# Modern Statistical Modeling

Lectured by Wei Lin LaTeXed by Chengxin Gong

2023 年 3 月 21 日

## 目录

# 1 Review of Linear Algebra

- Rank of $A \in \mathbb{R}^{m\times n}$: max # of linearly independent row/columns. Facts: (i) $0 \le \text{rank}(A) \le \min(m,n)$; (ii) $\text{rank}(A) = \text{rank}(A^T) = \text{rank}(AA^T) = \text{rank}(A^TA)$; (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) $\text{rank}(A) = \text{dim}\mathcal{C}(A)$; (ii) $\text{dim}\mathcal{C}(A) + \text{dim}\mathcal{N}(A) = n$; (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: $\langle A, B\rangle = \text{tr}(A^TB) = \text{tr}(AB^T) = \sum_i \sum_j a_{ij}b_{ij}$. It induces Frobenius norm: $||A||_F = \sqrt{\langle A, A\rangle} = (\sum_{i,j} a_{ij})^{1/2}$.

- 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 ($V^TV = VV^T = I$) and $\Lambda = \text{diag}(\lambda_1, \cdots, \lambda_n)$. (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 $\Sigma = \text{diag}(\sigma_1, \cdots, \sigma_r, 0, \cdots, 0)$,$\sigma_1 \ge \cdots \ge \sigma_r \ge 0$ and $\{u_i\}, \{v_i\}$ orthonormal. $\arg\min_{Y\in\mathbb{R}^{n\times p}, \text{rank}(Y)\le r} ||X - Y||_F = \sum_{i=1}^r \sigma_i u_i v_i^T$ (low rank-$r$ approximation). (3) QR decomposition: $A = QR$ where $Q$ is orthonormal and $R$ is upper-triangular. It corresponds to Garm-Schmidt orthogonalization process.

- Idempotent: $P^T = P$. Facts: (i) If $P$ is symmetric, then $P$ is idempotent of rank $r$ iff it has $r$ eignevalues 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 $P_X = X(X^TX)^-X^T$ is the orthogonal projection onto $\mathcal{C}(X)$. [$P_X$ does not depend on the choice of $(X^TX)^-$]

  **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, $P_Xv = P_Xx + P_Xw = P_Xx + X(X^TX)^-X^Tw = P_Xx$. We need to show $u^TX(X^TX)^-X^TX = u^TX, \forall u \in \mathbb{R}^n$.

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

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

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

# 2 Review of Probability Theory

- Distribution related to multivariate normal: $X \sim \mathcal{N}_p(\mu, \Sigma)$. Moment generating function: $M_X(t) = \mathbb{E}e^{t^TX} = \exp(t^T\mu + \frac{1}{2}t^T\Sigma t)$. Characteristic function: $\phi_X(t) = \mathbb{E}e^{it^TX} = \exp(it^T\mu - \frac{1}{2}t^T\Sigma t)$. Facts: (i) $A_{g\times p}X + b_{g\times 1} \sim \mathcal{N}_g(A\mu + b, A\Sigma A^T)$; (ii) $X \sim \mathcal{N}_p(\mu, \Sigma) \Leftrightarrow a^TX \sim \mathcal{N}(a^T\mu, a^T\Sigma a), \forall a \in \mathbb{R}^p$; (iii) $Y_1 = A_1X + b_1 \perp\!\!\!\perp Y_2 = A_2 + b_2 \Leftrightarrow \text{Cov}(Y_1, Y_2) = A_1\Sigma A_2^T = 0$.

