Modern Statistical Modeling

Lectured by Wei Lin LaTeXed by Chengxin Gong

June 15, 2023

Contents

1 Review of Linear Algebra and Probability

- 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) $\text{rank}(A) = \dim\mathcal{C}(A)$; (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: $\langle A, B\rangle = \text{tr}(A^TB) = \text{tr}(AB^T) = \sum_i \sum_j a_{ij}b_{ij}$. 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 $x \in \mathcal{C}(X), w \in \mathcal{C}(X)^\perp$. 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) $Y_1 = A_1X + b_1 \perp\!\!\!\perp Y_2 = A_2X + b_2 \Leftrightarrow \text{Cov}(Y_1, Y_2) = A_1\Sigma A_2^T = 0$.

- Noncentral χ^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 χ_{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 $\Rightarrow X^TX = \sum_i X^TA_iX \sim \chi^2_{\sum_i r_i}(\sum_i \lambda_i) \Rightarrow \sum_i r_i = p$. $\square$

- 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 $\mathcal{X} \subset [0, 1]^d$, ρ 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 η: η 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 η 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 $\mathcal{L}(h^*) = \mathbb{E}_x\{\eta(x) \wedge (1-\eta(x))\} \geq \mathbb{E}_x(\eta(x)(1-\eta(x)))$. □

- **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 $\leq \sum_{i=1}^r P(C_i)e^{-nP(C_i)} \leq r\max P(C_i)e^{-nP(C_i)} \leq \frac{r}{ne}$.□

- **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 $2\sqrt{d}(ne)^{-\frac{1}{d+1}} \leq 2\sqrt{d}n^{-\frac{1}{d+1}}$. □

- **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. □

- **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 θ is estimable if $\exists$ an unbiased estimator of θ. $c^T\beta$ 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 $c^T\hat{\beta} = l^TX\hat{\beta} = l^TX(X^TX)^-X^TY = l^TP_XY$ 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 β, $c = X^T\int \frac{y-X\beta}{(2\pi\sigma^2)^{\frac{n}{2}}\sigma^2}T(X,y)\exp\{-\frac{1}{2\sigma^2}||y-X\beta||^2\}dy$. $\square$

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, β 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 σ^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 $\beta_2\neq\beta_2^*$. β_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, $\mathrm{RSS}_{H_0} - \mathrm{RSS}\sim\sigma^2\chi_s^2(\lambda)$ 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 β_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 $\mathcal{C}(\tilde{X}^T) = \mathcal{C}((A,H)^{-1}X^T) \supset \mathcal{C}((A,H)^{-1}H) = \mathcal{C}(\begin{pmatrix}0\\ I_s\end{pmatrix})$, θ_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 γ_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 σ^2 is known; $T = \frac{\hat{\gamma}-\gamma_0}{\hat{\sigma}/\sqrt{D}} \sim t_{n-r}$ if σ^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 σ^2 is unknown, replacing σ^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.

- Multiple testing: Simultaneous confidence intervals of level $1-\alpha$.

- Bonferroni: Replace α by α/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: Consider $Y = X\beta + \epsilon, \epsilon \sim \mathcal{N}(0, \sigma^2I_n), \text{rank}(X) = r$ 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. $\hat{\gamma} = H\hat{\beta} \sim \mathcal{N}_s(\gamma, \sigma^2D)$ where $D = H^T(X^TX)^-H$, $\hat{\sigma}^2 = \frac{\text{RSS}}{n-r} \sim \sigma^2\chi^2_{n-r}$. For any fixed $u \in \mathbb{R}^s$, an $(1-\alpha)$ CI for $u^T\gamma : u^T\hat{\gamma} \pm t_{n-r,\frac{\alpha}{2}}\hat{\sigma}\sqrt{u^TDu}$. Now allow $u \in \mathbb{R}^s$ to vary arbitrarily. Since $\sup_{u\neq0}\frac{|u^T\hat{\gamma}-u^T\gamma|^2}{u^TDu} \overset{v=D^{\frac{1}{2}}u}{=} \sup_{v\neq0}\frac{|v^TD^{-\frac{1}{2}}(\hat{\gamma}-\gamma)|^2}{v^Tv} \overset{\text{Cauchy-Schwarz}}{=} (\hat{\gamma}-\gamma)D^{-1}(\hat{\gamma}-\gamma), P(\sup_{u\neq0}\frac{|u^T\hat{\gamma}-u^T\gamma|^2}{s\hat{\sigma}^2u^TDu} \le F_{s,n-r,\alpha}) = 1-\alpha$. Simultaneous CIs for $u^T\gamma, \forall u \in \mathbb{R}^s : u^T\hat{\gamma} \pm \hat{\sigma}\sqrt{sF_{s,n-r,\alpha}u^TDu}$. (Bonfernnoi: $t_{n-r,\alpha/(2m)}$)

- Tukey's method: Consider $y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \epsilon_{ij} \sim \mathcal{N}(0,\sigma^2)$ i.i.d., $j = 1, \cdots, m, i = 1, \cdots, k$ and test for $\alpha_i - \alpha_{i'}, \forall i, i' = 1, \cdots, k$. If $Z_1, \cdots, Z_n \sim \mathcal{N}(0,1), R^2 \sim \chi^2_v$, then $\frac{Z_{(n)}-Z_{(1)}}{\sqrt{R^2/v}} \sim q_{n,v}$ (studentized range distribution). Thus $\frac{\sqrt{m}}{\hat{\sigma}}\max_{i,i'}\{\bar{y_i} - \bar{y_{i'}} - (\alpha_i - \alpha_{i'})\} = \frac{\left\{\max_i\frac{\sqrt{m}(\bar{y_i}-\mu-\alpha_i)}{\sigma} - \min_i\frac{\sqrt{m}(\bar{y_i}-\mu-\alpha_i)}{\sigma}\right\}}{\sqrt{\frac{\text{RSS}/\sigma^2}{n-k}}} \sim q_{k,n-k}$. Simultaneous CIs: $\bar{y_i} - \bar{y_{i'}} \pm \frac{\hat{\sigma}}{\sqrt{m}}q_{k,n-k,\alpha}$. (Bonfernnoi: $t_{n-k,\alpha/[k(k-1)]}$, Scheffé: $\sqrt{kF_{k,n-k,\alpha}}$, Tukey: $q_{k,n-k,\alpha}/\sqrt{2}$ (the best/shortest length))

5 Exponential Families

- One parameter exponential families: $\mathscr{G} = \{g_\eta(y) = e^{\eta y - \psi(\eta)}g_0(y)d\nu(y), \eta \in A, y \in \mathcal{Y}\}$, or $\log g_\theta(x) = A(\theta)B(x) + C(\theta) + D(x)$. η: natural parameter; y: sufficient statistics; $\psi(\eta)$: normalizing function s.t. $\frac{\int e^{\eta y}g_0(y)d\nu(y)}{e^{\psi(\eta)}} = 1$; A: natural parameter space s.t. $\int e^{\eta y}g_0(y)d\nu(y) < \infty$. $e^{\eta y-\psi(\eta)}$: exponential tilting, a method of generating an additive distribution family.

- Mean and variance: $e^{\psi(\eta)} = \int_Y e^{\eta y}g_0(y)d\nu(y)$, differnetiating w.r.t. y, $\psi'(\eta)e^{\psi(\eta)} = \int_Y ye^{\eta y}g_0(y)d\nu(y), [\psi''(\eta) + \psi'(\eta)^2]e^{\psi(\eta)} = \int_Y y^2e^{\eta y}g_0(y)d\nu(y) \Rightarrow \psi'(\eta) = \mathbb{E}_\eta Y = \mu_\eta, \psi''(\eta) = \mathbb{E}_\eta Y^2 - \mu_\eta^2 = \text{Var}_\eta(Y) = V_\eta$.

- Cumulants: Let $\kappa_j, j = 1, 2, \cdots$ satisfy $\psi(\eta) - \psi(\eta_0) = \kappa_1(\eta - \eta_0) + \frac{\kappa_2}{2}(\eta - \eta_0)^2 + \frac{\kappa_3}{3!}(\eta - \eta_0)^3 + \cdots$. $\psi'''(\eta_0) = \kappa_3 = \mathbb{E}_0(Y - \mu_0)^3, \psi''''(\eta_0) = \kappa_4 = \mathbb{E}_0(Y - \mu_0)^4 - 3\kappa_2^2$. They correspond to central/noncentral moments. Skewness(偏度): $\gamma = \frac{\kappa_3}{\kappa_2^{3/2}} = \frac{\mathbb{E}(Y-\mathbb{E}Y)^3}{(\mathrm{Var}(Y))^{3/2}}$. Kurtosis(峰度): $\delta = \frac{\kappa_4}{\kappa_2^2} = \frac{\mathbb{E}(Y-\mathbb{E}Y)^4}{(\mathrm{Var}(Y))^2} - 3$.

- If $y \sim g_\eta(\cdot)$ in an exponential family, then $y \sim [\psi', \psi''^{1/2}, \psi'''/\psi''^{3/2}, \psi''''/\psi''^2]$(expectation, SD, skewness, kurtosis). e.g. Poisson: $\psi = e^\eta = \mu, \phi' = \cdots = \phi'''' = \mu, y \sim [\mu, \sqrt{\mu}, 1/\sqrt{\mu}, 1/\mu]$.

- **Theorem** 5.1 $P(Y \leq \mathrm{median}(Y)) \approx 0.5 + \frac{1}{6\sqrt{2\pi}}\mathrm{skewness}(Y)$.

- **Lemma** 5.1 $Y = [y_0, y_1]$, then $\mathbb{E}_\eta[-l_0'(y)] = \eta - (g_\eta(y_1) - g_\eta(y_0))$ where $l_0(y) = \log g_0(y)$ and $l_0'(y) = \frac{dl_0(y)}{dy}$.

  Proof Integration by parts. □

- MLEs in exponential family: $Y_i \sim g_\eta$ i.i.d. for $i = 1, \cdots, n$. $g_\eta^{(n)}(y) = e^{n(\eta\bar{y} - \psi(\eta))}\prod_{i=1}^n g_0(y_i), \eta^{(n)} = n\eta, \psi^{(n)}(y) = n\psi(\eta^{(n)}/n)$. log-likelihood: $l_\eta(y) = \log g_\eta^{(n)}(y) = n(\eta\bar{y} - \psi(\eta)) + C$, score: $l_\eta'(y) = n(\bar{y} - \mu_\eta)$, score equation: $l_{\hat{\eta}}'(y) = 0 \Rightarrow \mu_{\hat{\eta}} = \bar{y}$. Since $\frac{d\mu}{d\eta} = \psi''(\eta) = V_\eta > 0$, we can solve $\hat{\eta}$ by $\hat{\eta} = \psi'^{-1}(\hat{\mu})$. e.g. (1) Poisson: $\hat{\eta} = \log(\bar{y})$; (2) Binomial: $\hat{\eta} = \log(\frac{\bar{y}}{1-\bar{y}})$.

- Fisher information: $I_\eta^{(n)} = nI_\eta = nV_\eta, I_\mu^{(n)} = nI_\mu = \frac{n}{V_\eta}$. C-R lower bound: $\xi = h(\eta)$, any unbiased estimator $\bar{\xi}$ of ξ, $\mathrm{Var}(\bar{\xi}) \geq \frac{1}{I_\mu^{(n)}(\xi)} = \frac{(h'(\eta))^2}{nV_\eta}$. In particular, $\xi = \mu$, then $\mathrm{Var}(\hat{\mu}) \geq \frac{V_\eta}{n}$.

- Important distributions: (1) Normal: $\mathcal{N}(\eta, 1), \psi(\eta) = \frac{1}{2}\eta^2, g_0(y) = \frac{1}{\sqrt{2\pi}}e^{-y^2/2}$; (2) Binomial: $g_\eta(y) = C_N^y\pi^y(1-\pi)^{N-y} = C_N^y e^{y\log\pi + (N-y)\log(1-\pi)}, y = 0, 1, \cdots, N, \eta = \log\frac{\pi}{1-\pi}, \pi = \frac{1}{1+e^{-\eta}} = \frac{e^\eta}{1+e^\eta}, \psi(\eta) = N\log(1+e^\eta)$; (3) Gamma$(k, \theta)$(shape, scale), $\chi_k^2 = \mathrm{Gamma}(k/2, 2)$; (4) Negative Binomial: $\mathrm{NB}(k, \theta) = \#$ tails until kth head. $g_\eta(y) = C_{y+k-1}^{k-1}(1-\theta)^y\theta^k = C_{y+k-1}^{k-1}e^{y\log(1-\theta) + k\log\theta}, y = 0, 1, 2, \cdots, \theta \in (0, 1), \eta = \log(1-\theta), \psi(\eta) = k\log(1-e^\eta), \mu = k\frac{1-\theta}{\theta}, V = \frac{\mu}{\theta}$ (property: $k \to \infty, \mu$ fixed, $Y \to \mathrm{Poisson}(\mu)$).

