\beta}_2 - \beta_2^*)] - X_2 \beta_2^* = (I - P_{X_1}) X_2 (\hat{\beta}_2 - \beta_2^*)$ .
- **Theorem 4.3**  $\mathcal{C}(Z_2) = \mathcal{C}(X_1)^\perp \cap \mathcal{C}(X)$ , where  $Z_2 = (I - P_{X_1})X_2 = X_2 - P_{X_1} X_2$ .

**Proof**  $\mathcal{C}(Z_2) \subset \mathcal{C}(I - P_{X_1}) = \mathcal{C}(X_1)^\perp$ . Since  $\mathcal{C}(P_{X_1} X_2) \subset \mathcal{C}(X_1)$ ,  $\mathcal{C}(Z_2) = \mathcal{C}(X_2 - P_{X_1} X_2) \subset \mathcal{C}(X)$ . Conversely, if  $X = X_1 b_1 + X_2 b_2 \in \mathcal{C}(X)$  and  $X \perp \mathcal{C}(X_1)$ , then  $X = (I - P_{X_1})X = (I - P_{X_1})X_2 b_2 \in \mathcal{C}(Z_2)$ .  $\square$

**Corollary 4.1**  $P_{Z_2} = P_X - P_{X_1}$ .

- Now  $\hat{Y} - \tilde{Y} = (I - P_{X_1})[X_2(\hat{\beta}_2 - \beta_2^*) + X_1 \hat{\beta}_1] = (I - P_{X_1})(P_X Y - X_2 \beta_2^*) = (I - P_{X_1})P_X (Y - X_2 \beta_2^*) = P_{Z_2}(Y - X_2 \beta_2^*)$ . In view of  $\mathbb{R}^n = \mathcal{C}(X)^\perp \oplus \mathcal{C}(X)$ ,  $Y - \tilde{Y} = (Y - \hat{Y}) + (\hat{Y} - \tilde{Y})$ .  $\text{RSS}_{H_0} = \|Y - \tilde{Y}\|^2 = \|Y - \hat{Y}\|^2 + \|\hat{Y} - \tilde{Y}\|^2$ ,  $\text{RSS} = \|Y - \hat{Y}\|^2 = \|(I - P_X)Y\|^2 = \|(I - P_X)(Y - X_2 \beta_2^*)\|^2$ .  $\text{RSS}_{H_0} - \text{RSS} = \|\hat{Y} - \tilde{Y}\|^2 = \|Z_2(\hat{\beta}_2 - \beta_2^*)\|^2 = \|P_{Z_2}(Y - X_2 \beta_2^*)\|^2$ . By Cochran's theorem,  $\text{RSS}_{H_0} - \text{RSS} \sim \chi_s^2(\lambda)$  with  $\lambda = \|P_{Z_2}(X\beta - X_2 \beta_2^*)\|^2$ .
- Wald's statistics:  $(\hat{\theta} - \theta_0) \text{Var}(\hat{\theta})^{-1} (\hat{\theta} - \theta_0)$ . Since  $\beta_2$  is estimable,  $\exists C \in \mathbb{R}^{s \times n}$ ,  $(0_{s \times p-s}, I_s) = CX = (CX_1, CX_2) \Rightarrow CP_{X_1} = CX_1(X_1^T X_1)^{-1} X_1^T = 0$ ,  $CZ_2 = C(I_n - P_{X_1})X_2 = CX_2 - CP_{X_1} X_2 = I_s \Rightarrow Z_2$  has full column rank.  $\hat{\beta}_2 = (0, I)\hat{\beta} = CX\hat{\beta} = CP_X Y = C(P_{X_1} + P_{Z_2})Y = CP_{Z_2} Y$ . Thus,  $\text{Var}(\hat{\beta}_2) = \text{Var}(CP_{Z_2} Y) = CP_{Z_2} \sigma^2 I_n P_{Z_2} C^T = \sigma^2 CZ_2(Z_2^T Z_2)^{-1} Z_2^T C^T = \sigma^2(Z_2^T Z_2)^{-1}$ .  $(\hat{\beta}_2 - \beta_2^*) \text{Var}(\hat{\beta}_2)^{-1} (\hat{\beta}_2 - \beta_2^*) = \|Z_2(\hat{\beta}_2 - \beta_2^*)\|^2 / \sigma^2 = \frac{\text{RSS}_{H_0} - \text{RSS}}{\sigma^2}$ .

- Inference:  $H = (h_1, \dots, h_s) \in \mathbb{R}^{p \times s}, \xi = \mathbb{R}^s$ . General linear hypothesis:  $H_0 : H^T \beta = \xi$  ( $s$  constraints). Assume (1)  $\mathcal{C}(H) \subset \mathcal{C}(X^T)$ , so that  $H^T \beta$  is estimable; (2)  $H$  has full column rank,  $s = \text{rank}(H) \leq \text{rank}(X) = r \leq p$ .
- Reparameterization: Choose  $A \in \mathbb{R}^{p \times (p-s)}$  s.t.  $\mathcal{C}(A) = \mathcal{C}(H)^\perp$ . Let  $\theta = \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} = \begin{pmatrix} A^T \beta \\ H^T \beta \end{pmatrix}$  and  $\tilde{X} = X \begin{pmatrix} A^T \\ H^T \end{pmatrix}^{-1} = (\tilde{X}_1, \tilde{X}_2)$ . The reparameterized model  $Y = \tilde{X}\theta + \epsilon$ . Since  $\mathcal{C}(\tilde{X}^T) = \mathcal{C}((A, H)^{-1}X^T) \supset \mathcal{C}((A, H)^{-1}H) = \mathcal{C}\begin{pmatrix} 0 \\ I_s \end{pmatrix}$ ,  $\theta_2$  is estimable.  $\hat{\theta}$  solves the normal equation  $\tilde{X}^T \tilde{X} \hat{\theta} = \tilde{X}^T Y$ . Under  $H_0$ ,  $\tilde{Y} = \tilde{X}_1 \tilde{\theta}_1 + \tilde{X}_2 \xi = \tilde{X}_1 \hat{\theta}_1 + P_{\tilde{X}_1} \tilde{X}_2 (\hat{\theta}_2 - \xi) + \tilde{X}_2 \xi$ ,  $\text{RSS}_{H_0} - \text{RSS} = \|Y - \tilde{Y}\|^2 - \|Y - \hat{Y}\|^2 = \|\hat{Y} - \tilde{Y}\|^2 = \sigma^2 (\hat{\theta}_2 - \xi)^T \text{Var}(\hat{\theta}_2)^{-1} (\hat{\theta}_2 - \xi)$ . Substituting into the original model,  $\hat{\theta}_2 = H^T \hat{\beta}$ ,  $\text{Var}(\hat{\theta}_2) = \sigma^2 H^T (X^T X)^{-1} H$ . Since  $\mathbb{E}(X^T A X) = \text{tr}(A \Sigma) + \mu^T A \mu$  where  $\mu = \mathbb{E}X, \Sigma = \text{Var}(X)$ ,  $\mathbb{E}\|\hat{Y} - \tilde{Y}\|^2 / \sigma^2 = \text{tr}(\text{Var}(\hat{\theta}_2)^{-1} \text{Var}(\hat{\theta}_2)) + (H^T \beta - \xi)^T \text{Var}(H^T \beta)^{-1} (H^T \beta - \xi)$ .  $Y - \hat{Y} = (I_n - P_{\tilde{X}})(Y - \tilde{X}_2 \xi)$ ,  $\hat{Y} - \tilde{Y} = \tilde{Z}_2 (H^T \hat{\beta} - \xi) = P_{\tilde{Z}_2} (Y - \tilde{X}_2 \xi)$ . By Cochran's thm,  $\frac{\|Y - \hat{Y}\|^2}{\sigma^2} \sim \chi_{n-r}^2$  and  $\frac{\|\hat{Y} - \tilde{Y}\|^2}{\sigma^2} \sim \chi_s^2(\lambda)$  are independent with  $\lambda = (H^T \beta - \xi)^T \text{Var}(H^T \beta)^{-1} (H^T \beta - \xi)$ . Hence,  $\frac{(\text{RSS}_{H_0} - \text{RSS})/s}{\text{RSS}/(n-r)} \sim F_{s, n-r}(\lambda)$ .
- Let  $\gamma = H^T \beta$  and  $\gamma_0 = \xi$ . Test  $H_0 : \gamma = \gamma_0$  can be regarded as a weighted distance between  $\hat{\gamma}$  and  $\gamma_0$ . To see this, let  $\hat{\gamma} = H^T \hat{\beta} \sim \mathcal{N}_s(\gamma, \sigma^2 D)$  where  $D = H^T (X^T X)^{-1} H$  and  $\hat{\sigma}^2 = \frac{\text{RSS}}{n-r}$ . Under  $H_0$ , (1)  $s = 1$ :  $Z = \frac{\hat{\gamma} - \gamma_0}{\hat{\sigma} \sqrt{D}} \sim \mathcal{N}(0, 1)$  if  $\sigma^2$  is known;  $T = \frac{\hat{\gamma} - \gamma_0}{\hat{\sigma} / \sqrt{D}} \sim t_{n-r}$  if  $\sigma^2$  is unknown. Confidence interval:  $\hat{\gamma} \pm t_{n-r, \alpha/2} \hat{\sigma} \sqrt{D}$ . (2)  $s \geq 1$ : Mahalanobis distance  $\|\hat{\gamma} - \gamma_0\|_{(\sigma^2 D)^{-1}} = \sqrt{(\hat{\gamma} - \gamma_0)^T (\sigma^2 D)^{-1} (\hat{\gamma} - \gamma_0)}$ ,  $\|\hat{\gamma} - \gamma_0\|_{(\sigma^2 D)^{-1}}^2 = (\hat{\gamma} - \gamma_0)^T (\sigma^2 D)^{-1} (\hat{\gamma} - \gamma_0) \sim \chi_s^2(\lambda)$  where  $\lambda = (\gamma - \gamma_0)^T D^{-1} (\gamma - \gamma_0) / \sigma^2$ . Thus  $\mathbb{E}(\hat{\gamma} - \gamma_0)^T D^{-1} (\hat{\gamma} - \gamma_0) / s = (s + \lambda) \sigma^2 / s = (1 + \lambda/s) \sigma^2 \geq \sigma^2$  with equality holding just when  $\gamma = \gamma_0$ . One may reject  $H_0$  if  $(\hat{\gamma} - \gamma_0)^T D^{-1} (\hat{\gamma} - \gamma_0) / (s \sigma^2)$  is large. If  $\sigma^2$  is unknown, replacing  $\sigma^2$  with  $\hat{\sigma}^2$  yields  $\frac{(\hat{\gamma} - \gamma_0)^T D^{-1} (\hat{\gamma} - \gamma_0)}{s \hat{\sigma}^2} = \frac{\|\hat{Y} - \tilde{Y}\|^2 / s}{\|Y - \hat{Y}\|^2 / (n-r)} \sim F_{s, n-r}(\lambda)$ , where  $\lambda = 0$  iff  $H_0$  is true.
- Multiple testing: Simultaneous confidence intervals of level  $1 - \alpha$ .
- Bonferroni: Replace  $\alpha$  by  $\alpha/m$ :  $P(E_j) = 1 - \alpha_j, j = 1, \dots, m$ , then  $P(\cap_j E_j) = 1 - P(\cup_j E_j^c) \geq 1 - \sum_j P(E_j) = 1 - \sum_j \alpha_j = 1 - \alpha$ .
- Scheffé's method: Consider  $Y = X\beta + \epsilon, \epsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ ,  $\text{rank}(X) = r$  and test for  $u^T \gamma, \forall u \in \mathbb{R}^s$ , where  $\gamma = H^T \beta$  is estimable and  $H$  is of full column rank.  $\hat{\gamma} = H^T \hat{\beta} \sim \mathcal{N}_s(\gamma, \sigma^2 D)$  where  $D = H^T (X^T X)^{-1} H$ ,  $\hat{\sigma}^2 = \frac{\text{RSS}}{n-r} \sim \sigma^2 \chi_{n-r}^2$ . For any fixed  $u \in \mathbb{R}^s$ , an  $(1 - \alpha)$  CI for  $u^T \gamma : u^T \hat{\gamma} \pm t_{n-r, \frac{\alpha}{2}} \hat{\sigma} \sqrt{u^T D u}$ . Now allow  $u \in \mathbb{R}^s$  to vary arbitrarily. Since  $\sup_{u \neq 0} \frac{|u^T \hat{\gamma} - u^T \gamma|^2}{u^T D u} \stackrel{v=D^{\frac{1}{2}} u}{=} \sup_{v \neq 0} \frac{|v^T D^{-\frac{1}{2}} (\hat{\gamma} - \gamma)|^2}{v^T v} \stackrel{\text{Cauchy-Schwarz}}{=} (\hat{\gamma} - \gamma)^T D^{-1} (\hat{\gamma} - \gamma), P(\sup_{u \neq 0} \frac{|u^T \hat{\gamma} - u^T \gamma|^2}{s \hat{\sigma}^2 u^T D u} \leq F_{s, n-r, \alpha}) = 1 - \alpha$ . Simultaneous CIs for  $u^T \gamma, \forall u \in \mathbb{R}^s : u^T \hat{\gamma} \pm \hat{\sigma} \sqrt{s F_{s, n-r, \alpha} u^T D u}$ . (Bonferroni:  $t_{n-r, \alpha/(2m)}$ )
- Tukey's method: Consider  $y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$  i.i.d.,  $j = 1, \dots, m, i = 1, \dots, k$  and test for  $\alpha_i - \alpha_{i'}, \forall i, i' = 1, \dots, k$ . If  $Z_1, \dots, Z_n \sim \mathcal{N}(0, 1), R^2 \sim \chi_v^2$ , then  $\frac{Z_{(n)} - Z_{(1)}}{\sqrt{R^2/v}} \sim q_{n,v}$  (studentized range distribution). Thus  $\frac{\sqrt{m}}{\hat{\sigma}} \max_{i, i'} \{\bar{y}_i - \bar{y}_{i'} - (\alpha_i - \alpha_{i'})\} = \frac{\{\max_i \frac{\sqrt{m}(\bar{y}_i - \mu - \alpha_i)}{\hat{\sigma}} - \min_i \frac{\sqrt{m}(\bar{y}_i - \mu - \alpha_i)}{\hat{\sigma}}\}}{\sqrt{\frac{\text{RSS}/\sigma^2}{n-k}}} \sim q_{k, n-k}$ . Simultaneous CIs:  $\bar{y}_i - \bar{y}_{i'} \pm \frac{\hat{\sigma}}{\sqrt{m}} q_{k, n-k, \alpha}$ . (Bonferroni:  $t_{n-k, \alpha/[k(k-1)]}$ , Scheffé:  $\sqrt{k F_{k, n-k, \alpha}}$ , Tukey:  $q_{k, n-k, \alpha} / \sqrt{2}$  (the best/shortest length))

## 5 Generalized Linear Models

- Data types for response  $y$  : 
$$\left\{ \begin{array}{l} \text{numerical:} \left\{ \begin{array}{l} \text{continuous: Box-Cox transformation:} \left\{ \begin{array}{ll} \frac{x^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \log x, & \lambda = 0 \end{array} \right. \\ \text{discrete: count} \end{array} \right. \\ \text{categorical:} \left\{ \begin{array}{l} \text{nominal:} \left\{ \begin{array}{l} \text{binary} \\ \text{multinomial} \end{array} \right. \\ \text{ordinal} \end{array} \right. \end{array} \right.$$

- Example 1 (Dilution assays): density  $\rho_0$ , at the  $x$ -th dilution  $\rho_x = \rho_0 2^{-x}$ ,  $x = 0, 1, 2, \dots$ , proportion of infected plates  $y_x = \frac{r_x}{m_x}$ ,  $Y = I(\text{infected})$ ,  $\mathbb{E}(Y|x) = P(Y = 1|x) = \pi_x$ , # organism on a plate:  $N_x \sim \text{Poisson}(\rho_x v)$ ,  $\pi_x = P(N_x \geq 1) = 1 - e^{-\rho_x v} = 1 - e^{-\rho_0 v 2^{-x}}$ , link function  $g(\pi_x) = \log(-\log(1 - \pi_x)) = \log v + \log \rho_0 - x \log 2$ .
- Example 2 (Dose response): dose level  $x$ , survival rate  $\pi_x$ , cell  $j$ , dose level  $x_j$ ,  $y_j$  survive out of  $m_j$  animals. (1) Probit model:  $\pi_x = \Phi(\alpha + \beta x)$ , where  $\Phi$  is the c.d.f. of  $\mathcal{N}(0, 1)$ , link function  $g = \Phi^{-1}$ . (2) Logisitic/Logit model:  $\pi_x = \text{expit}(\alpha + \beta x) = \frac{1}{1 + e^{-(\alpha + \beta x)}}$ , link function  $g(\pi_x) = \text{logit}(\pi_x) = \log \frac{\pi_x}{1 - \pi_x}$ .