- Noncentral $\chi^2$: $X \sim \mathcal{N}_p(\mu, I_p)$. Then $X^TX \sim \chi_p^2(\lambda)$ with noncenteral 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 Poisson($\frac{\lambda}{2}$)-weighted mixture of $\chi_{p+2k}^2$. M.g.f.: $M_X(t; p, \lambda) = \frac{1}{(1-2t)^{p/2}}\exp(\frac{\lambda t}{1-2t})$. 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 $(X-\mu)^T\Sigma^{-1}(X-\mu) \sim \chi^2_p$ and $X^T\Sigma^{-1}X \sim \chi^2_p(\mu^T\Sigma^{-1}\mu)$; (ii) Additivity: If $X \sim \chi^2_{p_i}(\lambda_i)$ independent for $i = 1, \cdots, k$, then $\sum_{i=1}^n X_i \sim \chi^2_{\sum_i p_i}(\sum_i \lambda_i)$; (iii) Rank deficient: If $X \sim \mathcal{N}_p(\mu, I_p), A \in \mathbb{R}^{p\times p}$ symmetric, then $X^TAX \sim \chi^2_p(\lambda)$ with $\lambda = \mu^TA\mu \Leftrightarrow A$ 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 $X^TAX \perp\!\!\!\perp BX \Leftrightarrow B\Sigma A = 0_{q\times p}$; (v) $X^TAX \perp\!\!\!\perp X^TBX \Leftrightarrow A\Sigma B = 0_{p\times p}$.

- **Theorem** 2.1 (Cochran) $X \sim \mathcal{N}_p(\mu, I_p), X^TX = X^TA_1X + \cdots + X^TA_kX \equiv Q_1 + \cdots + Q_k, A_i \in \mathbb{R}^{p\times p}$ symmetric of rank $r_i$. Then $Q_i \sim \chi^2_{r_i}(\lambda_i)$ independent for $i = 1, \cdots, k \Leftrightarrow p = r_1 + \cdots + r_k$. In this case, $\lambda_i = \mu^TA_i\mu$ and $\lambda_1 + \cdots + \lambda_k = \mu^T\mu$.

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

  "$\Rightarrow$": $X^TA_i \sim \chi^2_{r_i}(\lambda_i)$ independent $\Rightarrow X^TX = \sum_i X^TA_iX \sim \chi^2_{\sum_i r_i}(\sum_i \lambda_i) \Rightarrow \sum_i r_i = p$. □

- Noncentral $F$: If $Q_1 \sim \chi^2_p(\lambda)$ and $Q_2 \sim \chi^2_q$ 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 $U_2 \sim \chi^2_q$ 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 $f : \mathcal{X} \to \mathcal{Y}$ that minimizes $\mathbb{E}_{X,Y}\mathcal{L}(f(X), Y) = \mathbb{E}\{\mathbb{E}(\mathcal{L}(f(X), Y)|X)\}$ where loss function $\mathcal{L} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$.

- 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 $\mathcal{L}(\hat{y}, y) = I(\hat{y} = y)$, $\mathbb{E}\{\mathcal{L}(h(X), Y)|X = x\} = \sum_{j\neq h(x)} P(Y = j|X = x) = 1 - P(Y = h(X)|X = x)$. 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: $h_{k\text{-NN}}(x) =$ 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 $\pi_1(x), \cdots, \pi_n(x)$ be an ordering of $\{1, \cdots, n\}$ 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 $C_1,\cdots,C_r$ 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) $\mathbb{E}_S\mathcal{L}(\hat{h}_S)\leq 2\mathcal{L}(h^*) + 2c\sqrt{d}n^{-\frac{1}{d+1}}$.

  **Proof** Take $C_i$ of the form $\{x : x_j\in[(\alpha_j-1)/T, \alpha_j/T], \forall j\}$, where $\alpha_1,\cdots,\alpha_d\in\{1,\cdots,T\}^d$.
  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, $\mathbb{E}_{S,x}||x-x_{\pi_1(x)}||\leq\mathbb{E}_S\{P(\cup_{i:C_i\cap S\neq\emptyset}C_i)\sqrt{d}\epsilon + P(\cup_{i:C_i\cap S=\emptyset})\sqrt{d}\}\leq\sqrt{d}(\epsilon+\frac{r}{ne})$. Since $r = (\frac{1}{\epsilon})^d$, $\cdots\leq\sqrt{d}(\epsilon+\frac{1}{\epsilon^d ne})$. Matching the two terms gives $\epsilon = (\frac{1}{ne})^{\frac{1}{d+1}}$ and the optimal bound $2\sqrt{d}(ne)^{-\frac{1}{d+1}}\leq 2\sqrt{d}n^{-\frac{1}{d+1}}$. $\square$

- **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 $\eta : G_c^d\to\{0,1\}$ 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

- $Y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \epsilon_{n\times 1}$, $\mathbb{E}(\epsilon|X) = 0$, $\mathrm{Var}(\epsilon) = \sigma^2 I_n$ and $X$ fixed.