- Inverse Gaussian: $W(t)$: Wiener process with drift $1/\mu$. $W(t) = \frac{1}{\mu}t + B(t)$ and $W(t) \sim \mathcal{N}(t/\mu, t), \mathrm{Cov}(W(t), W(t+s)) = t$. Y = 1st passage time to $W(t) = 1$. Density of IG(μ): $g(y) = \frac{1}{\sqrt{2\pi y^3}}\exp\{-\frac{(y-\mu)^2}{2\mu^2 y}\} = \frac{1}{\sqrt{2\pi y^3}}\exp(-\frac{y}{2\mu^2} + \frac{1}{\mu} - \frac{1}{2y})$ with $\eta = -\frac{1}{2\mu^2}, \psi(\eta) = -\sqrt{2\eta}$ belongs to the exponential family.

- Tilted hypergeometric: Consider 2×2 talk (Table 1). Counts $X = (x_1, x_2, x_3, x_4) \sim \mathrm{Multinomial}(N, (\pi_1, \pi_2, \pi_3, \pi_4))$. Test: $H_0 : \theta = \log(\frac{\pi_1/\pi_2}{\pi_3/\pi_4}) = 0$. Under H_0, conditional distribution of x_1 given (r_1, r_2, c_1, c_2) is $g_0(x_1|r_1, r_2, c_1, c_2) = \frac{C_{r_1}^{x_1}C_{r_2}^{c_1-x_1}}{C_N^{c_1}} \sim$ hypergeometric with $\max(0, c_1 - r_2) \leq x_1 \leq \min(c_1, r_1)$. When H_0 is not true, $g_\theta(x_1|r_1, r_2, c_1, c_2) = \frac{g_0(x_1|r_1,r_2,c_1,c_2)e^{\theta x_1}C_N^{c_1}}{C(\theta)}$ belongs to the exponential family with $C(\theta) = \sum_{x_1}C_{r_1}^{x_1}C_{r_2}^{c_1-x_1}e^{\theta x_1}$.

Table 1: 2 × 2 talk

	Yes	No	
Male	x_1	x_2	r_1
Female	x_3	x_4	r_2
	c_1	c_2	N

- Deviance (Kullback-Leibler divergence): Generating Euclidean distance to exponential families, $2\mathrm{KL}(\eta_1, \eta_2) = D(\eta_1, \eta_2) := 2\int\eta_1(y)\log\frac{\eta_1(y)}{\eta_2(y)}d\nu(y) = 2\mathbb{E}_{\eta_1}[(\eta_1 - \eta_2)y - (\psi(\eta_1) - \psi(\eta_2))] = 2[(\eta_1 - \eta_2)\mu_1 - (\psi(\eta_1) - \psi(\eta_2))]$. Multual information: $D(f(x, y), f(x)f(y))/2$. Example: (1) $\mathcal{N}(\mu, 1) : D(\mu_1, \mu_2) = (\mu_1 - \mu_2)^2$; (2) Poisson$(\mu)$: $D(\mu_1, \mu_2) = 2\mu_1[\log(\frac{\mu_1}{\mu_2}) - (1 - \frac{\mu_2}{\mu_1})]$; (3) Binomial$(N, \pi) : D(\pi_1, \pi_2) = 2N[\pi_1\log(\frac{\pi_1}{\pi_2}) + (1 - \pi_1)\log(\frac{1-\pi_1}{1-\pi_2})]$.

- **Theorem** 5.2 (Hoeffding's formula) For $g_\eta(y) = e^{\eta y - \psi(\eta)}g_0(y)$, let $\hat{\eta}$ be the MLE of η and $\hat{\mu}$ be the MLE of μ. Then $g_\eta(y) = g_{\hat{\eta}}(y)e^{-D(\hat{\eta}, \eta)/2}, g_\mu(y) = g_{\hat{\mu}}(y)e^{-D(\hat{\mu}, \mu)/2}$.

Proof $\frac{g_\eta(y)}{g_{\hat\eta}(y)} = e^{(\eta-\hat\eta)y-(\psi(\eta)-\psi(\hat\eta))} \overset{y=\hat\mu}{=} e^{-D(\hat\eta,\eta)/2}$. □

- **Proposition** 5.1 $D(\eta_1,\eta_2) = I_{\eta_1} \times (\eta_2-\eta_1)^2 + O((\eta_2-\eta_1)^3)$.

  Proof $\frac{\partial}{\partial\eta_2}D(\eta_1,\eta_2) = \frac{\partial}{\partial\eta_2}2[(\eta_1-\eta_2)\mu_1-(\psi(\eta_1)-\psi(\eta_2))] = 2(-\mu_1+\mu_2) \Rightarrow \frac{\partial}{\partial\eta_2}D(\eta_1,\eta_2)|_{\eta_2=\eta_1} = 0$. $\frac{\partial^2}{\partial\eta_2^2}D(\eta_1,\eta_2) = 2\frac{\partial\mu_2}{\partial\eta_2} \Rightarrow \frac{\partial^2}{\partial\eta_2^2}D(\eta_1,\eta_2)|_{\eta_2=\eta_1} = 2V_{\eta_1}$. Taylor expansion: $D(\eta_1,\eta_2) = 2V_{\eta_1}\frac{(\eta_2-\eta_1)^2}{2} + O((\eta_2-\eta_1)^3) = I_{\eta_1}(\eta_2-\eta_1)^2 + O((\eta_2-\eta_1)^3)$. □

- Deviance residuals: Exponential family analogue of normal residuals $y-\mu$: $\text{sgn}(y-\mu)\sqrt{D(y,\mu)}$. Let $y_i \sim g_\mu(\cdot)$ i.i.d. for $i=1,\cdots,n$. Define the deviance residual $R = \text{sgn}(\bar y-\mu)\sqrt{nD(\bar y,\mu)} = \text{sgn}(\bar y-\mu)\sqrt{D^{(n)}(\bar y,\mu)}$. The hope is that R will be nearly $\mathcal{N}(0,1)$, at least closer to normal than the more obvious "Pearson residual" $R_p = \frac{\bar y-\mu}{\sqrt{V_\mu/n}}$.

- **Theorem** 5.3 $R \dot\sim \mathcal{N}(-a_n,(1+b_n)^2)$ where $a_n = \frac{\gamma_\mu/6}{\sqrt n}$ and $b_n = \frac{\frac{7}{36}\gamma_\mu^2-\delta_\mu}{n}$ (recall γ_μ,δ_μ is skewness and kurtosis of g_μ). The constants a_n and b_n are called "Bartlett corrections". More precisely, $P(\frac{R+a_n}{1+b_n} > z_\alpha) = \alpha + O(n^{-3/2})$.

  Corollary 5.1 $D^{(n)}(\bar y,\mu) = R^2 \dot\sim (1+\frac{5\gamma_\mu^2-3\delta_\mu}{12n})\chi_1^2$.

- We wish to approximate the density under $g_\mu^{(n)}$ of the sufficient statistic $\hat\mu = \bar y$. Normal approximation: $g_\mu^{(n)}(\hat\mu) = \sqrt{\frac{n}{2\pi V_\mu}}e^{-\frac{n(\hat\mu-\mu)^2}{2V_\mu}}$. Saddlepoint approximation: $g_\mu^{(n)}(\hat\mu) = \sqrt{\frac{n}{2\pi\hat V}}e^{-nD(\hat\mu,\mu)/2}$.

- Lugananni-Rice Formula: Observing $\bar y = \hat\mu$, p-value $\alpha(\mu) = \int_{\hat\mu}^\infty g_\mu^{(n)}(t)d\nu(t) \approx 1-\Phi(R)-\phi(R)(\frac{1}{R}-\frac{1}{Q})+O(n^{-3/2})$ where Φ and ϕ are cdf/pdf of $\mathcal{N}(0,1)$, $R = \text{sgn}(\hat\mu-\mu)\sqrt{nD(\hat\mu,\mu)}$ is the deviance residual, and $Q = \sqrt{n\hat V}(\hat\eta-\eta)$ is the crude form of the Pearson residual based on the canonical parameter.

- Transformation: $\zeta = H(\mu), \hat\zeta = H(\hat\mu)$, $\hat\mu$ the MLE of μ, $H'(\mu) = V_\mu^{\delta-1}, 0 \le \delta \le 1$.

$\delta =$	0	$\frac{1}{3}$	$\frac{1}{2}$	$\frac{2}{3}$	1
$\zeta =$	Canonical parameter η	Normal likelihood	Stabilized variance	Normal density	Expectation parameter μ

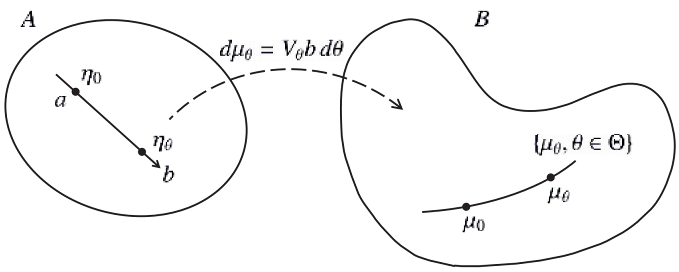
  Example (when $\delta = \frac{1}{2}$): (1) Poisson(μ), $H'(\mu) = \mu^{-1/2}, H(\mu) = 2\sqrt\mu, 2\sqrt y \dot\sim \mathcal{N}(2\sqrt\mu,1)$; (2) Binomial($N,\pi$), $\hat\zeta = 2\sqrt N\sin^{-1}\sqrt{\frac{Np+3/8}{N+3/4}}$.

- Multiparameter exponential families: A p-parameter exponential family $\mathscr{G} = \{g_\eta(y) : \eta \in A \subset \mathbb{R}^p, y \in \mathcal{Y} \subset \mathbb{R}^p\}$ with $g_\eta(y) = e^{\eta^Ty-\psi(\eta)}g_0(y)d\nu(y), \mu = \mathbb{E}_\eta Y = \psi'(\eta), V = \text{Var}_\eta(Y) = \psi''(\eta), d\mu = Vd\eta, d\eta = V^{-1}d\mu$. Assume V will be positive definite for all η in $A = \{\eta : \int_{\mathcal{Y}} e^{\eta^Ty}g_0(y)d\nu < \infty\}$. Let $B = \{\mu = \mathbb{E}_\eta Y, \eta \in A\}$.

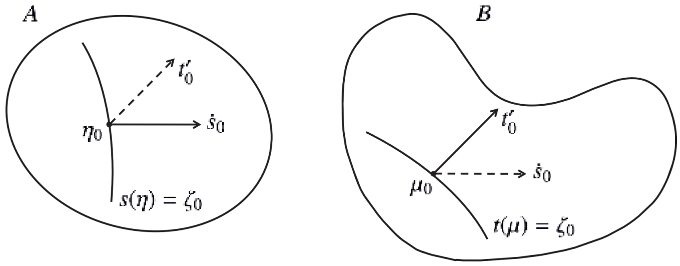
- Facts: (1) A is convex; (2) $B \subset$ convex hull of $\mathcal{Y}$; (3) Angle$(d\eta,d\mu) < \frac{\pi}{2}$ $(d\eta^Td\mu = d\eta^TVd\eta > 0)$.

- Transformation: $\zeta = h(\eta) = H(\mu) \in \mathbb{R}, \eta,\mu \in \mathbb{R}^p, D = \frac{d\eta}{d\mu} = V^{-1}$. Then $H'(\mu) = Dh'(\eta), H''(\mu) = Dh''(\eta)D^T + D_2h'(\eta)$ where $D_2 = (\frac{\partial^2\eta_k}{\partial\mu_i\partial\mu_j})_{i,j,k}$.

- One-parameter subfamilies: $\eta_\theta = a+b\theta, \theta \in \Theta \subset \mathbb{R}, a,b \in \mathbb{R}^p, \mathscr{F} = \{f_\theta(y) = g_{\eta_\theta}(y) = e^{(a+b\theta)^Ty-\psi(a+b\theta)}g_0(y)d\nu, \theta \in \Theta\}$. Still a one-parameter exponential family, natural parameter θ, sufficient statistics $x = b^Ty$. MLE of θ (score equation): $l'_\theta(\bar y) = 0 \Rightarrow b^T(\bar y-\mu_\theta) = 0$.



- Stein's least favorable subfamily: $\zeta = s(\eta) = t(\mu), \zeta_0 = s(\eta_0) = t(\mu_0), s_0' = \frac{\partial s(\eta)}{\partial \eta}|_{\eta_0}, t_0' = \frac{\partial t(\mu)}{\partial \mu}|_{\mu_0}$. Define the LFF: $\eta_\theta = \eta_0 + t_0'\theta, \theta \in$ neighborhood of 0.