- Least squares estimator (LSE) solves the normal equation $X^TX\hat{\beta} = X^TY, \hat{\beta} = (X^TX)^-X^TY$.

- ANOVA: $y_{ij} = \mu+\alpha_j+\epsilon_{ij}, i = 1,\cdots,n_j, j = 1,\cdots,J$. $\sum_j n_j = n, \sum_j\alpha_j = 0$.

- **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. $\mathbb{E}(l^TY) = c^T\beta, \forall\beta\in\mathbb{R}^p\Leftrightarrow c = X^Tl\in\mathcal{C}(X^T)$.

- **Theorem** 4.1 (1) If $c^T\hat{\beta}$ is unique, then $c\in\mathcal{C}(X^TX) = \mathcal{C}(X^T)$.
  (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 $b\in\mathcal{C}(X^TX)^\perp$ be arbitrary, then $X^TY = X^TX\hat{\beta} = X^TX(\hat{\beta}+b)\Rightarrow c^T\hat{\beta} = c^T(\hat{\beta}+b)\Rightarrow c^Tb = 0$.
  (2) $c = X^Tl$ for some $l\in\mathbb{R}^n$, then $c^T\hat{\beta} = lX^T\hat{\beta} = lX^T(X^TX)^-X^TY = lP_XY$ is unique. $\mathbb{E}(c^T\hat{\beta}) = l^TP_x\mathbb{E}Y = l^TP_XX\beta = l^TX\beta = c^T\beta$.

(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: $\hat{\beta} = (X^TX)^-X^TY$.
- **Proposition** 4.1 For any estimable $A\beta$ and $B\beta$, $\mathrm{Cov}(A\hat{\beta},B\hat{\beta}) = \sigma^2A(X^TX)^-B^T$, $\mathrm{Var}(A\hat{\beta}) = \sigma^2A(X^TX)^-A^T$.

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

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

- Residual $\hat{\epsilon} = Y-\hat{Y} = (I-P_X)Y\in\mathcal{C}(X)^\perp$, $\mathbb{E}\hat{\epsilon}(I-P_X)\mathbb{E}Y = (I-P_X)X\beta = 0$, $\mathrm{Var}(\hat{\epsilon}) = \sigma^2(I-P_X)^2 = \sigma^2(I-P_X)$, $\mathrm{Cov}(\hat{\epsilon},\hat{Y}) = \mathrm{Cov}((I-P_X)Y,P_XY) = (I-P_X)(\sigma^2I)P_X = 0$.
- Residual sum of squares (RSS): $||\hat{\epsilon}||^2 = \hat{\epsilon}^T\hat{\epsilon} = Y^T(I-P_X)Y$. $\mathbb{E}(\mathrm{RSS}) = \mathbb{E}\mathrm{tr}(\hat{\epsilon}\hat{\epsilon}^T) = \mathrm{tr}(\mathbb{E}(\hat{\epsilon}\hat{\epsilon}^T)) = \mathrm{tr}\{(I-P_X)\sigma^2\} = \sigma^2(n-\mathrm{rank}(X))$. $\hat{\sigma}^2 = \frac{\mathrm{RSS}}{n-r}$ is an unbiased estimator of $\sigma^2$.
- Restricted LSE: $Y = X\beta+\epsilon, \mathbb{E}\epsilon = 0, \mathrm{Var}(\epsilon) = \sigma^2I, \mathrm{rank}(X) = r, X = (\underset{n\times(p-s)}{X_1}, \underset{n\times s}{X_2}), \beta = (\underset{1\times(p-s)}{\beta_1^T}, \underset{1\times s}{\beta_2^T})^T$. $H_0$ : $\beta_2 = \beta_2^*$ vs $\beta_2\neq\beta_2^*$. $\beta_2$ is estimable $\Rightarrow\mathrm{rank}(X_2) = s, \mathrm{rank}(X_1) = r-s$ 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 $\mathrm{rank}(X_2) = s$ and $\mathrm{rank}(X_1) = r-s$. If $X_1b_1 = X_2b_2$ then $b_2 = CX_1b_1 = 0$. $\square$

- 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: $X_1^TX_1\tilde{\beta}_1 = X_1^T(Y-X_2\beta_2^*)$. $\mathcal{C}(X_1)\subset\mathcal{C}(X)\Rightarrow P_{X_1}P_X = P_{X_1}$. Since $P_XY = \hat{Y} = X\hat{\beta} = X_1\hat{\beta}_1+X_2\hat{\beta}_2$, we have $X_1\tilde{\beta}_1 = P_{X_1}(Y-X_2\beta_2^*) = P_{X_1}(P_XY-X_2\beta_2^*) = P_{X_1}(X_1\hat{\beta}_1+X_2(\hat{\beta}_2-\beta_2^*)) = X_1\hat{\beta}_1+P_{X_1}X_2(\hat{\beta}_2-\beta_2^*)$. Let $\tilde{Y} = X_1\tilde{\beta}_1+X_2\beta_2^*$ the fitted valued 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 $X = X_1b_1+X_2b_2\in\mathcal{C}(X)$ and $X\perp\mathcal{C}(X_1)$, then $X = (I-P_{X_1})X = (I-P_{X_1})X_2b_2\in\mathcal{C}(Z_2)$. $\square$

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

- Now $\hat{Y}-\tilde{Y} = (I-P_{X_1})[X_2(\hat{\beta}_2-\beta_2^*)+X_1\hat{\beta}_1] = (I-P_{X_1})(P_XY-X_2\beta_2^*) = (I-P_{X_1})P_X(Y-X_2\beta_2^*) = P_{Z_2}(Y-X_2\beta_2^*)$. In view of $\mathbb{R}^n = \mathcal{C}(X)^\perp\oplus\mathcal{C}(X)$, $Y-\tilde{Y} = (Y-\hat{Y})+(\hat{Y}-\tilde{Y})$. $\mathrm{RSS}_{H_0} = ||Y-\tilde{Y}||^2 = ||Y-\hat{Y}||^2+||\hat{Y}-\tilde{Y}||^2$, $\mathrm{RSS} = ||Y-\hat{Y}||^2 = ||(I-P_X)Y||^2 = ||(I-P_X)(Y-X_2\beta_2^*)||^2$. $\mathrm{RSS}_{H_0}-\mathrm{RSS} = ||\hat{Y}-\tilde{Y}||^2 = ||Z_2(\hat{\beta}_2-\beta_2^*)||^2 = ||P_{Z_2}(Y-X_2\beta_2^*)||^2$. By Cochran's theorem, $\mathrm{RSS}_{H_0}-\mathrm{RSS}\sim\chi_s^2(\lambda)$ with $\lambda = ||P_{Z_2}(X\beta-X_2\beta_2^*)||^2$.
- Wald's statistics: $(\hat{\theta}-\theta_0)\mathrm{Var}(\hat{\theta})^{-1}(\hat{\theta}-\theta_0)$. Since $\beta_2$ is estimable, $\exists C\in\mathbb{R}^{s\times n}$, $(0_{s\times p-s}, I_s) = CX = (CX_1, CX_2)\Rightarrow CP_{X_1} = CX_1(X_1^TX_1)^-X_1^T = 0, CZ_2 = C(I_n-P_{X_1})X_2 = CX_2-CP_{X_1}X_2 = I_s\Rightarrow Z_2$ has full column rank. $\hat{\beta}_2 = (0,I)\hat{\beta} = CX\hat{\beta} = CP_XY = C(P_{X_1}+P_{Z_2})Y = CP_{Z_2}Y$. Thus, $\mathrm{Var}(\hat{\beta}_2) = \mathrm{Var}(CP_{Z_2}Y) = CP_{Z_2}\sigma^2I_nP_{Z_2}C^T = \sigma^2CZ_2(Z_2^TZ_2)^-Z_2^TC^T = \sigma^2(Z_2^TZ_2)^{-1}$. $(\hat{\beta}_2-\beta_2^*)\mathrm{Var}(\hat{\beta}_2)^{-1}(\hat{\beta}_2-\beta_2^*) = ||Z_2(\hat{\beta}_2-\beta_2^*)||^2/\sigma^2 = \frac{\mathrm{RSS}_{H_0}-\mathrm{RSS}}{\sigma^2}$.