- **Theorem** 5.4 The 1-parameter CRLB for estimating ζ in LFF evaluated at $\theta = 0$ is the same as the p-parameter CRLB for estimating ζ in $\mathscr{G}$ at $\eta = \eta_0$, which equals $t_0' V_0 t_0$, where V_0 is the variance evaluated at η_0 or μ_0.

  Remark 5.1 In other words, the reduction to the LFF does not make it any easier to estimate ζ. It can be shown that any choice other than $b = t_0'$ for the family $\eta_\theta = \eta_0 + b\theta$ makes the one-parameter CRLB smaller than the p-parameter CRLB. Stein's construction is useful when some statistical property is easily calculated only in the one-parameter case.

- Examples: (1) $\mathcal{N}(\lambda, \Gamma) : g(x) = \frac{1}{\sqrt{2\pi\Gamma}} \exp(-\frac{x^2}{2\Gamma} + \frac{\lambda}{\Gamma}x - \frac{\lambda^2}{2\Gamma}), \eta = (\lambda/\Gamma, -\frac{1}{2\Gamma})^T, y = (x, x^2)^T, \mu = (\lambda, \lambda^2 + \Gamma)^T$; (2) $\text{Beta}(\alpha, \beta) : g(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)} = \exp\{\alpha \log x + \beta \log(1-x) - \log B(\alpha, \beta)\}\frac{1}{x(1-x)}, \eta = (\alpha, \beta)^T, y = (\log x, \log(1-x))^T$; (3) $\text{Dirichlet}(\alpha_1, \cdots, \alpha_p), g_\alpha(x) = \frac{1}{B(\alpha)} \prod_{i=1}^p x_i^{\alpha_i - 1}, B(\alpha) = \frac{\prod_{i=1}^p \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^p \alpha_i)}, x \in \mathbb{S}^{p-1}$; (4) Graph/Degree model: $Y_{ij} = I(i—j), \pi_{ij} = P(Y_{ij} = 1) = \frac{e^{\theta_i + \theta_j}}{1 + e^{\theta_i + \theta_j}}, \theta_i = \beta^T x_i$ where x_i's are optional predictors. Sufficient statistics is degree of node i. (5) Bradley-Terry model: $\pi_{ij} = \frac{e^{\theta_i}}{e^{\theta_i} + e^{\theta_j}} = \frac{e^{\theta_i - \theta_j}}{1 + e^{\theta_i - \theta_j}}, w_{ij} \sim \text{Binomial}(n_{ij}, \pi_{ij}), g_\theta \propto \exp(\sum_{i,j}(\theta_i - \theta_j)w_{ij}) = \exp(\sum_i \theta_i \sum_j w_{ij} - \sum_j \theta_j \sum_i w_{ij}) = \exp\{\sum_i \theta_i[\#\text{win}(i) - \#\text{lose}(i)]\}$.

- Truncated data: $y \sim g_\eta(y) = e^{\eta^T y - \psi(\eta)} g_0(y)$, observed only if y falls in $\mathcal{Y}_0 \subset \mathcal{Y}$. Conditional density: $g_\eta(y|\mathcal{Y}_0) = e^{\eta^T y - \psi(\eta)} \frac{g_0(y)}{G_\eta(\mathcal{Y}_0)}$, where $G_\eta(\mathcal{Y}_0) = \int_{\mathcal{Y}_0} g_\eta(y)dy$.

- **Lemma** 5.2 Partition $\eta = (\eta_1, \eta_2), y = (y_1, y_2). y_1|y_2 \sim g_{\eta_1}(y_1|y_2) = e^{\eta_1^T y_1 - \psi(\eta_1|\eta_2)} dG_0(y_1|y_2), y_2 \sim g_{\eta_1,\eta_2}(y_2) = e^{\eta_2^T y_2 - \psi_{\eta_1}(\eta_2)} dG_{\eta_1,0}(y_2)$.

  Proof $g_\eta(y_2) = \int_{\mathcal{Y}_1} e^{\eta_1^T y_1 + \eta_2^T y_2 - \psi(\eta)} g_0(y_1|y_2) g_0(y_2) dy_1 = e^{\eta_2^T y_2 - \psi(\eta)} (\int_{\mathcal{Y}_1} e^{\eta_1^T y_1} g_0(y_1|y_2)dy_1) g_0(y_2) \Rightarrow g_\eta(y_1|y_2) = \frac{g_\eta(y)}{g_\eta(y_2)} = \frac{e^{\eta_1^T y_1 + \eta_2^T y_2 - \psi(\eta)} g_0(y)}{e^{\eta_2^T y_2 - \psi(\eta) + \psi(\eta_1|\eta_2)} g_0(y_2)} = e^{\eta_1^T y_1 - \psi(\eta_1|\eta_2)} dG_0(y_1|y_2)$. □

  Remark 5.2 Usually after a transformation $M \in \mathbb{R}^{p \times p}$ nonsingular, $\tilde{\eta} = (M^{-1})^T \eta, \tilde{y} = My$.

- Examples: (1) Fisher's exact test for 2×2 talk (Recall Table 1), $H_0 : \theta = \log(\frac{\pi_1/\pi_2}{\pi_3/\pi_4}) = 0$. The natural parameter is $\eta = (\log \pi_1, \cdots, \log \pi_4)$. Let $(M^{-1})^T = \begin{pmatrix} 1 & -1 & -1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & 1 & 1 & 1 \end{pmatrix}$, so that $M = \frac{1}{4}(M^{-1})^T, \tilde{y} = Mx, \tilde{y}_1 = \frac{1}{4}(x_1 - x_2 - x_3 + x_4) = x_1 - \frac{r_1}{2} - \frac{c_1}{2} + \frac{N}{4}$. (2) Wishart statistics: $x_1, \cdots, x_n \sim \mathcal{N}_d(\lambda, \Gamma)$ independent, $y_1 = \bar{x}, y_2 = \frac{1}{n}\sum_{i=1}^n x_i x_i^T$. Wishart statistics $W = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})^T = y_2 - y_1 y_1^T$. $y_2|y_1$ is in a $\frac{d(d+1)}{2}$-dim exponential family. (3) Poisson trick: $s = (s_1, \cdots, s_L), s_l \sim \text{Poisson}(\mu_L)$ independent $\Rightarrow s|n = \sum_{l=1}^L s_l \sim \text{Multinomial}_L(n, \pi)$ where $\pi_l = \frac{\mu_l}{\sum_j \mu_j}$. Conversely, if $s|n \sim \text{Multinomial}(n, \pi)$ and $n \sim \text{Poisson}(\mu_+)$, then $s_l \sim \text{Poisson}(\mu_+ \pi_l)$ i.i.d.

- Rotational speeds of stars: Bimodal: $f(x) = w\frac{\phi(x/c_1)}{c_1} + (1-w)\frac{\phi(x/c_2)}{c_2}$. Two competing candidates candidates for $\phi(x) : \phi_1(x) = 2xe^{-x^2}, \phi_2(x) = 4x^2 e^{-x^2}\pi^{-1/2}$. We take the bin partitions and set y_l to be the count and π_l be the probability of bin l. $y_l \sim \text{Poisson}(\mu_l), \mu_l = n\pi_l$. Any choice of (w, c_1, c_2) produces estimates of π_l and μ_l.

6 Generalized Linear Models

6.1 Basic Concepts

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

- Three components of GLMs: (1) Random: distribution of Y with $\mathbb{E}Y = \mu$; (2) Systematic: $\eta = \sum_{j=1}^p x_j \beta_j$; (3) Link: $g(\mu) = \eta$.

- Example 1 (Dilution assays): density ρ_0, at the x-th dilation $\rho_x = \rho_0 2^{-x}, x = 0, 1, 2, \cdots$, proportion of infected plates $y_x = \frac{r_x}{m_x}$, $Y = I(\text{infected})$, $\mathbb{E}(Y|x) = P(Y = 1|x) = \pi_x$, # orgamism 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 π_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 Φ is the c.d.f. of $\mathcal{N}(0, 1)$, link function $g = \Phi^{-1}$. (2) Logistic/Logit model: $\pi_x = \text{expit}(\alpha + \beta x) = \frac{1}{1 + e^{-(\alpha + \beta x)}}$, link function $g(\pi_x) = \text{logit}(\pi_x) = \log \frac{\pi_x}{1 - \pi_x}$.

- Random component: Y has a distribution in an exponential family: $f(y; \theta, \phi) = \exp\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\}$ where ϕ is dispersion parameter. Usually $a(\phi) = \phi / w_i$. log-likelihood: $l(\theta; y) = \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)$. $\frac{\partial l}{\partial \theta} = \frac{y - b'(\theta)}{a(\phi)}, \frac{\partial^2 l}{\partial \theta^2} = -\frac{b''(\theta)}{a(\phi)}$. $\mathbb{E}\frac{\partial l}{\partial \theta} = 0, \mathbb{E}(\frac{\partial l}{\partial \theta})^2 = -\mathbb{E}\frac{\partial^2 l}{\partial \theta^2}$, $\mathbb{E}Y = \mu = b'(\theta), \text{Var}(Y) = a(\phi) b''(\theta)$.

- Systematic component: predictors $(x_1, \cdots, x_p), \eta = x^T \beta$.

- Canonical link function: $g = b'^{-1}(\mu)$ so that $\eta = g(\mu) = b'^{-1}(b'(\theta)) = \theta$.

- Goodness of fit: Null model: one parameter, μ common mean. Full model: n parameters, one per oberservation. Idea: Measure discrepancy between an intermediate model and the full model.

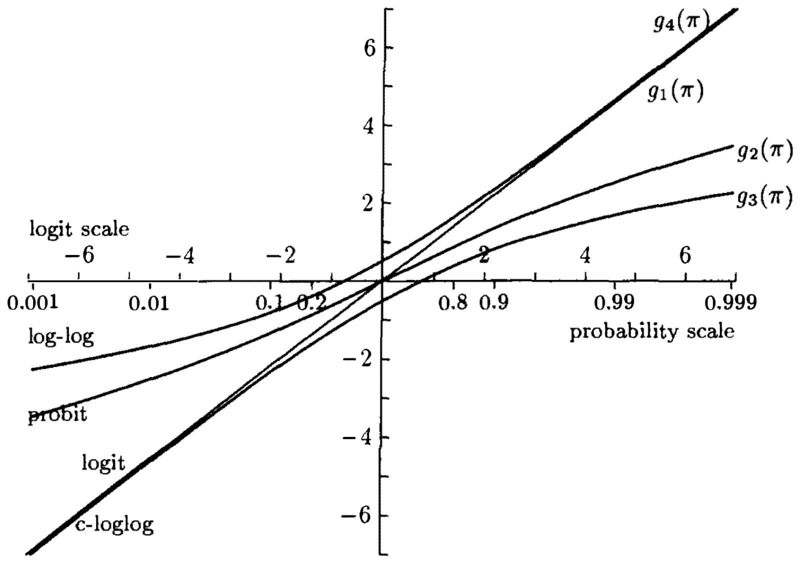
- Assume $l(y, \phi; y), l(\hat{\mu}, \phi; y)$ maximize log-likelihood over β with fixed ϕ, g_1/g_2 is full/current model respectively, $\tilde{\theta}/\hat{\theta} = \theta(y)/\theta(\hat{\mu})$ and $a_i(\phi) = \phi/w_i$. $2\mathbb{E}_{P_n} \log \frac{l(y, \phi; y)}{l(\hat{\mu}, \phi; y)} = 2\sum_{i=1}^n \frac{w_i}{\phi}[(\tilde{\theta}_i - \hat{\theta}_i) y_i - b(\tilde{\theta}_i) + b(\hat{\theta}_i)] := \frac{D(y, \hat{\mu})}{\phi}$. Under suitable regularity conditions, if the fitted model is correct, $D(y, \hat{\mu})/\phi \dot{\sim} \chi^2_{n-p}$ where p is the dimension of β.

- Pearson's χ^2-statistic: $\chi^2 = \sum_{i=1}^n \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)/w_i}$ where $V(\mu) = b''(b'^{-1}(\mu))$. Under suitable regularity conditions, if the model is correct, $\chi^2/\phi \dot{\sim} \chi^2_{n-p}$.

- Residuals: (1) Deviance residual: $r_D = \text{sgn}(y - \hat{\mu})\sqrt{d_i}$ where $d_i = 2w_i[(\tilde{\theta}_i - \hat{\theta}_i) y_i - b(\tilde{\theta}_i) + b(\hat{\theta}_i)]$; (2) Pearson residual: $r_p = \frac{y - \hat{\mu}}{\sqrt{V(\hat{\mu})/w_i}}$; (3) Anscombe residual: $\delta = \frac{2}{3}, H'(\mu) = V_\mu^{-\frac{1}{3}}, A = \int \frac{d\mu}{V^{1/3}(\mu)}$. For Poisson distribution, $A = \frac{3}{2}\mu^{2/3}$, and we must scale by dividing by the SD of $A(Y)$, i.e. $A'(\mu)\sqrt{V(\mu)} \Rightarrow r_A = \frac{\frac{3}{2}(y^{2/3} - \mu^{2/3})}{\mu^{1/6}}$.

- Algorithms for fitting GLMs: $l(\beta)$ log-likelihood, $u(\beta) = \frac{\partial}{\partial \beta} l(\beta), H(\beta) = \frac{\partial^2}{\partial \beta \partial \beta^T} l(\beta)$. The MLE of $\hat{\beta}$ solves the estimating equation. $0 = u(\hat{\beta}) \approx u(\beta^{(0)}) + H(\beta^{(0)})(\hat{\beta} - \beta^{(0)})$ giving the update $\beta^{(t+1)} = \beta^{(t)} - H(\beta^{(t)})^{-1} u(\beta^{(t)})$. Fisher scoring: $\beta^{(t+1)} = \beta^{(t)} + I(\beta^{(t)})^{-1} u(\beta^{(t)})$ (since $I(\beta) = -\mathbb{E}H(\beta)$). In a GLM, $l = \sum_{i=1}^n l_i, l_i = \frac{y_i \theta_i - b(\theta_i)}{a_i(\phi)} + c_i(y_i, \phi)$, and $u_{ir} = \frac{\partial l_i}{\partial \beta_r} = \frac{\partial l_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_r} = \frac{y_i - \mu_i}{a_i(\phi)}\frac{1}{V(\mu_i)}\frac{1}{g'(\mu_i)} x_{ir} = \frac{(y_i - \mu_i) x_{ir}}{a_i(\phi) V(\mu_i) g'(\mu_i)} = (y - \mu)^T W \frac{d\eta}{d\mu} x_{(r)}$ where $W = \text{diag}(\frac{1}{a_i(\phi) V(\mu_i) g'(\mu_i)^2})$. Since $\text{Cov}(u_r, u_s) = \sum_{i=1}^n \frac{\text{Var}(y_i) x_{ir} x_{is}}{a_i(\phi)^2 V(\mu_i)^2 g'(\mu_i)^2} = \sum_{i=1}^n \frac{x_{ir} x_{is}}{a_i(\phi) V(\mu_i) g'(\mu_i)^2} \Rightarrow I(\beta) = \text{Var}(u(\beta)) = X^T W X, u(\beta) = X^T W \frac{d\eta}{d\mu}(y - \mu)$ where $X = (x_{ir})_{n \times p}$. $H(\beta) = -X^T W X + X^T\{\frac{\partial}{\partial \beta^T}(W \frac{d\eta}{d\mu})\}(y - \mu)$.

- Under what conditions $-H(\beta) = I(\beta)$? Take canonical link $\eta_i = b^{-1}(\mu_i) = \theta_i$, $V(\mu_i) = b''(\theta_i) = \frac{\partial \mu_i}{\partial \theta_i} = \frac{\partial \mu_i}{\partial \eta_i}, w_{ii} = \frac{1}{a_i(\phi)V(\mu_i)g'(\mu_i)^2} = \frac{1}{a_i(\phi)\frac{\partial \eta_i}{\partial \mu_i}} \Rightarrow W\frac{d\eta}{d\mu} = \mathrm{diag}(\frac{1}{a_i(\phi)}) \Rightarrow \frac{\partial}{\partial \beta^T}(W\frac{d\eta}{d\mu}) = 0$.
- Substituting back, $\beta^{(t+1)} = \beta^{(t)} + (X^TW^{(t)}X)^{-1}X^TW^{(t)}\frac{d\eta}{d\mu}(y-\mu) = (X^TW^{(t)}X)^{-1}X^TW^{(t)}[X\beta^{(t)} + \frac{d\eta}{d\mu}(y-\mu)] = (X^TW^{(t)}X)^{-1}X^TW^{(t)}[\eta^{(t)} + \frac{d\eta}{d\mu}|_{\mu^{(t)}}(y-\mu^{(t)})]$ (iteratively reweighted least squares).
- Inference about β: $I(\beta)^{\frac{1}{2}}(\hat{\beta}-\beta) \Rightarrow \mathcal{N}_p(0, I), \widehat{\mathrm{Var}}(\hat{\beta}) = (X^TW(\hat{\beta})X)^{-1}, h(\hat{\beta}) \stackrel{.}{\sim} \mathcal{N}(h(\beta), h'(\beta)^TI(\beta)^{-1}h'(\beta)), \hat{\eta} = x^T\hat{\beta} \stackrel{.}{\sim} \mathcal{N}(x^T\beta, x^TI(\beta)^{-1}x)$.
- CI for x that gives rise to a specified mean response μ_0: $\{x : \frac{x^T\hat{\beta}-g(\mu_0)}{\sqrt{x^TI(\hat{\beta})^{-1}x}} < z_{\alpha/2}\}$ (Fieller's method).
- Binary responses: $g(\pi_i) = \eta_i = x_i^T\beta, g : (0,1) \to \mathbb{R}$, link functions: $g_1 = \log(\frac{\pi}{1-\pi})$, $g_2 = \Phi^{-1}(\pi)$, $g_3 = \log(-\log(1-\pi))$ (complementary log-log), $g_4 = \log(-\log \pi)$ (log-log). These g_i's are from the inverse of the cdfs: $f_1 = \frac{e^x}{(1+e^x)^2}$ (logistic), $f_3 = e^{x-e^x}$, i.e. $\log X, X \sim \mathrm{Exp}(1)$, $f_4 = e^{-x+e^x}$, i.e. $-\log X, X \sim \mathrm{Exp}(1)$ (Gumbel).



- Application: Many epidemiological studies have the goal of comparing distinct groups, e.g., assessing risk factors for some disease. Denote D = disease status, X = exposure status.

	$\overline{D}$	D	
$\overline{X}$	π_{00}	π_{01}	$\pi_{0\cdot}$
X	π_{10}	π_{11}	$\pi_{1\cdot}$
	$\pi_{\cdot 0}$	$\pi_{\cdot 1}$	1

Sampling probabilities: $P(D|x) = \frac{e^{\alpha+x^T\beta}}{1+e^{\alpha+x^T\beta}}, \pi_0 = P(Z=1|D), \pi_1 = P(Z=1|\overline{D})$ where Z is indicator of being sampled. This is because $|D|$ may be much smaller than $|\overline{D}|$ and we need more data on D (i.e. $\pi_0 >> \pi_1$). Then $P(D|Z=1,x) = \frac{P(Z=1|D,x)P(D|x)}{P(Z=1|D,x)P(D|x)+P(Z=1|\overline{D},x)P(\overline{D}|x)} = \frac{\pi_0 e^{\alpha+x^T\beta}}{\pi_0 e^{\alpha+x^T\beta}+\pi_1} = \frac{e^{\alpha+x^T\beta+\log(\pi_0/\pi_1)}}{1+e^{\alpha+x^T\beta+\log(\pi_0/\pi_1)}} := \frac{e^{\alpha^*+x^T\beta}}{1+e^{\alpha^*+x^T\beta}}$ by Bayes formula. Thus, the "biased" random sampling of D and $\overline{D}$ does not impact the value of β, and only translates α to $\alpha + \log(\pi_0/\pi_1)$. We can conduct logistic regression on the new dataset.

6.2 Binomial Regression

- $Y_i \sim \mathrm{Binomial}(m_i, \pi_i), i = 1, , \cdots, n$. For simplicity, $m_i = m, \forall i$. The log-likelihood $l(\pi; y) = \sum_{i=1}^n [y_i \log \frac{\pi_i}{1-\pi_i} + m\log(1-\pi_i)] + C(y)$. Under logistic link, $\log\frac{\pi_i}{1-\pi_i} = x_i^T\beta$, or $\pi_i = \frac{e^{x_i^T\beta}}{1+e^{x_i^T\beta}}$, so that $l(\beta; y) = \sum_{i=1}^n [y_i x_i^T\beta - m\log(1+e^{x_i^T\beta})]$. Exponential family has the form $l(\theta; y) = \sum_{i=1}^n [\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)]$, so $\eta_i = \theta_i = x_i^T\beta, b(\theta_i) =$

$m\log(1+e^{x_i^T\beta}), a_i(\phi) \equiv 1$. General likelihood equation $u(\beta) = X^T W \frac{d\eta}{d\mu}(y-\mu) = 0$ where $W\frac{d\eta}{d\mu} = \text{diag}(\frac{1}{a_i(\phi)})$ under canonical link. Now $a_i(\phi) \equiv 1$, so $u(\beta) = X^T(y-\mu) = 0$. The weight matrix $W = \frac{d\mu}{d\eta} = m\frac{d\pi}{d\eta} = \text{diag}\{m\pi_i(1-\pi_i)\}$. The working response $z_i = \eta_i + \frac{d\eta_i}{d\mu_i}(y_i - \mu_i) = \eta_i + \frac{y_i - m_i\pi_i}{m_i}\frac{d\eta_i}{d\pi_i} = \eta_i + \frac{y_i - m_i\pi_i}{m_i\pi_i(1-\pi_i)}$. Solve $X^T W X\hat{\beta} = X^T W Z$.

- **Theorem** 6.1 (Wedderburn, 1976) If the link function is log concave and $0 < y_i < m_i, \forall i$, then $\hat{\beta}$ is finite and the log-likelihood has a unique maximum at $\hat{\beta}$.
- **Theorem** 6.2 (Shao, Ex 4.117) For logistic regression, if $\sum_{i=1}^n x_i x_i^T$ is positive definite, $\forall n \geq n_0$, then the log-likelihood equation has at most one solution when $n \geq n_0$ and a solution exists with probability $\to 1$.
- Deviance: The fitted log-likelihood $l(\hat{\pi}; y) = \sum_{i=1}^n [y_i \log(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}) + m_i \log(1-\hat{\pi}_i)] = \sum_{i=1}^n [y_i \log \hat{\pi}_i + (m_i - y_i)\log(1-\hat{\pi}_i)]$, maximum achievable log-likelihood $l(\tilde{\pi}_i; y)$ where $\tilde{\pi}_i = \frac{y_i}{m_i}$, $D(y, \hat{\pi}) = 2l(\tilde{\pi}; y) - 2l(\hat{\pi}; y) = 2\sum_{i=1}^m [y_i \log(\frac{y_i}{\hat{\mu}_i}) + (m_i - y_i)\log(\frac{m_i - y_i}{m_i - \hat{\mu}_i})]$. Asymptotic properties: $D(y, \hat{\pi}) \dot{\sim} \chi^2_{n-p}$ (assumptions: no overdispersion; $m_i \to \infty$ with n fixed). Note that if $n \to \infty$ while m_i fixed, D is not independent of $\hat{\pi}$ and large $D \not\Rightarrow$ poor fit.
- Extrapolation: predict x_0 corresponding to π_0. Using Fieller's method, $|\frac{\hat{\beta}_0 + \hat{\beta}_1 x_0 - g(\pi_0)}{V(x_0)}| \leq Z_{\alpha/2}$ where $V(x_0)^2 = \text{Var}(\hat{\beta}_0) + 2x_0\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) + x_0^2\text{Var}(\hat{\beta}_1)$.
- Overdispersion: "nominal" variance: $m\pi(1-\pi)$. $\text{Var}(y) > / < m\pi(1-\pi)$: over/under dispersion. Mechanism: clustering is the population. Assume m subjects from m/k clusters, each of size k. $Z_i \sim \text{Binomial}(k, \pi_i)$ and $Y = Z_1 + \cdots + Z_{m/k}$. If $\mathbb{E}\pi_i = \pi$ and $\text{Var}(\pi_i) = \tau^2\pi(1-\pi)$, then $\mathbb{E}Y = \mathbb{E}(\mathbb{E}(Y|\pi)) = \mathbb{E}[k(\pi_1 + \cdots + \pi_{m/k})] = m\pi, \text{Var}(Y) = \mathbb{E}[\text{Var}(Y|\pi)] + \text{Var}[\mathbb{E}(Y|\pi)] = m\pi(1-\pi)(1-\tau^2) + m\tau^2\pi(1-\pi) = m\pi(1-\pi)[1+(k-1)\pi^2]$. Since $0 \leq \tau^2 \leq 1$ ($\text{Var}(\pi_i) = \mathbb{E}\pi_i^2 - \pi^2 \leq \mathbb{E}\pi_i - \pi^2 = \pi(1-\pi)$), $1 \leq \sigma^2 := 1 + (k-1)\tau^2 \leq k \leq m$.
- Estimation of σ^2 wth overdispersion: Case 1 (with replication): For the same x-value, observe $(y_1, m_1), \cdots, (y_r, m_r)$, $\tilde{\pi} = \frac{\sum_{i=1}^r y_i}{\sum_{i=1}^r m_i}$, $\mathbb{E}[\sum_{j=1}^r \frac{(y_j - m_j\tilde{\pi})^2}{m_j}] = (r-1)\sigma^2\pi(1-\pi), \hat{\sigma}^2 = \frac{1}{r-1}\sum_{j=1}^r \frac{(y_j - m_j\tilde{\pi})^2}{m_j\tilde{\pi}(1-\tilde{\pi})}$ approximately unbiased for σ^2. Case 2 (without replication): Using the fitted $\hat{\pi}_i$, $\hat{\sigma}^2 = \frac{1}{n-p}\sum_{i=1}^n \frac{(y_i - m_i\hat{\pi}_i)^2}{m_i\hat{\pi}_i(1-\hat{\pi}_i)} \dot{\sim} \chi^2_{n-p}, \text{Var}(\hat{\beta}) \approx \hat{\sigma}^2(X^T W X)^{-1}$.

6.3 Poisson Regression

- $Y \sim$ counts of events that occur over a period of time or a region at a constant rate. log-link: $\log \mu_i = \eta_i = x_i^T\beta$. Nominal variance $\text{Var}(y_i) = \mu_i$. More generally, let $\text{Var}(y_i) = \sigma^2\mu_i$. Over/Under dispersion: $\sigma^2 > / < 1$.
- Mechanism for overdispersion: clustered Poisson process. $Y = Z_1 + \cdots + Z_N, Z_i$ i.i.d., $N \sim$ Poisson independent of Z_i. $\mathbb{E}Y = \mathbb{E}N\mathbb{E}Z, \text{Var}(Y) = \mathbb{E}[\text{Var}(Y|N)] + \text{Var}(\mathbb{E}(Y|N)) = \mathbb{E}N\text{Var}(Z) + \text{Var}(N)(\mathbb{E}Z)^2 = \mathbb{E}N\mathbb{E}Z^2 (> \mathbb{E}Y$ if $\mathbb{E}Z^2 > \mathbb{E}Z)$. Estimation of σ^2: $\hat{\sigma}^2 = \frac{1}{n-p}\sum_{i=1}^n \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}$.

6.4 Gamma Regression

- Motivation: $\text{Var}(Y) = $ Const – linear; $\text{Var}(Y) \propto \mathbb{E}Y$ – Poisson; $\text{Var}(Y) \propto (\mathbb{E}Y)^2$ – Gamma. $\sigma := \frac{\sqrt{\text{Var}(Y)}}{\mathbb{E}Y} = $ const: coefficient of variance.
- Mechanism: (1) Multiplicative error: $Y = \mu(1+\epsilon), \mathbb{E}\epsilon = 0, \text{Var}(\epsilon) = \sigma^2, \mathbb{E}Y = \mu, \text{Var}(Y) = \mu^2\sigma^2$. (2) log-transformed additive: $\log Y = \mu + \epsilon, \mathbb{E}\epsilon = 0, \text{Var}(\epsilon) = \sigma^2, Y = e^\mu e^\epsilon, \mathbb{E}Y = e^\mu\mathbb{E}(e^\epsilon), \text{Var}(Y) = e^{2\mu}\text{Var}(e^\epsilon), \frac{\text{Var}(Y)}{(\mathbb{E}Y)^2} = \frac{\text{Var}(e^\epsilon)}{(\mathbb{E}e^\epsilon)^2} \approx \frac{\text{Var}(1+\epsilon)}{(1+\mathbb{E}\epsilon)^2} = \text{Var}(\epsilon) = \sigma^2$.
- Parameterization of gamma: $\text{Gamma}(k, \theta)$ pdf $\frac{1}{\Gamma(k)\theta^k}y^{k-1}e^{-y/\theta}dyI(y > 0), \mathbb{E}Y = k\theta, \text{Var}(Y) = k\theta^2, \sigma^2 = \frac{1}{k}$; $\text{Gamma}(\mu, \nu)$ pdf $\frac{1}{\Gamma(\nu)}(\frac{\nu y}{\mu})^\nu e^{-\nu y/\mu}d(\log y)I(y > 0)$.
- Choice of link function: (1) Canonical link $g(\mu) = \frac{1}{\mu}$. Example: Plants density experiments: x density, yield per plant $\propto \frac{1}{\beta_0 x + \beta_1}$, yield per unit area $\propto \frac{x}{\beta_0 x + \beta_1}, \mu = \eta^{-1} = \frac{x}{\beta_0 + \beta_1}$ or $\eta = \beta_0 + \frac{\beta_1}{x}$. (2) log link: $\eta = \log\mu = x^T\beta$. (3) identity link: $\eta = \mu = x^T\beta$.

- Estimation of σ^2: $\nu = \frac{1}{\sigma^2}, D(y, \hat{\mu}) = 2n[\log \hat{\nu} - \frac{\Gamma'(\hat{\nu})}{\Gamma(\hat{\nu})}]$.
- Example (Rainfall data): Daily rainfall skewed to the right with a spike around 0. Two stages: (1) wet/dry day: Markove chain and logistic; (2) rainfall on wet days: gamma/log-normal. Stage 1: $\pi_0(t) = P(\text{day } t \text{ is wet}|\text{day } t-1 \text{ is wet}), \pi_1(t) = P(\text{day } t \text{ is wet}|\text{day } t-1 \text{ is dry})$. logistic model: $\text{logit}(\pi_i(t)) = \alpha_i + \alpha_{i,1}\sin(\frac{2\pi t}{365}) + \beta_{i,1}\cos(\frac{2\pi t}{365})$. Stage 2: $\log(\mu(t)) = \text{const} + $ harmonic terms where $\mu(t)$ is mean rainfall on day $t|$wet day.

6.5 Categorical Data and Multinomial Regression

- Types of measurement scales: $\begin{cases} \text{nominal: exchangeable} \\ \text{ordinal: ordered but no measure of distance} \\ \text{interval: numerical scores} \\ \text{cardinal: counts} \end{cases}$
- Ordinal: Response probabilities: $\pi_1, \cdots, \pi_k$; cumulative probabilities: $\gamma_j = \sum_{i=1}^{j} \pi_i, j = 1, 2, \cdots, k-1, \gamma_k \equiv 1$. Principle: Inferences should not essentially change by combinining adjacent categories. Proportional odds model: $\log(\frac{\gamma_j(x)}{1-\gamma_j(x)}) = \theta_j - \beta^T x, j = 1, 2, \cdots, k-1$ (parallel regressions), $\frac{\text{odds}(Y \leq j|x_1)}{\text{odds}(Y \leq j|x_2)} = e^{-\beta^T(x_1 - x_2)}$. Why "$-$"? Latent variable interpretation: $Z \sim \text{logistic}(\beta^T x, 1), Z - \beta^T x \sim \text{logistic}(0, 1)$. Let $Y = j$ whenever $\theta_{j-1} < Z \leq \theta_j$. Then $P(Y \leq j) = P(Z \leq \theta_j) = P(Z - \beta^T x \leq \theta_j - \beta^T x) = \frac{e^{\theta_j - \beta^T x}}{1 + e^{\theta_j - \beta^T x}}$. Extensions: (1) Nonparallel: $\text{logit}(\gamma_j(x)) = \theta_j - \beta_j^T x$; (2) Scale modeling: $\text{logit}(\gamma_j(x)) = \frac{\theta_j - \beta^T x}{e^{\tau^T x}}$.
- Interval: Features: (1) The categories are of interest in themselves and are not chosen arbitrarily; (2) It does not normally make sense to form a new category by amalgamating adjacent categories; (3) difference between scores s_j is a measure of distance. Modeling strategies: (1) Extend proportional odds model: $\theta_j = \xi_0 + \xi_1(\frac{s_j + s_{j+1}}{2})$ and replace $\beta^T x$ by $\beta^T x + \xi^T x(c_j - \bar{c}), c_j = \frac{s_j + s_{j+1}}{2}$ or $\text{logit}(\frac{s_j + s_{j+1}}{2})$; (2) Model log-probabilities: $\eta_{ij} = \log \pi_j(x_i), \pi_j(x_i) = \frac{e^{\eta_{ij}}}{\sum_j e^{\eta_{ij}}}, \eta_j(x_i) = \eta_j + \alpha_i$ or $\eta_j(x_i) = \eta_j + (\beta^T x_i)s_j + \alpha_i$, $\frac{\pi_j}{\pi_{j'}} \nearrow$ by a factor $e^{s_j - s_{j'}}$ with a unit change in $\beta^T x$; (3) Model the scores $\mathbb{E}(S|x_i) = \sum_{j=1}^{k} \pi_j(x_i)s_j = \beta^T x_i$.
- Nominal: model π_j or η_j: e.g. $\eta_j(x_i) = \eta_j(x_0) + \beta_j^T(x_i - x_0) + \alpha_i$.
- Multinomial: Data $(y_1, \cdots, y_n), y_i = (y_{i1}, \cdots, y_{ik}), \sum_j y_{ij} = m_i$ fixed, parameters $\pi_i, i = 1, \cdots, n, \pi_i = (\pi_{i1}, \cdots, \pi_{ik})$ s.t. $\sum_j \pi_{ij} = 1$. Log-likelihood: $l(\pi, y) = \sum_{i,j} y_{ij} \log \pi_{ij}$. Use the method of Lagrange multiplies $\Rightarrow \mathscr{L}_\lambda(\pi, y) = \sum_{i,j} y_{ij} \log \pi_{ij} - \sum_i \lambda_i(\sum_j \pi_{ij} - 1), \frac{\partial \mathscr{L}}{\partial \pi_{ij}} = \frac{y_{ij}}{\pi_{ij}} - \lambda_i = 0 \Rightarrow m_i = \lambda_i \Rightarrow \pi_{ij} = \frac{y_{ij}}{m_i}$. Let $\Sigma_i = m\{\text{diag}(\pi_i) - \pi_i \pi_i^T\}$ (rank $k-1$) and $\Sigma_i^- = \text{diag}(\frac{1}{m\pi_{ij}})$. In matrix form, $\frac{\partial \mathscr{L}}{\partial \pi} = M\Sigma^-(y - \mu) = 0$ where $\Sigma = \text{diag}(\Sigma_1, \cdots, \Sigma_n)$ of rank $n(k-1)$ and $M = \text{diag}(\underbrace{m_1, \cdots, m_1}_{k \text{ times}}, \cdots, \underbrace{m_n, \cdots, m_n}_{k \text{ times}})$. Full score equation w.r.t. β_r: $\frac{\partial l}{\partial \pi_{ij}} \frac{\partial \pi_{ij}}{\partial \beta_r} = 0$. GLMs: (1) log-linear: $\log \pi_{ij} = x_{ij}^{*T} \beta^*$; (2) prop odds: $\text{logit}(\gamma_{ij}) = x_{ij}^{*T} \beta^*$ (γ_{ij}: cumulative probability). Overdispersion: $\mathbb{E}Y = m\pi, \text{Var}(Y) = \sigma^2 \Sigma, \hat{\sigma}^2 = \frac{X^2}{\text{residual d.f.} = n(k-1) - p}$ where X^2 is Pearson's statistic.

6.6 Quasi-Likelihood Estimation

- For a GLM, the inference depends on the assumed distribution for y_i only through the mean μ_i and the variance function $V(\cdot)$. In addition, a GLM specifies the independence of observations, which is not indispensable.
- More generally, suppose $\text{Var}(y) = \sigma^2 V(\mu)$. Independent: $V(\mu) = \text{diag}\{V_1(\mu), \cdots, V_n(\mu)\} = \{V_1(\mu_1), \cdots, V_n(\mu_n)\}$. Dependent: $V(\mu)$ nondiagnal. Quasi-score function: $u(\beta) = D^T V^{-1}(y - \mu)/\sigma^2$ where $D = \frac{\partial \mu}{\partial \beta^T} = (\frac{\partial \mu_i}{\partial \beta_r})_{n \times p}$. Facts: $\mathbb{E}u(\beta) = 0, \text{Var}(\beta) = D^T V^{-1} D/\sigma^2, \frac{\partial}{\partial \beta^T} u(\beta) = \frac{1}{\sigma^2}[\frac{\partial}{\partial \beta^T}(D^T V^{-1})(y - \mu) - D^T V^{-1} D], \mathbb{E}[\frac{\partial}{\partial \beta^T} u(\beta)] = -D^T V^{-1} D/\sigma^2$. Quasi-information matrix: $I(\beta) = \text{Var}(u(\beta)) = -\mathbb{E}\frac{\partial}{\partial \beta^T} u(\beta) = D^T V^{-1} D/\sigma^2$.

- Independent: $\frac{\partial u(\beta)}{\partial \beta^T}$ diagnal. Dependent: $\frac{\partial u_r(\beta)}{\partial \beta_s} \neq \frac{\partial u_s(\beta)}{\partial \beta_r}$ in general. Thm: symmetric $\frac{\partial u(\beta)}{\partial \beta^T} \Leftrightarrow$ the line integral (quasi-log-likelihood) $Q(\mu; y, \gamma(s)) = \frac{1}{\sigma^2}\int_{s_0}^{s_1}(y-\gamma)^T V(\gamma)^{-1} d\vec{\gamma}(s)$ along a path $\gamma : [s_0, s_1] \to \mathbb{R}^n$ from $\gamma(s_0) = y$ to $\gamma(s_1) = \mu$ is path-independent. Independent: $Q(\mu, y) = \sum_{i=1}^n Q_i(\mu_i, y_i) = \sum_{i=1}^n \int_{y_i}^{\mu_i} \frac{y_i - t}{\sigma^2 V_i(t)} dt$, quasi-deviance: $D(y, \mu) = 2[Q(y, y) - Q(\mu, y)]\sigma^2 = 2\sum_{i=1}^n \int_{\mu_i}^{y_i} \frac{y_i - t}{V_i(t)} dt$. Dependent: Solve $\sigma^2 u(\beta) = D^T V^{-1}(y - \mu) = 0$ (GEE).

- Fisher-scoring: $\beta^{(t+1)} = \beta^{(t)} + I(\beta^{(t)})^{-1} u(\beta^{(t)}) = \beta^{(t)} + [D(\beta^{(t)}) V(\beta^{(t)})^{-1} D(\beta^{(t)})]^{-1} D(\beta^{(t)})^T V(\beta^{(t)})^{-1}(y - \mu(\beta^{(t)})) = (X^T W^{(t)} X)^{-1} X^T W^{(t)} Z^{(t)}$ where $Z^{(t)} = X\beta^{(t)} + \frac{d\eta}{d\mu}|_{\mu(\beta^{(t)})}(y - \mu(\beta^{(t)}))$ and $W^{(t)} = (\frac{d\eta}{d\mu}|_{\mu(\beta^{(t)})})^{-1} V(\beta^{(t)})^{-1} (\frac{d\eta}{d\mu}|_{\mu(\beta^{(t)})})^{-1}$.

- Asymptotics: Under regularity conditions, $\hat{\beta}$ is consistent and asymptotic normal with variance $\mathrm{Var}(\hat{\beta}) = I(\beta)^{-1} = \sigma^2(D^T V^{-1} D)^{-1}$.

- Optimality: Estimating function $G(\beta; y)$ if $\mathbb{E}G(\beta; y) = 0, \forall \beta$. Linear estimating function $h(\beta) = H^T(y - \mu(\beta))$. Let $\tilde{\beta}$ solve $h(\beta) = 0$. Taylor expansion $\Rightarrow 0 = h(\tilde{\beta}) = h(\beta) + [\frac{\partial H^T}{\partial \beta}(y - \mu(\beta)) - H^T D](\tilde{\beta} - \beta)$. Take expectations $\Rightarrow 0 = h(\beta) - H^T D(\tilde{\beta} - \beta) \Rightarrow \tilde{\beta} \approx \beta + (H^T D)^{-1} h(\beta)$. Thus, $\mathbb{E}\tilde{\beta} \approx \beta, \mathrm{Var}(\tilde{\beta}) \approx (H^T D)^{-1} H^T V H (D^T H)^{-1}$. Since $\mathrm{Var}(\hat{\beta}) = \sigma^2(D^T V^{-1} D)^{-1}$, $\mathrm{Var}(\hat{\beta})^{-1} - \mathrm{Var}(\tilde{\beta})^{-1} \approx \sigma^{-2} D^T V^{-1} D - \sigma^{-2} D^T H (H^T V H)^{-1} H^T D = \sigma^{-2} D^T V^{-1/2}(I - P_{V^{1/2}H}) V^{-1/2} D \succeq 0 \Rightarrow \mathrm{Var}(\hat{\beta})$ is minimal.

6.7 Longitudinal Data and GLMMs

- Longitudinal study: individuals (subjects) are measured repeatedly over time.

- Questions: (1) average time course of response change; (2) degree of heterogeneity across individuals; (3) factors that predict response change.

- Two perspectives: (1) Marginal/population-averaged: linear, normal, $\mathbb{E}Y_{ij} = \beta_0 + \beta_1 t_{ij}$; (2) Subject-specific: $Y_{ij} = \beta_{0i} + \beta_{1i} t_{ij} + \epsilon_{ij}, \beta_i = \beta + b_i, \mathbb{E}b_i = 0 \Rightarrow \mathbb{E}(Y_{ij}|b_i) = \beta_0 + b_{0i} + (\beta_1 + b_{1i}) t_{ij}, \mathbb{E}Y_{ij} = \beta_0 + \beta_1 t_{ij}$.

- GLMMs: $Y_{ij}|b_i = f_{y_{ij}|b_i}(y_{ij}|b_i) = \exp\{\frac{y_{ij}\theta_{ij} - \psi(\theta_{ij})}{\phi} + c(y_{ij}, \phi)\}$ i.i.d., $\mathbb{E}(Y_{ij}|b_i) = \mu_i = \psi'(\theta_{ij}), \mathrm{Var}(Y_{ij}|b_i) = \phi\psi''(\theta_{ij}), g(\mu_{ij}) = \eta_{ij} = x_{ij}^T\beta + z_{ij}^T b_i$. Typically $b_i \sim \mathcal{N}(0, D)$.

- Mean of Y_{ij}: $\mathbb{E}Y_{ij} = \mathbb{E}[\mathbb{E}(Y_{ij}|b_i)] = \mathbb{E}\mu_{ij} = \mathbb{E}[g^{-1}(x_{ij}^T\beta + z_{ij}^T b_i)]$. Example: $g(\mu) = \log\mu, h(\mu) = g^{-1}(\mu) = e^\mu, z_{ij} \equiv 1, b_i \sim \mathcal{N}(0, \sigma_b^2)$. Then $\mathbb{E}Y_{ij} = e^{x_{ij}^T\beta} e^{\sigma_b^2/2}$.

- Variance of Y_{ij}: $\mathrm{Var}(Y_{ij}) = \mathrm{Var}[\mathbb{E}(Y_{ij}|b_i)] + \mathbb{E}[\mathrm{Var}(Y_{ij}|b_i)] = \mathrm{Var}(\mu_{ij}) + \mathbb{E}[\phi V(\mu_{ij})] = \mathrm{Var}[g^{-1}(x_{ij}^T\beta + z_{ij}^T b_i)] + \mathbb{E}[\phi V(g^{-1}(x_{ij}^T\beta + z_{ij}^T b_i))]$. Example (cont'd): $Y_{ij}|b_i \sim \text{Poisson}(\mu_{ij})$, so that $\mathbb{E}(Y_{ij}|b_i) = \mathrm{Var}(Y_{ij}|b_i) = \mu_{ij}$. Then $\mathrm{Var}(Y_{ij}) = \mathrm{Var}(\mu_{ij}) + \mathbb{E}\mu_{ij}$. Still assume $b_i \sim \mathcal{N}(0, \sigma_b^2)$. Then $\mathrm{Var}(Y_{ij}) = \mathrm{Var}(e^{x_{ij}^T\beta + b_i}) + \mathbb{E}(e^{x_{ij}^T\beta + b_i}) = e^{2x_{ij}^T\beta}(e^{2\sigma_b^2} - e^{\sigma_b^2}) + e^{x_{ij}^T\beta} e^{\sigma_b^2/2} = e^{x_{ij}^T\beta + \sigma_b^2/2}[e^{x_{ij}^T\beta} e^{\sigma_b^2/2}(e^{\sigma_b^2} - 1) + 1] > \mathbb{E}Y_{ij}$ (overdispersed).

- Estimation of GLMMs: (1) Maximum likelihood: log-likelihood $l(\beta, D) = \log\prod_{i=1}^N \int \varphi(b_i; 0, D)\prod_{j=1}^{n_i} f(y_{ij}|b_i) db_i$, or more generally, $\log\int\prod_{i,j} f_{y_{ij}|b_i}(y_{ij}|b_i) f_{b_i}(b_i) db_i$. Let $b_i = Qv_i$, where $v_i \sim \mathcal{N}_q(0, I_q)$ and $D = QQ^T$ (Cholesky decomposition), then $l(\beta, D) = \log\prod_i \int_{\mathbb{R}} \phi(v_{iq}) \cdots \int_{\mathbb{R}} \phi(v_{i1}) \prod_j f(y_{ij}|v_i) dv_i$. Approximate 1-D integral by Gauss-Hermite quadrature: $\int_{\mathbb{R}} e^{-u^2} f(u) du \approx \sum_{k=1}^K w_k f(u_k)$ where u_k are roots of Hermite polynomials $H_k(u)$ and $w_k = \frac{2^{k-1} k! \sqrt{\pi}}{k^2(H_{k-1}(u_k))^2}$. Back to our integral, $\int_{\mathbb{R}} \phi(v_{i1}) \prod_j f(y_{ij}|v_i) dv_{i1} \approx \sum_k \frac{w_k}{\sqrt{\pi}} \prod_j f(y_{ij}|(\sqrt{2}u_k, v_{i2}, \cdots, v_{iq}))$. Adaptive GH: For simplicity, $\int_{\mathbb{R}} \phi(v_i) \prod_j f(y_{ij}|v_i) dv_i \propto$ posterior density of $v_i|y_{ij}$. Approximate $f(v_i|y_{ij})$ by $\phi(v_i; \mu_i, \tau_i^2)$ and write $\int_{\mathbb{R}} \varphi(v_i; \mu_i, \tau_i^2)[\frac{\phi(v_i)\prod_j f(y_{ij}|v_i)}{\varphi(v_i; \mu_i, \tau_i^2)}] dv_i \approx w_{ik} \prod_j f(y_{ij}|u_{ik})$. u_{ik} and w_{ik} are subject-specific, $u_{ik} = \mu_i + \tau_i u_k, w_{ik} = \sqrt{2\pi}\tau_i e^{u_k^2/2} \phi(\mu_i + \tau_i u_k) w_k$. (2) Quasi-likelihood: Similar to IRLS, use current estimates $(\beta^k, D^k, V^k, b_i^k)$ to linearize the model for y_{ij}: $y_{ij} \approx h(\eta_{ij}^k) + x_{ij}^T(\beta - \beta^k) h'(\eta_{ij}^k) + z_{ij}^T(b_i - b_i^k) h'(\eta_{ij}^k) + \epsilon_{ij}$ where $\mathrm{Var}(\epsilon_{ij}) = \phi V(\mu_{ij}^k)$. Let the offset $o_{ij} = h(\eta_{ij}^k) - h'(\eta_{ij}^k) x_{ij}^T \beta^k - h'(\eta_{ij}^k) z_{ij}^T b_i^k$, total residual $\xi_{ij} = h'(\eta_{ij}^k) z_{ij}^T b_i + \epsilon_{ij} \Rightarrow y_{ij} = o_{ij} + h'(\eta_{ij}^k) x_{ij}^T \beta + \xi_{ij}$ (just an OLS).

- Prediction of random effects $b_i \sim \mathcal{N}(0, D)$: posterior $f(b_i|y_i, \beta, D, \phi) = \frac{f(y_i|\beta, D, \phi) f(b_i|D)}{\int f(y_i|\beta, D, \phi) f(b_i|D) db_i}$, estimate b_i by MAP.

- Marginal/Population-average models: Idea: specify only the mean and variance structure of Y_{ij} rather than the full likelihood.

- Generalized Estimating Eqs (GEE): (1) Mean response: $\mu_{ij} = \mathbb{E}Y_{ij} = h(x_{ij}^T\beta)$; (2) Variance: $\mathrm{Var}(Y_{ij}) = \phi V(\mu_{ij})$; (3) Within subject correlation structure, $\Gamma_i \in \mathbb{R}^{n_i\times n_i}$ correlation matrix for subject i, $\Sigma_i = \mathrm{Var}(Y_{ij}) = \phi T_i^{1/2}\Gamma_i T_i^{1/2}$ where $T_i = \mathrm{diag}(V(\mu_{i1}),\cdots,V(\mu_{in_i}))$. "Working correlation structure": $G_i = G_i(\alpha)$ so that working covariance $S_i = \phi T_i^{1/2}G_iT_i^{1/2}$. Choice of G_i may effect efficiency but not consistency. Solve $u(\beta) = \sum_{i=1}^N D_i^T\Sigma_i^{-1}(y_i-\mu_i(\beta)) = 0$ where $D = \frac{\partial\mu_i}{\partial\beta^T} \in \mathbb{R}^{n_i\times p}$. Possible choices of G_i: (1) Unsturctured: $G_i = (\rho_{ik})$, # parameters $\frac{n_i(n_i-1)}{2}$; (2) Independence: $G_i = I_{n_i}$; (3) Exchangeable: $G_i = \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1\end{pmatrix}$; (4) AR(1): $G_i = \begin{pmatrix} 1 & \rho & \cdots & \rho^{n_i-1} \\ \rho & 1 & \cdots & \vdots \\ \vdots & \vdots & \ddots & \rho \\ \rho^{n-1} & \cdots & \rho & 1\end{pmatrix}$; (5) Continuous AR(1) (spatial): $G_i = \begin{pmatrix} 1 & \cdots & \rho^{|t_j-t_k|} \\ & \ddots & \vdots \\ & & 1\end{pmatrix}$.

- Estimation of $\beta, \mathrm{Var}(\hat\beta), \phi, \alpha, \Sigma_i$: Find $\hat\phi$ and $\hat\alpha$: Let $r_{ij} = \frac{y_{ij}-\mu_{ij}(\hat\beta)}{V(\mu_{ij}(\hat\beta))^{1/2}}$, then $\hat\phi = \frac{1}{n-p}\sum_{i=1}^N\sum_{j=1}^{n_i} r_{ij}^2$ and $\hat\rho_{jk} = \frac{1}{N\hat\phi}\sum_{i=1}^N r_{ij}r_{ik}$ or $\frac{1}{N\hat\phi}\sum_{i=1}^{n_i}\frac{1}{n_i-1}\sum_{j=1}^{n_i-1} r_{ij}r_{i,j+1}$ (corresponds to G_i's in the above (1) and (3)).

- Iterative procedure: Step 1: Initialize β_0 (acquired by assuming complete independence). Repeat until convergence: Step 2: Given β, estimate ϕ and α for the working correlation structure; Step 3: Use current estimates of β, ϕ and $\hat\alpha$ to form $\hat\Sigma_i$ and solve GEE for a new $\hat\beta$.

- Inference: $\hat\beta \sim \mathcal{N}_p(\beta, V(\hat\beta))$ for n large where $V(\hat\beta) = \phi(\sum_{i=1}^n D_i^TS_i^{-1}D_i)$.

- Robust covariance estimation: In general, G_i are misspecified, then

$$\hat V_{\hat\beta}^{\text{robust}} = \hat\phi(\sum_{i=1}^N \hat D_i^T\hat S_i^{-1}\hat D_i)^{-1}(\sum_{i=1}^N \hat D_i^T\hat S_i^{-1}\hat C_i\hat S_i\hat D_i)(\sum_{i=1}^N \hat D_i^T\hat S_i^{-1}\hat D_i)^{-1}$$

  where $\hat C_i = (y_i-\mu_i(\hat\beta))(y_i-\mu_i(\hat\beta))^T$. ("sandwich estimator")

7 Nonparametric Statistics

- Nonparametric models: (1) Kernel smoothing (local polynomial regression); (2) Splines (e.g. B-splines): smoothing splines (+penalty), # knots $= n$; regression splines, # knots $<< n$; (3) Orthogonal series expansion; (4) Wavelets; (5) Sharp-constrained: isotonic, log-concavity.

- Model: $Y = r(X)+\epsilon_i, r(x) = \mathbb{E}(Y|X=x)$. Linear smoothers: $\hat r_n(x) = \sum_{i=1}^n l_i(x)y_i, (\hat r_n(x_1),\cdots,\hat r_n(x_n))^T := \hat r = LY, L = (l_j(x_i))_{ij}$, effective d.f. $v = \mathrm{rank}(L), \sum_{i=1}^n l_i(x) = 1$. Example: (1) Histogram regression: $\hat r_n(x) = \frac{1}{k_j}\sum_{i:x_i\in B_j} y_i = \sum_{i=1}^n l_i(x)y_i$, where $l(x) = (0,\cdots,0,1/k_j,\cdots,1/k_j,0,\cdots,0)^T$; (2) Local average: Let $B_x = \{i : |x_i - x| \le h\}, \hat r_n(x) = \frac{1}{n}\sum_{i\in B_x} y_i$.

- Leave-one-out CV: $\hat R(h) = \frac{1}{n}\sum_{i=1}^n (y_i - \hat r^{(-i)}(x_i))^2, R(h) = \mathbb{E}[\frac{1}{n}\sum_{i=1}^n(\hat r_n(x_i)-r(x_i))^2]$. Linear smoother: $R = \frac{1}{n}\sum_{i=1}^n(\frac{y_i-\hat r_n(x_i)}{1-l_{ii}})^2$. Generalized CV: Replace l_{ii} by v/n.

- Kernel Smoothing: N-W estimator: $\hat r_n(x) = \sum_{i=1}^n l_i(x)y_i$ where $l_i(x) = \frac{K(\frac{x-x_i}{h})}{\sum_{j=1}^n K(\frac{x-x_j}{h})}$. $f(x)$ is pdf of x, and MSE $R(\hat r_n, r) = \frac{h_n^4}{4}\left(\int x^2K(x)dx\right)^2\int\left(r''(x)+2r'(x)\frac{f'(x)}{f(x)}\right)^2 dx + \frac{\sigma^2\int K^2(x)dx}{nh_n}\int\frac{1}{f(x)}dx + \cdots$ as $h_n\to 0$ and $nh_n\to\infty$, optimal $h^* \sim n^{-1/5}, R(\hat r_n, r) \sim n^{-4/5}$.

- Local Polynomial/Linear Regression: $\min_c \sum_{i=1}^n w_i(x)(y_i-c)^2$ or $\min \sum_{i=1}^n w_i(x)(y_i-P_x(x_i;c))^2$. N-W boundary bias: $O(h_n)$; LLR: $O(h_n^2)$.

- Smoothing Splines: $\min M(\lambda)=\sum_{i=1}^n(y_i-\hat{r}_n(x_i))^2+\lambda J(r)$ where $J(r)=\int r''(x)^2dx$ or $\min ||Y-B\beta||^2+\lambda\beta^T\Omega\beta$ where $B_{ij}=B_j(x_i), \Omega_{jk}=\int B_j''(x)B_k''(x)dx$.

- Equivalence between kernel and spline estimators: spline estimator can be expressed by $\hat{r}_n(x)=\sum_{i=1}^n l_i(x)y_i$ with $l_i(x)\approx\frac{1}{f(x_i)h(x_i)}K(\frac{x-x_i}{h(x_i)})$ (e.g. $K(t)=\frac{1}{2}e^{-|t|/2}\sin(\frac{|t|}{\sqrt{2}}+\frac{\pi}{4})$).

- Regression splines: Regress Y_i directly (w/o penalty) on B: $\hat{r}=LY, L=B(B^TB)^{-1}B^T$. Tuning parameter: # knots of B.

- Variance estimation: $\hat{\sigma}^2=\frac{1}{n-2v+\tilde{v}}\sum_{i=1}^n(y_i-\hat{r}(x_i))^2$ with $v=\text{tr}(L)$ and $\tilde{v}=\text{tr}(L^TL)=\sum_{i=1}^n||l(x_i)||^2$ (quadratic form: $\hat{\sigma}^2=\frac{Y^T\Lambda Y}{\text{tr}(\Lambda)}$ and $\Lambda=(I-L)^T(I-L)=I-L-L^T+L^TL$). $\mathbb{E}\hat{\sigma}^2=\frac{\sigma^2\text{tr}(\Lambda)+r^T\Lambda r}{\text{tr}(\Lambda)}=\sigma^2+\frac{r^T\Lambda r}{\text{tr}(\Lambda)}$. If r is sufficiently smooth, $v=o(n)$ and $\tilde{v}=o(n)$, $\hat{\sigma}^2$ is consistent for σ^2. Another estimator: $\tilde{\sigma}^2=\frac{1}{2(n-1)}\sum_{i=1}^{n-1}(y_{i+1}-y_i)^2$.

- Heteroscedastic case: $y_i=r(x_i)+\sigma(x_i)\epsilon_i\Rightarrow(y_i-r(x_i))^2=\sigma^2(x_i)\epsilon_i^2\Rightarrow Z_i:=\log(y_i-r(x_i))^2=\log\sigma^2(x_i)+\log\epsilon_i^2$. Regress Z_i on x_i using a nonparametric method, denoted as $\hat{q}(x_i)$. Then $\hat{\sigma}^2(x)=e^{\hat{q}(x)}$.

- Confidence bands: $\hat{r}_n(x)\pm cs_n(x)$. $\bar{r}_n(x)=\mathbb{E}\hat{r}_n(x)$, $\frac{\hat{r}_n(x)-r(x)}{s_n(x)}=\frac{\hat{r}_n(x)-\bar{r}_n(x)}{s_n(x)}+\frac{\bar{r}_n(x)-r(x)}{s_n(x)}=\underbrace{Z_n(x)}_{\sim\mathcal{N}(0,1)}+\underbrace{\frac{\text{bias}(\hat{r}_n(x))}{\sqrt{\text{Var}(\hat{r}_n(x))}}}_{\text{smoothing bias}}$.

  Remedy: (1) CB for $\bar{r}_n(x)$ rather than $r(x)$; (2) undersmooth. To find c, consider $I(x)=(\hat{r}_n(x)-c\sigma||l(x)||,\hat{r}_n(x)+c\sigma||l(x)||), x\in[a,b]$. Assume σ is known, then $P(\bar{r}_n(x)\notin I(x)\text{ for some }x\in[a,b])=P(\sup_x\frac{|\hat{r}_n(x)-\bar{r}_n(x)|}{\sigma||l(x)||}>c)=P(\sup_x|\sum_{i=1}^n\frac{l_i(x)}{||l(x)||}\frac{\epsilon_i}{\sigma}|>c):=P(\sup_x|W_n(x)|>c)$.

  Lemma 7.1 $P(\sup_x|W_n(x)|>c)=2(1-\Phi(c))+\frac{\kappa_0}{\pi}e^{-c^2/2}$ where $\kappa_0=\int_a^b||T'(x)||dx$ and $T_i(x)=\frac{l_i(x)}{\sigma||l(x)||}$.

  Choose c to solve $2(1-\Phi(c))+\frac{\kappa_0}{\pi}e^{-c^2/2}=\alpha$. If σ is unknown, then $P(\sup_x|W_n(x)|>c)\approx P(|T_m|>c)+\frac{\kappa_0}{\pi}(1+\frac{c^2}{m})^{-m/2}, T_m\sim t_m, m=n-\text{tr}(L)$. Average coverage: $C=\frac{1}{b-a}\int_a^bP(\bar{r}_n(x)\in I(x))dx$.

- Local likelihood for nonparametric EM regression: Data $(x_i,y_i), i=1,\cdots,n, y_i\in\{0,1\}$. $Y_i\sim\text{Bernolli}(r(x_i))$. log-likelihood: $l(r)=\sum_{i=1}^n\log\{r(x_i)^{y_i}(1-r(x_i))^{1-y_i}\}=\sum_{i=1}^n\{y_i\log\frac{r(x_i)}{1-r(x_i)}+\log(1-r(x_i))\}:=\sum_{i=1}^nl(y_i,\xi(x_i))$ where $\xi(x_i)=\log\frac{r(x_i)}{1-r(x_i)}$. Local linear regression: approximate $\xi(x)$ around x_0 by $a_0+a_1(x-x_0)$. Local loglik $l_x(a)=\sum_{i=1}^nK(\frac{x-x_i}{h})l(y_i,a_0+a_1(x-x_0))$. MLE: $\hat{a}(x)=(\hat{a}_0(x),\hat{a}_1(x))\Rightarrow\hat{r}_n(x)=\frac{e^{\hat{a}_0(x)}}{1+e^{\hat{a}_0(x)}}$. Leave-one-out CV: $\text{CV}=\sum_{i=1}^nl(y_i,\hat{\xi}_{(-i)}(x_i))$.

- Multiple nonparametric regression: $y_i=r(x_i)+\epsilon_i, x_i=(x_{i1},\cdots,x_{id})^T$. Optimal rate of convergence: $n^{-\frac{2\beta}{2\beta+d}}$.

- Method 1: Local regression: multivariate kernel $K_H(x)=\frac{1}{\det(H)^{1/2}}K(H^{-1/2}X)$ where H is $d\times d$ bandwidth metrix. Univariate kernel: $h^{-d}K(\frac{||x||}{h})$. Minimise $\sum_{i=1}^nK(\frac{||x_i-x||}{h})(y_i-a_0-\sum_{j=1}^da_j(x_{ij}-\bar{x}_j))^2$ at $x=(x_1,\cdots,x_d)^T$. $\hat{a}=(X_x^TW_xX_x)^{-1}X_x^TW_xY$.

- Method 2: Splines: $d=2$ thin-plate: $\min_r\sum_{i=1}^n(y_i-r(x_i))^2+\lambda J(r), J(r)=\iint_{\mathbb{R}^2}(\frac{\partial^2r(x)}{\partial x_1^2}+2\frac{\partial^2r(x)}{\partial x_1\partial x_2}+\frac{\partial^2r(x)}{\partial x_2^2})dx_1dx_2$.

- Method 3: Additive model: $y_i=\alpha+\sum_{j=1}^dr_j(x_{ij})+\epsilon_i$, with $\sum_{i=1}^nr_j(x_{ij})=0$. Backfitting algorithm: Set $\hat{\alpha}=\bar{y}$ and initialize $\hat{r}_1,\cdots,\hat{r}_d$; Repeat for $j=1,\cdots,d,1,\cdots,d,\cdots$ until convergence: $\tilde{y}_i=y_i-\hat{\alpha}-\sum_{k\neq j}r_k(x_{ik}),\forall i$; Regress $\tilde{y}_i$ on x_j by a 1-d smoother to set $\hat{r}_j$; $\hat{r}_j(x)\leftarrow\hat{r}_j(x)-\frac{1}{n}\sum_i\hat{r}_j(x_i)$.

8 Survival Analysis

- Survival data: Survival/failure time/time-to-event, binary data subject to censoring.

- Examples: (1) Heart transplant. Response: survival time from date of admission. Covariates (might be time-varying): surgery, age, donor/receipient variable. Censoring: drop-out (e.g. withdrawal, loss to follow up, death due to irrelevant reasons). Goal: effect of heart transplant (prediction) on survival. (2) Accelerated life test:

Temp	Hours to Failure
150℃	0 until 8064h
170℃	7 failures + 3 successes
190℃	5 + 5
220℃	5

  Question: Failure time distribution at 130℃?

- Censoring type: (1) predetermined time; (2) # failures reached; (3) right/left/interval censoring.

- Survival function for continuous T: $S(t) = P(T > t) = 1-F(t)$. Hazard function: $\lambda(t) = \lim_{h\to 0+} \frac{P(t\le T<t+h|T\ge t)}{h} = \lim_{h\to 0+} \frac{P(t\le T<t+h)}{hP(T\ge t)} = \frac{f(t)}{S(t)} = -\frac{d}{dt}\log S(t) \Rightarrow S(t) = \exp(-\int_0^t \lambda(s)ds) := e^{-\Lambda(t)}$ where $\Lambda(t)$ is defined as cumulative hazard function. $f(t) = \lambda(t)e^{-\Lambda(t)}$. Assumption: $\exists t = \int_0^t \lambda(s)ds < \infty$ and $\int_0^\infty \lambda(s)ds = \infty$.

- Independent censoring: censoring time C, at-risk indicator $y(t) = \begin{cases} 1, & \text{at risk} \\ 0, & \text{failed or censored} \end{cases}$, covariates x. Then $\lambda(t) = \lim_{h\to 0+} \frac{P(T\in[t,t+h)|x,T\ge t)}{h} = \lim_{h\to 0+} \frac{P(T\in[t,t+h)|x,T\ge t,y(t)=1)}{h} = \lim_{h\to 0+} \frac{P(T\in[t,t+h)|x,T\ge t,C\ge t)}{h}$.

- Nonparameteric estimation of survival function: w/o censoring: $F_n(t) = \frac{\#\{i:T_i\le t\}}{n}$. How to deal with censoring? At-risk set $R(t) = \{i : y_i(t) = 1\}$. $t_0 = 0, t_{k+1} = \infty, t_1 < \cdots < t_k$ from $n = n_0$ subjects. d_j failed at t_j, m_j censored in $[t_j, t_{j+1})$, $n_j = \#$ at risk just prior to t_j and $n_j = \sum_{l=j}^k (d_l + m_l)$. $P(T = t_j) = S(t_{j-}) - S(t_j), P(T > t_{jl}) = S(t_{jl})$. Assume independent censoring, the likelihood $L = \prod_{j=0}^k \{S(t_{j-}) - S(t_j)\}^{d_j} \prod_{l=1}^{m_j} S(t_{jl})$. NPMLE: $\hat{S} = \arg\max_S L(S)$. Facts: (1) discontinuous at all t_j; (2) $S(t_{jl}) = S(t_j), \forall j, l$. $\hat{S}(t)$ with hazard components $\hat\lambda_1, \cdots \hat\lambda_k$ at $t_1, \cdots, t_k$, respectively: $\hat{S}(t_j) = \prod_{l=1}^j (1-\hat\lambda_l), \hat{S}(t_{j-}) = \prod_{l=1}^{j-1}(1-\hat\lambda_l)$. To determine the values of λ_j, maximize $L(\lambda) = \prod_{j=1}^k \lambda_k^{d_j} \prod_{l=1}^{j-1}(1-\lambda_l)^{d_j} \prod_{l=1}^{j-1}(1-\lambda_l)^{m_j} = \prod_{j=1}^k \lambda_j^{d_j}(1-\lambda_j)^{n_j-d_j}$. $\frac{\partial}{\partial\lambda_j}\log L(\lambda) = 0 \Rightarrow \hat\lambda_j = \frac{d_j}{n_j}$. NPMLE of $S(t) : \hat{S}(t) = \prod_{j:t_j\le t} \frac{n_j-d_j}{n_j}$ (Kplan-Meier product limit estimator).

- Inference: $\widehat{\text{Var}}(\hat\lambda_j) = \frac{\hat\lambda_j(1-\hat\lambda_j)}{n_j} = \frac{d_j(n_j-d_j)}{n_j^3}$. For $\log\hat{S}(t) = \sum_{j:t_j\le t}\log(1-\hat\lambda_j)$, by the delta method, $\widehat{\text{Var}}(\log\hat{S}(t)) = \sum_{j:t_j\le t}\frac{1}{(1-\hat\lambda_j)^2}\widehat{\text{Var}}(\hat\lambda_j) = \sum_{j:t_j\le t}\frac{d_j}{n_j(n_j-d_j)}$ and $\widehat{\text{Var}}(\hat{S}(t)) = (\hat{S}(t))^2\sum_{j:t_j\le t}\frac{d_j}{n_j(n_j-d_j)}$. Hence $\hat\Lambda(t) = \int_0^t\hat\lambda(s)ds = \sum_{j:t_j\le t}\hat\lambda_j$ (Nelson-Aalen estimator).

- Counting process: $N_i(t) = \#$ failures in $[0,t]$, $y_i(t)$ is at-risk process, and $N.(t) = \sum_{i=1}^n N_i(t), y.(t) = \sum_{i=1}^n y_i(t)$. Nelson-Aalen: $\hat\Lambda(t) = \int_0^t \frac{I(y.(u)>0)}{y.(u)}dN(u)$.

- Cox's proportional hazards model: Conditional hazard function $\lambda(t|z) = \lim\limits_{h\to 0+}\frac{P(t\le T<t+h|T>t,z)}{h} = \lambda_0(t)e^{z(t)^T\beta} = \lambda_0(t)r(t,z)$. Other forms: (1) additive hazards $\lambda(t|z) = \lambda_0(t) + z(t)^T\beta$; (2) NP/SP Cox: $r(t,z) = \sum_{j=1}^d g_j(z_j(t))$.

- Estimation of β: Partial likelihood: Suppose the information of y is constrained in $A_1, B_1, \cdots, A_m, B_m$. The joint density $\prod_{j=1}^m f(b_j|b^{(j-1)}, a^{(j-1)}; \theta, \beta)\prod_{j=1}^m f(a_j|b^{(j)}, a^{(j-1)}; \beta)$. In our case, $B_j =$ censoring information in $[t_{j-1}, t_j]$ + information that 1 individual fails in $[t_j, t_j + dt_j)$ and $A_j =$ individual j fails in $[t_j, t_j + dt_j)$. The j-th failure contributes $L_j(\beta) = f(a_j|b^{(j)}, a^{(j-1)}; \beta) = \frac{\lambda(t_j|z_j)dt_j}{\sum_{l\in R(t_j)}\lambda(t_j|z_l)dt_j} \overset{\text{Cox model}}{=} \frac{e^{z_j(t_j)^T\beta}}{\sum_{l\in R(t_j)}e^{z_l(t_j)^T\beta}} \Rightarrow L(\beta) = \prod_{j=1}^n \frac{e^{z_j(t_j)^T\beta}}{\sum_{l\in R(t_j)}e^{z_l(t_j)^T\beta}}, \hat\beta = \arg\max_\beta L(\beta)$. Partial score equation:

$$0 = u(\beta) = \frac{\partial\log L}{\partial\beta} = \sum_{j=1}^m\left\{z_j(t_j) - \sum_{l\in R(t_j)} z_l(t_j)\frac{e^{z_l(t_j)^T\beta}}{\sum_{k\in R(t_j)}e^{z_k(t_j)^T\beta}}\right\} := \sum_{j=1}^m\{z_j(t_j) - \bar{z}(\beta, t_j)\}$$

Furthermore, the observed information matrix

$$I_n(\beta) = -\frac{\partial^2 \log L}{\partial\beta\partial\beta^T} = \sum_{j=1}^{k}\sum_{l\in R(t_j)} z_l(t_j)\left\{z_l(t_j) - \bar{z}(\beta, t_j)\right\}^T p_l(\beta, t_j) = \sum_{j=1}^{k}\sum_{l\in R(t_j)} \left\{z_l(t_j) - \bar{z}(\beta, t_j)\right\}^{\otimes 2} p_l(\beta, t_j)$$

where $p_l(\beta, t_j) = \frac{\lambda(t_j|z_l)dt_j}{\sum_{k\in R(t_j)}\lambda(t_j|z_k)dt_j}$ and $b^{\otimes 2} = bb^T$. Asymptotics: $\sqrt{n}(\hat{\beta} - \beta_0) \to_d \mathcal{N}_d(0, I(\beta_0)^{-1})$. Using counting process notation,

$$\begin{aligned}
l(\beta) &= \sum_{i=1}^{n}\int_0^\infty \left[z_i(t) - \log\left(\sum_{i=1}^{n} y_i(t)e^{z_l(t_j)^T\beta}\right)\right] dN_i(t)\\
u(\beta) &= \sum_{i=1}^{n}\int_0^\infty [z_i(t) - \bar{z}(\beta, t)]dN_i(t)\\
I_n(\beta) &= \sum_{i=1}^{n}\int_0^\infty [z_i(t) - \bar{z}(\beta, t)]^{\otimes 2} e^{z_i(t)^T\beta} dN_i(t)
\end{aligned}$$

- Breaking the ties: average partial likelihood over all possible permutations:

$$\prod_{j=1}^{k} e^{s_j(t_j)} \sum_{P\in Q_j}\prod_{r=1}^{d_j}\left\{\sum_{l\in R(t_j,P,r)} e^{z_l(t_j)^T\beta}\right\}^{-1}$$

  where Q_j is the set of $d_j!$ permutations and $s_j = \sum_{i=1}^{d_j} z_{j_i}(t_j)$.

- Approximations: (1) Breslow's:

$$L = \prod_{j=1}^{k}\frac{e^{s_j(t_j)^T\beta}}{(\sum_{l\in R(t_j)} e^{z_l(t_j)^T\beta})^{d_j}}$$

  (2) Efron's:

$$L = \prod_{j=1}^{k}\frac{e^{s_j(t_j)^T\beta}}{\prod_{r=0}^{d_j-1}\left(\sum_{k\in R(t_j)} e^{z_l(t_j)^T\beta} - r\bar{A}(\beta, t_j)\right)}$$

  where $\bar{A}(,\beta, t_j) = \frac{1}{d_j}\sum_{l\in D(t_j)} e^{z_l(t_j)^T\beta}$ and D_j is the set of failures at t_j.

- Estimation of $\Lambda_0(\cdot)$: NPMLE: $L = \prod_{j=1}^{k}\prod_{l\in D_j}\{S(t_j-|z_l) - S(t_j|z_l)\}\prod_{l\in C_j} S(t_j|z_l)$ where C_j is the set of censoring at t_j. Let $\Lambda_0(t) = \sum_{j:t_j\le t}(1-\alpha_j)$, maximise $\prod_{j=1}^{k}\prod_{l\in D_j}(1-\alpha_j^{e^{z_l(t_j)^T\beta}})\prod_{l\in C_j}\alpha_j^{e^{z_l(t_j)^T\beta}}$ over $\alpha = (\alpha_1, \cdots, \alpha_k)^T$. $\hat{\alpha}_j = \left\{1 - \frac{e^{z_j(t_j)^T\beta}}{\sum_{l\in R(t_j)} e^{z_l(t_j)^T\beta}}\right\} e^{-z_j(t_j)^T\beta}$, $\hat{\Lambda}_0(t) = \sum_{j:t_j\le t}(1 - \hat{\alpha}_j)$.