- Inference: $H = (h_1, \cdots, h_s) \in \mathbb{R}^{p\times s}, \xi = \mathbb{R}^s$. 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, $s = \text{rank}(H) \le \text{rank}(X) = r \le p$.

- 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 $\tilde{X}^T\tilde{X}\hat{\theta} = \tilde{X}^TY$. Under $H_0$, $\tilde{Y} = \tilde{X}_1\tilde{\theta}_1 + \tilde{X}_2\xi = \tilde{X}_1\hat{\theta}_1 + P_{\tilde{X}_1}\tilde{X}_2(\hat{\theta}_2 - \xi) + \tilde{X}_2\xi$, $\text{RSS}_{H_0} - \text{RSS} = ||Y-\tilde{Y}||^2 - ||Y-\hat{Y}||^2 = ||\hat{Y}-\tilde{Y}||^2 = \sigma^2(\hat{\theta}_2-\xi)\text{Var}(\hat{\theta}_2)^{-1}(\hat{\theta}_2-\xi)$. Substituting into the original model, $\hat{\theta}_2 = H^T\hat{\beta}, \text{Var}(\hat{\theta}_2) = \sigma^2H^T(X^TX)^-H$. Since $\mathbb{E}(X^TAX) = \text{tr}(A\Sigma) + \mu^TA\mu$ 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, $\frac{||Y-\hat{Y}||^2}{\sigma^2} \sim \chi^2_{n-r}$ and $\frac{||\hat{Y}-\tilde{Y}||^2}{\sigma^2} \sim \chi^2_s(\lambda)$ 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 been regarded as a weighted distance between $\hat{\gamma}$ and $\gamma_0$. To see this, let $\hat{\gamma} = H^T\hat{\beta} \sim \mathcal{N}_s(\gamma, \sigma^2D)$ where $D = H^T(X^TX)^-H$ and $\hat{\sigma}^2 = \frac{\text{RSS}}{n-r}$. Under $H_0$, (1) $s = 1$: $Z = \frac{\hat{\gamma}-\gamma_0}{\sigma\sqrt{D}} \sim \mathcal{N}(0,1)$ 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) $s \ge 1$: Mahalanobis distance $||\hat{\gamma}-\gamma_0||_{(\sigma^2D)^{-1}} = \sqrt{(\hat{\gamma}-\gamma_0)^T(\sigma^2D)^{-1}(\hat{\gamma}-\gamma_0)}$, $||\hat{\gamma}-\gamma_0||^2_{(\sigma^2D)^{-1}} = (\hat{\gamma}-\gamma_0)^T(\sigma^2D)^{-1}(\hat{\gamma}-\gamma_0) \sim \chi^2_s(\lambda)$ where $\lambda = (\gamma-\gamma_0)^TD^{-1}(\gamma-\gamma_0)/\sigma^2$. Thus $\mathbb{E}(\hat{\gamma}-\gamma_0)^TD^{-1}(\hat{\gamma}-\gamma_0)/s = (s+\lambda)\sigma^2/s = (1+\lambda/s)\sigma^2 \ge \sigma^2$ with equality holding just when $\gamma = \gamma_0$. One may reject $H_0$ if $(\hat{\gamma}-\gamma_0)^TD^{-1}(\hat{\gamma}-\gamma_0)/(s\sigma^2)$ is large. If $\sigma^2$ is unknown, replacing $\sigma^2$ with $\hat{\sigma}^2$ yields $\frac{(\hat{\gamma}-\gamma_0)^TD^{-1}(\hat{\gamma}-\gamma_0)}{s\hat{\sigma}^2} = \frac{||\hat{Y}-\tilde{Y}||^2/s}{||Y-\hat{Y}||^2/(n-r)} \sim F_{s,n-r}(\lambda)$, where $\lambda = 0$ iff $H_0$ is true.

# 5 Multiple Testing

- Simultaneous confidence intervals of level $1-\alpha$.

- Bonferroni: replace $\alpha$ by $\alpha/m$: $P(E_j) = 1-\alpha_j, j = 1, \cdots, m$, then $P(\cap_jE_j) = 1 - P(\cup_jE_j^c) \ge 1 - \sum_j P(E_j) = 1 - \sum_j \alpha_j = 1-\alpha$.

- Scheffé's method: