北京大學
PEKING UNIVERSITY

Notes for **Statistical Modeling and Methods**

lectured by Fang Yao* LaTeXed by T.-Y. Li†

2020 Spring

Abstract

The course deals with a variety of statistical models and methods that generalize classical linear regression to include many others that have been found useful in statistical analysis and applications. We will first review key concepts in linear regression analysis, and expand the scope in depth to generalized linear models, then head for several important directions: nonparametric regression and generalized additive models, linear and generalized linear mixed models, generalized estimating equations for correlated data, dimension reduction and so on (if time permits). The course is a mixture of theory and applications and includes computer projects featuring R programming.

 Please feel free to leave a message at https://zhuanlan.zhihu.com/p/106896222 (under construction). Any comments are welcome.

*Homepage: http://www.math.pku.edu.cn/teachers/yaof/
†E-Mail: kellty@pku.edu.cn

Contents

§1 Review of Linear Algebra (2020/2/19) · · · · · 1
§2 Review of Linear Algebra Cont'd (2020/2/26) · · · · · 1
§3 Review of Linear Algebra Cont'd (2020/2/28) · · · · · 2
§4 Noncentral Chi-square Distribution (2020/3/4) · · · · · 3
§5 Prediction Accuracy (2020/3/11) · · · · · 4
§6 Curse of Dimensionality & Estimability in Linear Models (2020/3/13) · · · · · 5
§7 Properties of Ordinary Least Squares (2020/3/18) · · · · · 6
§8 Restrictions on a Subset of Regression Coefficients (2020/3/25) · · · · · 7
§9 Inference for Linear Regression Models (2020/3/27) · · · · · 8
§10 Multiple Testing Techniques (2020/4/1) · · · · · 10
§11 Model Selection (2020/4/8) · · · · · 11
§12 Model Selection Cont'd (2020/4/10) · · · · · 13
§13 Model Assessment & Principle of Maximum Likelihood (2020/4/15) · · · · · 14
§14 Review of Likelihood Theory (2020/4/22) · · · · · 15
§15 Likelihood Cont'd & Intro to GLMs (2020/4/24) · · · · · 17
§16 Generalized Linear Models & Exponential Families (2020/4/29) · · · · · 19
§17 Justification of Exponential Families & Link Functions (2020/5/13) · · · · · 21
§18 Goodness of Fit & Algorithms for Model Estimation (2020/5/20) · · · · · 22
§19 Fitting and Inference Cont'd & Important Examples (2020/5/22) · · · · · 26
§20 Applications of GLMs (2020/5/27) · · · · · 28
§21 Diagnostics in GLMs & Quasi-Likelihood Methods (2020/6/3) · · · · · 31
§22 Nonparametric Regression & Additional Material (2020/6/5) · · · · · 34
§23 Mixed Effects & Generalized Estimating Equations · · · · · 59

References

[RTSH] C.R. Rao, H. Toutenburg, Shalabh, & C. Heumann (2008). *Linear Models and Generalizations: Least Squares and Alternatives* (3rd ed.). Springer.

[McCN] P. McCullagh, & J.A. Nelder (1989). *Generalized Linear Models* (2nd ed.). Chapman & Hall/CRC.

[HT] T.J. Hastie, & R.J. Tibshirani (1990). *Generalized Additive Models*. Chapman & Hall/CRC.

[FG] J. Fan, & I. Gijbels (1996). *Local Polynomial Modelling and Its Applications*. Springer.

[DHLZ] P.J. Diggle, P.J. Heagerty, K.-Y. Liang, & S.L. Zeger (2013). *Analysis of Longitudinal Data* (2nd ed.). Oxford University Press.

[F] J.J. Faraway (2016). *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models* (2nd ed.). Chapman & Hall/CRC.

[A] A. Agresti (2015). *Foundations of Linear and Generalized Linear Models*. Wiley.

[C] R. Christensen (2020). *Plane Answers to Complex Questions: The Theory of Linear Models* (5th ed.). Springer.

[DS] P.K. Dunn, & G.K. Smyth (2018). *Generalized Linear Models With Examples in R*. Springer.

[T] G. Tutz (2012). *Regression for Categorical Data*. Cambridge University Press.

[Z] *Topics in Matrix Theory* (in Chinese). https://zhuanlan.zhihu.com/Topics-in-Matrix-Theory

List of Figures

1 Orthogonal Projection in Linear Regression Models 6
2 Tradeoff between Parsimony and Fidelity 12
3 Prediction Error in Training and in Test 13
4 A Graphical Comparison of Link Functions with Logit 27
5 Local Polynomial Kernel Estimates 35

List of Tables

1 Overview of Generalized Linear Models 18
2 Commonly Used Exponential Families 21
3 Cross-Classification of Disease and Exposure 29
4 Quasi-Likelihoods Associated with Some Simple Variance Functions 33

Statisticians, like artists, have the bad habit of falling in love with their models.

—— George E. P. Box

§1 Review of Linear Algebra (2020/2/19)

We say that vectors $u_1, \dots, u_m \in \mathbb{R}^n$ are **linearly independent** if for scalars $c_1, \dots, c_m \in \mathbb{R}$,

$$c_1 u_1 + \cdots + c_m u_m = 0_n \implies c_1 = \cdots = c_m = 0.$$

The **rank** of a matrix $A \in \mathbb{R}^{m\times n}$ is the maximum number of its linearly independent rows/columns, which is at most $\min(m, n)$. The **column space** of $A = (a_{ij})_{1\le j\le n}^{1\le i\le m}$, spanned by its columns, is

$$C(A) = \mathrm{Col}(A) = \{Ax \in \mathbb{R}^m : x \in \mathbb{R}^n\},$$

the *range* of the linear transformation $(x_j)^{1\le j\le n} \mapsto (\sum_{j=1}^n a_{ij}x_j)^{1\le i\le m}$. Clearly $\mathrm{rank}\,(A) = \dim C(A)$. The **row space** of A is the column space of its transpose $A' = (a_{ji})_{1\le j\le m}^{1\le i\le n} \in \mathbb{R}^{n\times m}$.

For example, in the linear regression model $Y = X\beta + \varepsilon$, i.e., $y_i = x_i'\beta + \varepsilon_i$ for $i = 1, \cdots, n$, where $x_i = (x_{ij})^{1\le j\le p}$ characterizes the i^{th} observation/subject, it is usually thought that each of p variables $x_{(j)} = (x_{ij})^{1\le i\le n}$ contains useful information for modeling the response Y. So, naturally we may hope the data/design matrix $X = (x_{ij})_{1\le j\le p}^{1\le i\le n}$ to have full (column) rank.

Note that not all linear models feature full rank design matrices. Analysis of variance (ANOVA) focuses on the differences among treatment effects (group means), and the generalized linear model (GLM) allows the linear combination of predictors to be related to the response variable via a link function, not necessarily requiring full rank design matrices, just to mention a few. Rather than mathematics driven by models, what should be emphasized is statistical interpretation driven by data or problems.

For a square matrix $A = (a_{ij})_{1\le i,j\le n}$, two important quantities are its **trace**, denoted by

$$\mathrm{tr}\,(A) = \sum_{i=1}^n a_{ii},$$

and its **determinant**, denoted by

$$\det(A) = \sum_{\sigma\in\mathfrak{S}_n} \mathrm{sign}(\sigma) \prod_{i=1}^n a_{i\sigma(i)}.$$

If $Q' = Q^{-1}$, i.e., $Q'Q = QQ' = I_n$, then Q is said to be **orthogonal**, geometrically representing a *rotation* that preserves inner products in the sense that $(Qx)'(Qy) = x'y$ for $x, y \in \mathbb{R}^n$. Thus, $x \perp y \implies Qx \perp Qy$. If $A \in \mathbb{R}^{n\times n}$ is **symmetric**, i.e., $A' = A$, then its **eigen-decomposition**[i] takes the form

$$A = Q\Lambda Q' = (e_1, \dots, e_n)\,\mathrm{diag}(\lambda_1, \dots, \lambda_n)\,(e_1, \dots, e_n)' = \sum_{i=1}^n \lambda_i e_i e_i',$$

where λ_i's are eigenvalues and e_i's are orthonormal eigenvectors. Note that $e_ie_i' = \mathrm{proj}_{e_i}$ has rank one. The case $\lambda_j = \lambda_k$ for some $j \ne k$ (multiplicity) tends not to be considered in statistical analysis, since observations are assumed to be drawn randomly.

§2 Review of Linear Algebra Cont'd (2020/2/26)

For symmetric $A = \sum_{j=1}^{\mathrm{rank}(A)} \lambda_j e_j e_j' \in \mathbb{R}^{n\times n}$ and well-defined $g(\cdot)$, we have $g(A) = \sum_{j=1}^{\mathrm{rank}(A)} g(\lambda_j) e_j e_j'$. The matrix A is **positive (semi-)definite** if the corresponding **quadratic form**

$$Q_A(x) = x'Ax = \sum \lambda_j (e_j'x)^2, \quad x \in \mathbb{R}^n$$

is $> (\ge) 0$ for $x \ne 0_n$. Then, the **square root** $A^{1/2}$ of A exists and is unique.

The eigen-decomposition $A = Q\Lambda Q'$ of positive definite $A \in \mathbb{R}^{n\times n}$ gives a geometric interpretation. Let $y = Q'x$, then $x'Ax = y'\Lambda y$, whose contours are rotated *ellipsoids*. The orientation of $\{x \in \mathbb{R}^n : x'Ax = c\}$ consists of e_j's, with lengths $\sqrt{c/\lambda_j}$, respectively.

Note that positive definite $A \in \mathbb{R}^{n\times n}$ induces an **inner product** $\langle x, y\rangle_A = x'Ay$ for $x, y \in \mathbb{R}^n$ and the **Cauchy-Schwarz inequality** reads $\langle x, y\rangle_A^2 \le \langle x, x\rangle_A \langle y, y\rangle_A$, with equality holding if and only if $x // y$. Some variants include $(x'y)^2 \le \langle x, x\rangle_A \langle y, y\rangle_{A^{-1}}$ and $\langle y, y\rangle_{A^{-1}} = \sup_{x\ne 0_n} \frac{(x'y)^2}{\langle x,x\rangle_A}$.

[i] cf. https://zhuanlan.zhihu.com/p/75250722

The **singular value decomposition**[ii] (SVD) of an arbitrary $X \in \mathbb{R}^{n\times p}$ of rank r is

$$X = U\Sigma V' = (u_1, \ldots, u_r)\,\mathrm{diag}(\sigma_1, \ldots, \sigma_r)\,(v_1, \ldots, v_r)' = \sum_{k=1}^{r} \sigma_k u_k v_k'$$

such that $\sigma_k > 0$ decreases with k, and u_k's and v_k's form orthonormal bases of $C(X)$ and $C(X')$, respectively. Note that $Xv_k = \sigma_k u_k$ and $X'u_k = \sigma_k v_k$. From the perspective of matrix completion, SVD provides the best *low-rank approximation* in that

$$\sum_{k=1}^{m} \sigma_k u_k v_k' = \underset{Y\in\mathbb{R}^{n\times p}\,:\,\mathrm{rank}(Y)\le m}{\arg\min} \|X - Y\|_{\mathrm{F}}\,, \quad \forall m \le \mathrm{rank}\,(X)\,,$$

where $\|\bullet\|_{\mathrm{F}} = \sqrt{\mathrm{tr}\,(\bullet'\bullet)}$ denotes the *Frobenius norm*.

A matrix $P \in \mathbb{R}^{n\times n}$ is said to be **idempotent** if $P^2 = P$. Suppose that P is symmetric. The eigenvalues of P lie in $\{1, 0\}$, and thus $\mathrm{tr}\,(P) = \mathrm{rank}\,(P)$. Note that $P = \mathrm{proj}_{C(P)}$ and $I - P = \mathrm{proj}_{C(P)^\perp}$.

For any $A \in \mathbb{R}^{m\times n}$, we call $A^- \in \mathbb{R}^{n\times m}$ a **generalized inverse** of A if $AA^-A = A$. For instance, the **Moore-Penrose inverse** $X^+ = \sum_{k=1}^{r} \sigma_k^{-1} v_k u_k'$ is a generalized inverse of $X = \sum_{k=1}^{r} \sigma_k u_k v_k'$, which also satisfies that $X^+XX^+ = X^+$, and XX^+ and X^+X are symmetric.

Theorem. The matrix $P_X = X(X'X)^-X'$ does not depend on the choice of $(X'X)^-$. Indeed, P_X is the *orthogonal projection* onto the column space $\mathrm{Col}(X)$ of $X \in \mathbb{R}^{n\times p}$.

Proof. For any $v \in \mathbb{R}^n$, write $v = x + w$ for $x \in \mathrm{Col}(X)$ and $w \in \mathrm{Col}(X)^\perp$, which exist and are unique since $\mathbb{R}^n = \mathrm{Col}(X) \bigoplus \mathrm{Col}(X)^\perp$. It follows from $X'w = 0_p$ that $P_X w = 0_n$ and $P_X v = P_X x$. To prove $P_X x = x$, it is equivalent to show $\boxed{X(X'X)^-X'X = X}$, or $u'X(X'X)^-X'X = u'X$, $\forall u \in \mathbb{R}^n$. Now that $X'X(X'X)^-X'X = X'X$ by definition, it suffices to point out that $u'X = z'X'X$ for some $z \in \mathbb{R}^p$. What we need is exactly that $\boxed{\mathrm{Col}(X') = \mathrm{Col}(X'X)}$. Clearly $\mathrm{Col}(X'X) \subset \mathrm{Col}(X')$, so the well-known relation $\mathrm{rank}\,(X'X) = \mathrm{rank}\,(X) = \mathrm{rank}\,(X')$ completes the proof. □

§3 Review of Linear Algebra Cont'd (2020/2/28)

Let V be a linear subspace of $\mathbb{R}^n$. The **orthogonal projection** of $x \in \mathbb{R}^n$ onto V is

$$\mathrm{proj}_V(x) = \underset{v\in V}{\arg\min} \|x - v\|\,,$$

where $\|\bullet\| = \sqrt{\bullet'\bullet}$ denotes the Euclidean norm. One can see that

$$\|x - (\mathrm{proj}_V(x) + u)\|^2 - \|x - \mathrm{proj}_V(x)\|^2 = \|u\|^2 - 2(x - \mathrm{proj}_V(x))'u \ge 0, \quad \forall u \in V$$

entails $x - \mathrm{proj}_V(x) \in V^\perp$ and therefore proj_V is a linear transform uniquely defined via $\mathbb{R}^n = V \bigoplus V^\perp$. Note that $\mathrm{proj}_{\mathrm{span}(v)}$ is often abbreviated as proj_v for a single $v \in \mathbb{R}^n$. It's clear that

$$\mathrm{proj}_v \bullet = \left\langle \bullet, \tfrac{v}{\|v\|} \right\rangle \tfrac{v}{\|v\|} = \tfrac{v}{\|v\|}\tfrac{v'}{\|v\|}\bullet\,.$$

If $v_1, \ldots, v_k$ are orthogonal, then $\mathrm{proj}_{\mathrm{span}(v_1,\ldots,v_k)} = \sum_{i=1}^{k} \mathrm{proj}_{v_i}$.

The **orthogonalization** for $X = (x_1, \ldots, x_p) \in \mathbb{R}^{n\times p}$ of rank r aims to find $Z = (z_1, \ldots, z_r) \in \mathbb{R}^{n\times r}$ with orthogonal columns such that $\mathrm{span}\,(z_1, z_2, \ldots, z_k) = \mathrm{span}\left(x_1, x_2, \ldots, x_{\varsigma(k)}\right)$ for $1 \le k \le r$, where $\varsigma(k) = \min\{j : x_j \notin \mathrm{span}\left(x_1, x_2, \ldots, x_{\varsigma(k-1)}\right)\}$ are recursively determined from $\varsigma(0) := 0$. The so called **Gram-Schmidt process** works. To be explicit, let $z_1 = x_{\varsigma(1)}$, and

$$z_k = x_{\varsigma(k)} - \mathrm{proj}_{\mathrm{span}(z_1, z_2, \ldots, z_{k-1})}\, x_{\varsigma(k)}, \quad k = 2, 3, \cdots, r.$$

It follows immediately that $x_k = \sum_{i=1}^{\min(k,r)} \mathrm{proj}_{z_i}\, x_k$ for $1 \le k \le p$, which gives rise to $X = Z\Gamma$, where $\Gamma \in \mathbb{R}^{r\times p}$ has $\gamma_{ij} = z_i'x_j/\|z_i\|^2$ as its (i, j)-entry. Note that $\gamma_{ij} = 0$ for $i > j$, so Γ is upper-triangular.

[ii] cf. https://zhuanlan.zhihu.com/p/75283604

Putting $D = (Z'Z)^{1/2} = \operatorname{diag}(\|z_1\|, \dots, \|z_r\|)$, the **QR decomposition** of X can be $X = QR$, where $Q = ZD^{-1} \in \mathbb{R}^{n\times r}$ has orthonormal columns and $R = D\Gamma \in \mathbb{R}^{r\times p}$ is an upper-triangular matrix. Note that $P_X = X(X'X)^- X' = \operatorname{proj}_{\mathrm{Col}(X)} = \operatorname{proj}_{\mathrm{Col}(Q)} = QQ'$. Write $Q = (q_1, \dots, q_r) = (z_1/\|z_1\|, \dots, z_r/\|z_r\|)$, then

$$P_X y = \sum q_k q_k' y = \sum \langle y, q_k\rangle\, q_k, \quad \forall y \in \mathbb{R}^n.$$

The above results hold generally in a separable Hilbert space, where we may cut off a pre-specified basis.

As a matter of fact, statistical modeling often derives orthogonal decomposition. Note that, however, the projection can be expressed by *any* orthonormal basis. Using the SVD $X = U\Sigma V'$, one can see that $(X'X)^+ = V\Sigma^{-2}V'$ and then $P_X = X(X'X)^+X' = UU'$.

§4 Noncentral Chi-square Distribution (2020/3/4)

Recall that $X \sim \mathcal{N}_p(\mu, \Sigma)$ has

$$M_X(t) = \mathbb{E}\exp(t'X) = \exp(t'\mu + \tfrac{1}{2}t'\Sigma t), \quad t \in \mathbb{R}^p$$

as its *moment generating function* (m.g.f.), and

$$\phi_X(t) = \mathbb{E}\exp(\sqrt{-1}t'X) = \exp(\sqrt{-1}t'\mu - \tfrac{1}{2}t'\Sigma t), \quad t \in \mathbb{R}^p$$

as its *characteristic function* (ch.f.), which corresponds to the distribution uniquely by the *inversion formula*.

✍ It follows that $AX + b \sim \mathcal{N}_q(A\mu + b, A\Sigma A')$ for any $A \in \mathbb{R}^{q\times p}$ and $b \in \mathbb{R}^q$.

✍ Moreover, $X \sim \mathcal{N}_p(\mu, \Sigma)$ if and only if $a'X \sim \mathcal{N}(a'\mu, a'\Sigma a)$ for all $a \in \mathbb{R}^p$.

✍ Besides, $Y_1 = A_1X + b_1$ and $Y_2 = A_2X + b_2$ are independent if and only if $\operatorname{Cov}(Y_1, Y_2) = A_1\Sigma A_2'$ vanishes.

Given $X \sim \mathcal{N}_p(\mu, I_p)$, we say that $X'X$ obeys the **noncentral chi-square distribution** with p degrees of freedom and non-centrality parameter $\lambda = \mu'\mu$, denoted by $\chi_p^2(\lambda)$. The *probability density function* (p.d.f.) is

$$f(u; p, \lambda) = \sum_{k=0}^{\infty} \frac{\mathrm{e}^{-\lambda/2}(\lambda/2)^k}{k!} f(u; p+2k, 0), \quad f(u; r, 0) = \frac{u^{r/2-1}\mathrm{e}^{-\frac{u}{2}}}{2^{r/2}\Gamma(r/2)}\mathbb{1}_{[u>0]},$$

a Poisson$(\frac{\lambda}{2})$-weighted mixture of central chi-square densities. The m.g.f./ch.f. of $\chi_p^2(\lambda)$ is given by

$$M(t; p, \lambda) = \mathbb{E}\exp(tX'X) = \frac{1}{(1-2t)^{p/2}}\exp\left(\frac{\lambda t}{1-2t}\right), \quad t \in \mathbb{C},\ \operatorname{Re} t < \frac{1}{2}.$$

✍ If $X \sim \mathcal{N}_p(\mu, \Sigma)$ is non-degenerate, then $(X-\mu)'\Sigma^{-1}(X-\mu) \sim \chi_p^2$ and $X'\Sigma^{-1}X \sim \chi_p^2(\mu'\Sigma^{-1}\mu)$.

✍ For independent $\xi \sim \chi_p^2(\lambda)$ and $\eta \sim \chi_q^2(\nu)$, we have $\xi + \eta \sim \chi_{p+q}^2(\lambda + \nu)$.

✍ If $X \sim \mathcal{N}_p(\mu, I_p)$ and $A \in \mathbb{R}^{p\times p}$ is symmetric, then

$$X'AX \sim \chi_r^2(\lambda) \text{ with } \lambda = \mu'A\mu \text{ if and only if } A \text{ is an idempotent matrix of rank } r.$$

✍ If $X \sim \mathcal{N}_p(\mu, \Sigma)$ is non-degenerate, $A \in \mathbb{R}^{p\times p}$ is symmetric and $B \in \mathbb{R}^{q\times p}$, then

$$X'AX \text{ and } BX \text{ are independent if and only if } B\Sigma A = 0_{q\times p}.$$

✍ If $X \sim \mathcal{N}_p(\mu, \Sigma)$ is non-degenerate and $A, B \in \mathbb{R}^{p\times p}$ are symmetric, then

$$X'AX \text{ and } X'BX \text{ are independent if and only if } A\Sigma B = 0_{p\times p}.$$

The nontrivial proofs[iii] are left as exercise 7 in homework 1.

The following **Fisher-Cochran theorem** plays an important role in statistical inference.

Suppose $X \sim \mathcal{N}_p(\mu, I_p)$, and $Q_1, \dots, Q_k$ are quadratic forms in X such that $X'X = Q_1 + \cdots + Q_k$, i.e., $Q_j = X'A_jX$ for symmetric $A_j \in \mathbb{R}^{p\times p}$, $j = 1, \cdots, k$, such that $I_p = A_1 + \cdots + A_k$. Denote $\operatorname{rank}(A_j) = r_j$. A sufficient and necessary condition for $Q_j \sim \chi_{r_j}^2(\lambda_j)$ and Q_j's are independent is that $p = r_1 + \cdots + r_k$, in which case $\lambda_j = \mu'A_j\mu$. Indeed, A_j's should be projections onto mutually orthogonal subspaces.

If $Q_1 \sim \chi_m^2(\lambda)$ and $Q_2 \sim \chi_n^2$ are independent, then $\frac{Q_1/m}{Q_2/n} \sim F_{m,n}(\lambda)$, the **noncentral *F*-distribution**.

[iii] cf. https://zhuanlan.zhihu.com/p/85314322

§5 Prediction Accuracy (2020/3/11)

Given an input X and an output Y, statistical procedures concerning prediction attempt to find a function $f : \mathcal{X} \to \mathcal{Y}$ so that the **risk**[iv)]

$$\mathbb{E}[L(f(X), Y)] = \mathbb{E}[\mathbb{E}[L(f(X), Y)|X]]$$

is minimized, where $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ is the **loss function** quantifying accuracy. Whether the design is fixed or random, we may set the *optimal predictor* conditional on $x \in \mathcal{X}$ to be

$$\hat{f}(x) = \underset{f(x)\in\mathcal{Y}}{\arg\min}\, \mathbb{E}[L(f(x), Y)|X = x]$$

- Suppose that the response variable is numerical, then the quadratic loss

  $$L(\hat{y}, y) = |\hat{y} - y|^2$$

  is commonly used. The **mean squared error** (MSE) decomposes into

  $$\mathbb{E}[L(f(X), Y)|X] = \mathbb{E}\big[(Y - f(X))^2\big|X\big] = \mathbb{E}\big[(Y - \mathbb{E}[Y|X])^2\big|X\big] + \big(\mathbb{E}[Y|X] - f(X)\big)^2.$$

  Hence, the best predictor is

  $$\hat{f}(x) = \mathbb{E}[Y|X = x].$$

  – In parametric models, $\mathbb{E}[Y|X]$ is determined by a finite-dimensional parameter. For example, the linear model is $Y = x'\beta + \varepsilon$, where $\beta \in \mathbb{R}^k$, and the error ε is assumed $\mathbb{E}[\varepsilon|x] = 0$.

  – In nonparametric models, $\mathbb{E}[Y|X]$ involves infinite-dimensional parameters. For example, the predictor in nonparametric regression does not take a predetermined form but is constructed according to information derived from the data, say, $\mathbb{E}[Y|X = x] = m(x)$ where $m(\cdot)$ is assumed minimal conditions such as smoothness.

- Suppose that the response variable is categorical, then the zero-one loss

  $$L(\hat{y}, y) = \mathbb{1}_{[\hat{y}\neq y]}$$

  is commonly used. Keep the classification problem in mind. Minimization of

  $$\mathbb{E}[L(g(x), Y)|X = x] = \sum_{j\neq g(x)} \mathbb{P}(Y = j|X = x) = 1 - \mathbb{P}(Y = g(x)|X = x)$$

  yields the **naïve Bayesian classifier**

  $$\hat{g}(x) = \underset{j\in\mathcal{Y}}{\arg\max}\, \mathbb{P}(Y = j|X = x),$$

  which takes the most likely class given $X = x$ and is applicable to prior information. Also, we may treat Y as a vector of *dummy variables*

  $$Y_j = \mathbb{1}_{[y=j]}, \quad j \in \mathcal{Y},$$

  which add up to 1 and thus the degree of freedom is $\#\mathcal{Y}-1$. Henceforth, $\hat{g}(x) = \arg\max_j \mathbb{E}[Y_j|X = x]$.

Nonparametrically, the estimation of $\mathbb{E}[Y|X]$ can be based on the **nearest neighbor**. Given a training dataset $(x_i, y_i)_{1\leq i\leq n}$, one may use data around x to approximate $m(x) = \mathbb{E}[Y|X = x]$. Ideally,

$$\hat{m}(x) = \frac{1}{\#N(x)} \sum_{i\in N(x)} y_i,$$

where $N(x) = \{i : x_i = x\}$. Let $N_k(x)$ be the index set that marks the closest k inputs to x, then

$$\hat{m}_k(x) = \frac{1}{k} \sum_{i\in N_k(x)} y_i$$

is called the ***k*-NN estimator**. We may tune the effective sample size k to enhance the estimation quality. An extremely small k results in little data reduction, and hereby the variance of $\hat{m}_k$ tends to explode even though the bias may vanish. The case when k is too large is often accompanied with an overwhelming bias.

[iv)] cf. http://stat.wharton.upenn.edu/~skakade/courses/stat928/lectures/lecture0.pdf

§6 Curse of Dimensionality & Estimability in Linear Models (2020/3/13)

To guarantee the consistency of an estimator, the effective sample size $k = k(n)$ should increase to infinity as the sample size $n \to \infty$. Usually, $k(n) \sim n^\alpha$ for some constant $\alpha \in (0,1)$. At this point, the dimension p of the covariates is crucial. In classical setting, $p \ll n$; but in high-dimensional case, $p \gg n \gg q$, where q denotes the number of true signals. The *curse of dimensionality* turns out to be more serious when the models considered are nonparametric, compared with parametric cases. Denote the density function by $f(\cdot)$ and the bandwidth by $h = h(n) \in (0,1)$, then $k \sim n \int_{N(x;h)} f(u)\,\mathrm{d}u \sim nh^p$ decays polynomially with h and exponentially with p. Thus, a large p incurs expensive computation. For example, the unit ball in $\mathbb{R}^p$ almost concentrates near its surface as $p \to \infty$. See also exercise 6 in homework 1.

Consider the linear model

$$\underset{(n\times 1)}{Y} = \underset{(n\times p)}{X}\underset{(p\times 1)}{\beta} + \underset{(n\times 1)}{\varepsilon}, \qquad \mathbb{E}[\varepsilon] = 0_n,\ \mathrm{Var}(\varepsilon) = \sigma^2 I_n.$$

The first issue is *model identifiablity* —— the estimability of parameters. Note that the solution $\hat\beta$ to the **normal equation**

$$X'X\hat\beta = X'Y,$$

i.e., the **least squares estimator**

$$\hat\beta = (X'X)^- X'Y,$$

depends on the generalized inverse $(X'X)^-$ and is possibly not unique. For example, in one-way ANOVA

$$Y_{\ell r} = \mu + \alpha_\ell + \varepsilon_{\ell r}, \quad 1 \le r \le n_\ell,\ 1 \le \ell \le g,$$

$(\mu, \alpha_1, \dots, \alpha_g)'$ is not estimable, but the *contrasts*[v] such as $\alpha_1 - \alpha_g$ may be estimable. As a matter of fact, the estimators $\hat\alpha_\ell = \overline{Y}_{\ell\cdot} - a$ yield $\hat\alpha_1 - \hat\alpha_g = \overline{Y}_{1\cdot} - \overline{Y}_{g\cdot}$ whichever value a takes.

In many applications, we are interested in estimating some linear functions $\theta = c'\beta$ of β, where $c \in \mathbb{R}^p$. Recall that a parameter ϑ is said to be **estimable** if and only if there exists an *unbiased* estimator of ϑ. We say that $c'\beta$ is *linearly estimable* if, there exists some $l \in \mathbb{R}^n$ such that $\mathbb{E}[l'Y] = c'\beta,\ \forall \beta \in \mathbb{R}^p$. Since $\mathbb{E}[Y] = X\beta$, the linear estimablity of $c'\beta$ is equivalent to $c = X'l \in \mathrm{Col}(X')$.

Theorem. (1) If $c'\hat\beta$ is unique, then $c \in \mathrm{Col}(X'X) = \mathrm{Col}(X')$.
(2) If $c \in \mathrm{Col}(X')$, then $c'\hat\beta$ is unique and unbiased for $c'\beta$.
(3) If $c'\beta$ is estimable and $\varepsilon \sim \mathcal{N}_n(0_n, \sigma^2 I_n)$, then $c \in \mathrm{Col}(X')$.

Proof. (1) Let $b \in \mathrm{Col}(X'X)^\perp$, then

$$X'Y = X'X\hat\beta = X'X(\hat\beta + b) \implies c'(\hat\beta + b) = c'\hat\beta \implies c \perp b.$$

(2) Suppose that $c = X'l$ for some $l \in \mathbb{R}^n$, then

$$c'\hat\beta = l'X(X'X)^- X'Y = l'P_X Y$$

is unique, and

$$\mathbb{E}[c'\hat\beta] = l'P_X\mathbb{E}Y = l'P_X X\beta = l'X\beta = c'\beta.$$

(3) If there is an estimator $T(Y, X)$ unbiased for $c'\beta$, then

$$c'\beta = \int_{\mathbb{R}^n} T(y,X)\frac{1}{(2\pi\sigma^2)^{n/2}}\exp\left\{-\frac{1}{2\sigma^2}\|y - X\beta\|^2\right\}\mathrm{d}y.$$

Differentiation with respect to β yields

$$c = X'\int_{\mathbb{R}^n} T(y,X)\frac{y - X\beta}{(2\pi\sigma^2)^{n/2}\sigma^2}\exp\left\{-\frac{1}{2\sigma^2}\|y - X\beta\|^2\right\}\mathrm{d}y,$$

and thus $c \in \mathrm{Col}(X')$. □

In summary, $A\beta$ with $A \in \mathbb{R}^{q\times p}$ is estimable iff $\mathrm{Col}(A') \subset \mathrm{Col}(X')$, i.e., $A = A_* X$ for some $A_* \in \mathbb{R}^{q\times n}$. Particularly, β is estimable if and only if X has full column rank.

[v] cf. https://stats.stackexchange.com/a/221861

§7 Properties of Ordinary Least Squares (2020/3/18)

Let $\hat{\beta} = (X'X)^- X'Y$ be a choice of the ordinary least squares (OLS) estimator for β. A useful fact is that for any estimable $A\beta$ and $B\beta$, where A and B are compatible constant matrices, we have

$$\mathrm{Cov}(A\hat{\beta}, B\hat{\beta}) = \sigma^2 A(X'X)^- B', \quad \& \quad \mathrm{Var}(A\hat{\beta}) = \sigma^2 A(X'X)^- A'.$$

To see this, we may write $A = A_* X$ and $B = B_* X$. Since the vector of **fitted values**

$$\hat{Y} := X\hat{\beta} = X(X'X)^- X'Y = P_X Y$$

has variance-covariance matrix

$$\mathrm{Var}(\hat{Y}) = P_X \,\mathrm{Var}(Y) P_X' = P_X(\sigma^2 I_n) P_X = \sigma^2 P_X,$$

it follows that $\mathrm{Cov}(A_*\hat{Y}, B_*\hat{Y}) = A_* \,\mathrm{Var}(\hat{Y}) B_*' = \sigma^2 A_* P_X B_*'$.

Theorem (Gauss-Markov)**.** If $c'\beta$ is estimable, where $c \in \mathbb{R}^p$, then $c'\hat{\beta}$ has the minimum variance among all linear unbiased estimators. Thus, the least squares estimator is the *best linear unbiased estimator* (BLUE).

Proof. Let $l'Y$ be an unbiased estimator of $c'\beta$, where $l \in \mathbb{R}^n$. It must hold that $X'l = c$, so $\underline{c'\hat{\beta} = l'\hat{Y}}$. Therefore,

$$\mathrm{Var}(l'Y) - \mathrm{Var}(c'\hat{\beta}) = l'[\mathrm{Var}(Y) - \mathrm{Var}(\hat{Y})]l = \sigma^2 l'(I_n - P_X)l \geq 0,$$

where $I_n - P_X = \mathrm{proj}_{\mathrm{Col}(X)^\perp}$ is positive semi-definite. □

Sometimes the projection matrix P_X is also called **hat matrix**, denoted by H, as it "puts a hat on Y".

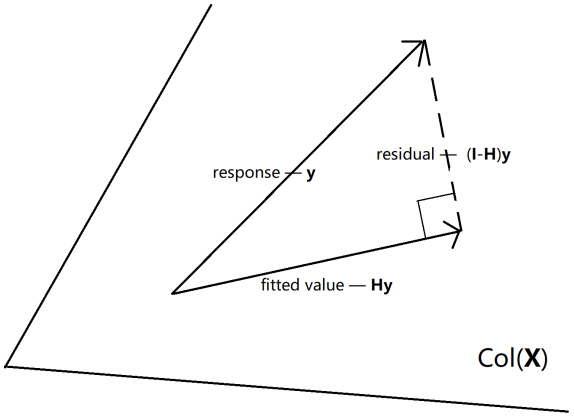

Figure 1: Orthogonal Projection in Linear Regression Models

The **residual** vector is

$$\hat{\varepsilon} := Y - \hat{Y} = (I_n - P_X)Y,$$

which falls into $\mathrm{Col}(X)^\perp$ and satisfies that

$$\mathbb{E}[\hat{\varepsilon}] = (I_n - P_X)X\beta = 0_n, \quad \mathrm{Var}(\hat{\varepsilon}) = \mathbb{E}[\hat{\varepsilon}\hat{\varepsilon}'] = (I_n - P_X)(\sigma^2 I_n)(I_n - P_X)' = \sigma^2(I_n - P_X),$$

and

$$\mathrm{Cov}(\hat{\varepsilon}, \hat{Y}) = (I_n - P_X)(\sigma^2 I_n) P_X' = \sigma^2(I_n - P_X)P_X = 0_{n\times n}.$$

The **residual sum of squares**, denoted by

$$R_0^2 := \|\hat{\varepsilon}\|^2 = \hat{\varepsilon}'\hat{\varepsilon} = Y'(I_n - P_X)Y,$$

has mean

$$\mathbb{E}[R_0^2] = \mathbb{E}\,\mathrm{tr}\,(\hat{\varepsilon}\hat{\varepsilon}') = \mathrm{tr}\,(\mathbb{E}[\hat{\varepsilon}\hat{\varepsilon}']) = \mathrm{tr}\,(I_n - P_X)\,\sigma^2 = [n - \mathrm{rank}\,(X)]\sigma^2.$$

This gives $\hat{\sigma}^2 := R_0^2/[n - \mathrm{rank}\,(X)]$ as a method of moments estimator for σ^2.

§8 Restrictions on a Subset of Regression Coefficients (2020/3/25)

In order to tackle the problem of testing a general linear hypothesis, let's begin with a simpler restriction. Suppose that the full model is

$$\underset{(n\times 1)}{\boldsymbol{Y}} = \underset{(n\times p)}{\boldsymbol{X}}\underset{(p\times 1)}{\boldsymbol{\beta}} + \underset{(n\times 1)}{\boldsymbol{\varepsilon}}, \qquad \mathbb{E}[\boldsymbol{\varepsilon}] = \mathbf{0}_n,\ \mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2\boldsymbol{I}_n,\ \mathrm{rank}\,(\boldsymbol{X}) = r.$$

Write $\boldsymbol{X} = (\underset{n\times(p-s)}{\boldsymbol{X}_1}, \underset{n\times s}{\boldsymbol{X}_2})$ and $\boldsymbol{\beta} = \begin{pmatrix}\boldsymbol{\beta}_1\\ \boldsymbol{\beta}_2\end{pmatrix}$, where $\boldsymbol{\beta}_1 \in \mathbb{R}^{p-s}$ and $\boldsymbol{\beta}_2 \in \mathbb{R}^s$. The hypothesis of interest is

$$H_0 : \boldsymbol{\beta}_2 = \boldsymbol{\beta}_2^* \quad \text{vs.} \quad H_1 : \boldsymbol{\beta}_2 \neq \boldsymbol{\beta}_2^*$$

where $\boldsymbol{\beta}_2^*$ is known, e.g., $\mathbf{0}_s$. It should be required that $\boldsymbol{\beta}_2$ is estimable, which implies

$$\mathrm{rank}\,(\boldsymbol{X}_2) = s,\ \mathrm{rank}\,(\boldsymbol{X}_1) = r - s,\ \text{and}\ \mathrm{Col}(\boldsymbol{X}_1) \cap \mathrm{Col}(\boldsymbol{X}_2) = \{\mathbf{0}_n\}.$$

Proof. There exists some $\boldsymbol{C} \in \mathbb{R}^{s\times n}$ such that $(\mathbf{0}_{s\times(p-s)}, \boldsymbol{I}_s) = \boldsymbol{C}\boldsymbol{X} = (\boldsymbol{C}\boldsymbol{X}_1, \boldsymbol{C}\boldsymbol{X}_2)$. Consequently, if $\boldsymbol{X}_1\boldsymbol{b}_1 = \boldsymbol{X}_2\boldsymbol{b}_2$, then $\boldsymbol{b}_2 = \boldsymbol{C}\boldsymbol{X}_2\boldsymbol{b}_2 = \boldsymbol{C}\boldsymbol{X}_1\boldsymbol{b}_1 = \mathbf{0}_s$. □

Under H_0, the restricted model becomes

$$\boldsymbol{Y} - \boldsymbol{X}_2\boldsymbol{\beta}_2^* = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}.$$

Then the restricted normal equation is

$$\boldsymbol{X}_1'\boldsymbol{X}_1\tilde{\boldsymbol{\beta}}_1 = \boldsymbol{X}_1'(\boldsymbol{Y} - \boldsymbol{X}_2\boldsymbol{\beta}_2^*),$$

and any restricted least squares estimator $\tilde{\boldsymbol{\beta}}_1$ gives

$$\boldsymbol{X}_1\tilde{\boldsymbol{\beta}}_1 = \boldsymbol{P}_{\boldsymbol{X}_1}(\boldsymbol{Y} - \boldsymbol{X}_2\boldsymbol{\beta}_2^*).$$

Note that $\mathrm{Col}(\boldsymbol{X}_1) \subset \mathrm{Col}(\boldsymbol{X})$ implies $\boldsymbol{P}_{\boldsymbol{X}_1}\boldsymbol{P}_{\boldsymbol{X}} = \boldsymbol{P}_{\boldsymbol{X}_1}$. With

$$\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{Y} = \hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1 + \boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2,$$

one can see that[vi)]

$$\boldsymbol{X}_1\tilde{\boldsymbol{\beta}}_1 = \boldsymbol{P}_{\boldsymbol{X}_1}(\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{Y} - \boldsymbol{X}_2\boldsymbol{\beta}_2^*) = \boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1 + \boldsymbol{P}_{\boldsymbol{X}_1}\boldsymbol{X}_2(\hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2^*).$$

Let $\tilde{\boldsymbol{Y}} = \boldsymbol{X}_1\tilde{\boldsymbol{\beta}}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2^*$ be fitted values of the restricted model. We have

$$\hat{\boldsymbol{Y}} - \tilde{\boldsymbol{Y}} = \boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2 - \boldsymbol{P}_{\boldsymbol{X}_1}\boldsymbol{X}_2(\hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2^*) - \boldsymbol{X}_2\boldsymbol{\beta}_2^* = (\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}_1})\boldsymbol{X}_2(\hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2^*).$$

Theorem. The column space of $\boldsymbol{Z}_2 := (\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}_1})\boldsymbol{X}_2 = \boldsymbol{X}_2 - \boldsymbol{P}_{\boldsymbol{X}_1}\boldsymbol{X}_2$ is $\mathrm{Col}(\boldsymbol{X}_1)^\perp \cap \mathrm{Col}(\boldsymbol{X})$.

Corollary. It follows that $\boldsymbol{P}_{\boldsymbol{Z}_2} = \boldsymbol{P}_{\boldsymbol{X}} - \boldsymbol{P}_{\boldsymbol{X}_1}$.

Proof. Clearly $\mathrm{Col}(\boldsymbol{Z}_2) \subset \mathrm{Col}(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}_1}) = \mathrm{Col}(\boldsymbol{X}_1)^\perp$. Since $\mathrm{Col}(\boldsymbol{P}_{\boldsymbol{X}_1}\boldsymbol{X}_2) \subset \mathrm{Col}(\boldsymbol{X}_1)$, it also holds that $\mathrm{Col}(\boldsymbol{Z}_2) \subset \mathrm{Col}(\boldsymbol{X}_2) + \mathrm{Col}(\boldsymbol{X}_1) = \mathrm{Col}(\boldsymbol{X})$. As for the converse, if $\boldsymbol{x} = \boldsymbol{X}_1\boldsymbol{b}_1 + \boldsymbol{X}_2\boldsymbol{b}_2 \in \mathrm{Col}(\boldsymbol{X})$ and $\boldsymbol{x} \perp \mathrm{Col}(\boldsymbol{X}_1)$, then $\boldsymbol{x} = (\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}_1})\boldsymbol{x} = (\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}_1})\boldsymbol{X}_2\boldsymbol{b}_2 \in \mathrm{Col}(\boldsymbol{Z}_2)$. □

Furthermore,

$$\begin{aligned}\hat{\boldsymbol{Y}} - \tilde{\boldsymbol{Y}} &= (\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}_1})[\boldsymbol{X}_2(\hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2^*) + \boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1]\\ &= (\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}_1})(\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{Y} - \boldsymbol{X}_2\boldsymbol{\beta}_2^*)\\ &= (\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}_1})\boldsymbol{P}_{\boldsymbol{X}}(\boldsymbol{Y} - \boldsymbol{X}_2\boldsymbol{\beta}_2^*) \quad = \boldsymbol{P}_{\boldsymbol{Z}_2}(\boldsymbol{Y} - \boldsymbol{X}_2\boldsymbol{\beta}_2^*).\end{aligned}$$

In view of $\mathbb{R}^n = \mathrm{Col}(\boldsymbol{X})^\perp \bigoplus \mathrm{Col}(\boldsymbol{X})$, the restricted residuals can be written as

$$\boldsymbol{Y} - \tilde{\boldsymbol{Y}} = (\boldsymbol{Y} - \hat{\boldsymbol{Y}}) + (\hat{\boldsymbol{Y}} - \tilde{\boldsymbol{Y}}),$$

[vi)] orthogonal designs may be encouraged since $\boldsymbol{P}_{\boldsymbol{X}_1}\boldsymbol{X}_2 = \mathbf{0}$ if and only if $\boldsymbol{X}_1'\boldsymbol{X}_2 = \mathbf{0}$, i.e., $\mathrm{Col}(\boldsymbol{X}_1) \perp \mathrm{Col}(\boldsymbol{X}_2)$.

whose sum of squares is

$$R_1^2 := \|\boldsymbol{Y} - \tilde{\boldsymbol{Y}}\|^2 = \|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\|^2 + \|\hat{\boldsymbol{Y}} - \tilde{\boldsymbol{Y}}\|^2.$$

Recall that

$$R_0^2 = \|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\|^2 = \|(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{Y}\|^2 = \|(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}})(\boldsymbol{Y} - \boldsymbol{X}_2\boldsymbol{\beta}_2^*)\|^2.$$

The difference

$$R_1^2 - R_0^2 = \|\hat{\boldsymbol{Y}} - \tilde{\boldsymbol{Y}}\|^2 = \|\boldsymbol{Z}_2(\hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2^*)\|^2 = \|\boldsymbol{P}_{\boldsymbol{Z}_2}(\boldsymbol{Y} - \boldsymbol{X}_2\boldsymbol{\beta}_2^*)\|^2$$

plays an important role in statistical inference. From the perspective of Cochran's theorem (see §4), the orthogonal decomposition

$$\boldsymbol{I}_n = \boldsymbol{P}_{\boldsymbol{X}_1} + \boldsymbol{P}_{\boldsymbol{Z}_2} + (\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}})$$

applies to Gaussian errors, in which case

$$\boldsymbol{Y} - \boldsymbol{X}_2\boldsymbol{\beta}_2^* \sim \mathcal{N}_n(\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{X}_2\boldsymbol{\beta}_2^*, \sigma^2\boldsymbol{I}_n).$$

Indeed, $R_1^2 - R_0^2$ could be connected with the *Wald statistic*[vii].

Proof. Due to the estimability of $\boldsymbol{\beta}_2$, we have $(\boldsymbol{0}_{s\times(p-s)}, \boldsymbol{I}_s) = \boldsymbol{C}\boldsymbol{X}$ for some $\boldsymbol{C} \in \mathbb{R}^{s\times n}$. It follows that

$$\boldsymbol{C}\boldsymbol{P}_{\boldsymbol{X}_1} = \boldsymbol{C}\boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{X}_1)^-\boldsymbol{X}_1' = \boldsymbol{0}_{s\times n},$$

and

$$\boldsymbol{C}\boldsymbol{Z}_2 = \boldsymbol{C}(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}_1})\boldsymbol{X}_2 = \boldsymbol{C}\boldsymbol{X}_2 = \boldsymbol{I}_s.$$

Hence, $\boldsymbol{Z}_2 \in \mathbb{R}^{n\times s}$ has full column rank. The variance-covariance matrix of $\hat{\boldsymbol{\beta}}_2 = \boldsymbol{C}\boldsymbol{X}\hat{\boldsymbol{\beta}}$ is

$$\begin{aligned}\mathrm{Var}(\hat{\boldsymbol{\beta}}_2) &= \sigma^2\boldsymbol{C}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{X}'\boldsymbol{C}' \\ &= \sigma^2\boldsymbol{C}(\boldsymbol{P}_{\boldsymbol{X}_1} + \boldsymbol{P}_{\boldsymbol{Z}_2})\boldsymbol{C}' \\ &= \sigma^2(\boldsymbol{C}\boldsymbol{Z}_2)(\boldsymbol{Z}_2'\boldsymbol{Z}_2)^{-1}(\boldsymbol{C}\boldsymbol{Z}_2)' \quad = \sigma^2(\boldsymbol{Z}_2'\boldsymbol{Z}_2)^{-1}.\end{aligned}$$

Therefore, $\sigma^2(\hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2^*)'\,\mathrm{Var}(\hat{\boldsymbol{\beta}}_2)^{-1}(\hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2^*) = \|\boldsymbol{Z}_2(\hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2^*)\|^2 = R_1^2 - R_0^2$. □

§9 Inference for Linear Regression Models (2020/3/27)

Let $\boldsymbol{H} = (\boldsymbol{h}_1, \ldots, \boldsymbol{h}_s) \in \mathbb{R}^{p\times s}$ and $\boldsymbol{\xi} \in \mathbb{R}^s$ be given. We are interested in *the general linear hypothesis*

$$H_0 : \boldsymbol{H}'\boldsymbol{\beta} = \boldsymbol{\xi},$$

where $\mathrm{Col}(\boldsymbol{H}) \subset \mathrm{Col}(\boldsymbol{X}')$, so $\boldsymbol{H}'\boldsymbol{\beta}$ is estimable. Assume that the matrix $\boldsymbol{H}$ is of full column rank to avoid redundancies. Hereby,

$$s = \mathrm{rank}\,(\boldsymbol{H}) \le \mathrm{rank}\,(\boldsymbol{X}) = r \le p.$$

In order to invoke the conclusions in §8, pick some $\boldsymbol{A} \in \mathbb{R}^{p\times(p-s)}$ such that $\mathrm{Col}(\boldsymbol{A}) = \mathrm{Col}(\boldsymbol{H})^\perp$, and put

$$\boldsymbol{\theta} = \begin{pmatrix}\boldsymbol{\theta}_1 \\ \boldsymbol{\theta}_2\end{pmatrix} = \begin{pmatrix}\boldsymbol{A}'\boldsymbol{\beta} \\ \boldsymbol{H}'\boldsymbol{\beta}\end{pmatrix}, \quad \& \quad \widetilde{\boldsymbol{X}} = \boldsymbol{X}\begin{pmatrix}\boldsymbol{A}' \\ \boldsymbol{H}'\end{pmatrix}^{-1} = (\underset{n\times(p-s)}{\widetilde{\boldsymbol{X}}_1}, \underset{n\times s}{\widetilde{\boldsymbol{X}}_2}).$$

Then

$$\mathbb{E}[\boldsymbol{Y}] = \boldsymbol{X}\boldsymbol{\beta} = \widetilde{\boldsymbol{X}}\boldsymbol{\theta} = \widetilde{\boldsymbol{X}}_1\boldsymbol{\theta}_1 + \widetilde{\boldsymbol{X}}_2\boldsymbol{\theta}_2.$$

The column space of $\widetilde{\boldsymbol{X}}' = (\boldsymbol{A}, \boldsymbol{H})^{-1}\boldsymbol{X}'$ contains the last s columns of $((\boldsymbol{A}, \boldsymbol{H})^{-1}\boldsymbol{A}, (\boldsymbol{A}, \boldsymbol{H})^{-1}\boldsymbol{H}) = \boldsymbol{I}_p$, and thus $\boldsymbol{\theta}_2 = (\boldsymbol{0}_{s\times(p-s)}, \boldsymbol{I}_s)\boldsymbol{\theta}$ is estimable in the *reparametrized* model $\boldsymbol{Y} = \widetilde{\boldsymbol{X}}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$. Henceforth, nothing is fresh. Let $\widehat{\boldsymbol{\theta}}$ be a solution to the normal equation $\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}}\widehat{\boldsymbol{\theta}} = \widetilde{\boldsymbol{X}}'\boldsymbol{Y}$, and

$$\widetilde{\boldsymbol{Y}} = \widetilde{\boldsymbol{X}}_1\widehat{\boldsymbol{\theta}}_1 + \boldsymbol{P}_{\widetilde{\boldsymbol{X}}_1}\widetilde{\boldsymbol{X}}_2(\widehat{\boldsymbol{\theta}}_2 - \boldsymbol{\xi}) + \widetilde{\boldsymbol{X}}_2\boldsymbol{\xi}$$

be fitted values of the restricted model. It can be seen that

$$\|\boldsymbol{Y} - \widetilde{\boldsymbol{Y}}\|^2 - \|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\|^2 = \|\hat{\boldsymbol{Y}} - \widetilde{\boldsymbol{Y}}\|^2 = \sigma^2(\widehat{\boldsymbol{\theta}}_2 - \boldsymbol{\xi})'\,\mathrm{Var}(\widehat{\boldsymbol{\theta}}_2)^{-1}(\widehat{\boldsymbol{\theta}}_2 - \boldsymbol{\xi}),$$

where $\widehat{\boldsymbol{\theta}}_2 = \boldsymbol{H}'\hat{\boldsymbol{\beta}}$ and $\mathrm{Var}(\widehat{\boldsymbol{\theta}}_2) = \sigma^2\boldsymbol{H}'(\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{H}$.

[vii] to test $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$, one may employ $W = (\hat{\boldsymbol{\theta}}^{\mathrm{MLE}} - \boldsymbol{\theta}_0)'\widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}}^{\mathrm{MLE}})^{-1}(\hat{\boldsymbol{\theta}}^{\mathrm{MLE}} - \boldsymbol{\theta}_0)$ as the test statistic, where $\hat{\boldsymbol{\theta}}^{\mathrm{MLE}}$ is the maximum likelihood estimator of $\boldsymbol{\theta}$. See also https://statlect.com/fundamentals-of-statistics/Wald-test.

Moreover,

$$\begin{aligned}\mathbb{E}\|\hat{\boldsymbol{Y}}-\widetilde{\boldsymbol{Y}}\|^2/\sigma^2 &= \mathbb{E}\,\mathrm{tr}\left(\mathrm{Var}(\widehat{\boldsymbol{\theta}}_2)^{-1}(\widehat{\boldsymbol{\theta}}_2-\boldsymbol{\xi})(\widehat{\boldsymbol{\theta}}_2-\boldsymbol{\xi})'\right)\\ &= \mathrm{tr}\left(\mathrm{Var}(\widehat{\boldsymbol{\theta}}_2)^{-1}\left[\mathrm{Var}(\widehat{\boldsymbol{\theta}}_2-\boldsymbol{\xi})+\mathbb{E}(\widehat{\boldsymbol{\theta}}_2-\boldsymbol{\xi})\mathbb{E}(\widehat{\boldsymbol{\theta}}_2-\boldsymbol{\xi})'\right]\right)\\ &= \mathrm{tr}\left(\boldsymbol{I}_s+\mathrm{Var}(\boldsymbol{H}'\hat{\boldsymbol{\beta}})^{-1}(\boldsymbol{H}'\boldsymbol{\beta}-\boldsymbol{\xi})(\boldsymbol{H}'\boldsymbol{\beta}-\boldsymbol{\xi})'\right)\\ &= s+(\boldsymbol{H}'\boldsymbol{\beta}-\boldsymbol{\xi})'\,\mathrm{Var}(\boldsymbol{H}'\hat{\boldsymbol{\beta}})^{-1}(\boldsymbol{H}'\boldsymbol{\beta}-\boldsymbol{\xi}).\end{aligned}$$

Note that $\mathrm{Col}(\widetilde{\boldsymbol{X}})=\mathrm{Col}(\boldsymbol{X})$ and $\mathrm{Col}(\widetilde{\boldsymbol{X}}_1)^\perp\cap\mathrm{Col}(\widetilde{\boldsymbol{X}})$ is spanned by the columns of $\widetilde{\boldsymbol{Z}}_2:=(\boldsymbol{I}_n-\boldsymbol{P}_{\widetilde{\boldsymbol{X}}_1})\widetilde{\boldsymbol{X}}_2$. According to Cochran's theorem, it follows from

$$\boldsymbol{Y}-\hat{\boldsymbol{Y}}=(\boldsymbol{I}_n-\boldsymbol{P}_{\widetilde{\boldsymbol{X}}})(\boldsymbol{Y}-\widetilde{\boldsymbol{X}}_2\boldsymbol{\xi})$$

and

$$\hat{\boldsymbol{Y}}-\widetilde{\boldsymbol{Y}}=\widetilde{\boldsymbol{Z}}_2(\boldsymbol{H}'\hat{\boldsymbol{\beta}}-\boldsymbol{\xi})=\boldsymbol{P}_{\widetilde{\boldsymbol{Z}}_2}(\boldsymbol{Y}-\widetilde{\boldsymbol{X}}_2\boldsymbol{\xi})$$

that

$\|\boldsymbol{Y}-\hat{\boldsymbol{Y}}\|^2/\sigma^2\sim\chi^2_{n-r}$ and $(\|\boldsymbol{Y}-\widetilde{\boldsymbol{Y}}\|^2-\|\boldsymbol{Y}-\hat{\boldsymbol{Y}}\|^2)/\sigma^2\sim\chi^2_s(\lambda)$ are independent, provided that $\boldsymbol{\varepsilon}\sim\mathcal{N}_n(\boldsymbol{0}_n,\sigma^2\boldsymbol{I}_n)$, where $\lambda=(\boldsymbol{H}'\boldsymbol{\beta}-\boldsymbol{\xi})'\,\mathrm{Var}(\boldsymbol{H}'\hat{\boldsymbol{\beta}})^{-1}(\boldsymbol{H}'\boldsymbol{\beta}-\boldsymbol{\xi})$. Therefore,

$$\frac{(\|\boldsymbol{Y}-\widetilde{\boldsymbol{Y}}\|^2-\|\boldsymbol{Y}-\hat{\boldsymbol{Y}}\|^2)/s}{\|\boldsymbol{Y}-\hat{\boldsymbol{Y}}\|^2/(n-r)}\overset{H_0}{\sim}F_{s,n-r}.$$

Such a statistic is the ratio of two scaled sums of squares reflecting different sources of variability, giving rise to an ***F*-test**. The test statistic tends to be greater when H_0 is not true.

To improve the comprehensibility, denote $\gamma=\boldsymbol{H}'\boldsymbol{\beta}$ and $\gamma_0=\boldsymbol{\xi}$. Generally speaking, a reasonable test procedure would reject $H_0:\gamma=\gamma_0$ if the *distance* between $\hat{\gamma}$ and γ_0 is large, taking the variability of $\hat{\gamma}$ into consideration. Notice that $\hat{\gamma}=\boldsymbol{H}'\hat{\boldsymbol{\beta}}$ satisfies $\mathbb{E}\hat{\gamma}=\gamma$ and $\mathrm{Var}(\hat{\gamma})=\sigma^2D$, where

$$D:=\boldsymbol{H}'(\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{H}$$

does not depend on $(\boldsymbol{X}'\boldsymbol{X})^-$ and is positive definite. For the sake of inference, assume that $\boldsymbol{\varepsilon}\sim\mathcal{N}_n(\boldsymbol{0}_n,\sigma^2\boldsymbol{I}_n)$. Thus, $\hat{\gamma}\sim\mathcal{N}_s(\gamma,\sigma^2D)$ is independent of $\hat{\sigma}^2=R_0^2/(n-r)$, where $R_0^2=\|(\boldsymbol{I}_n-\boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{Y}\|^2\sim\sigma^2\chi^2_{n-r}$.

- Consider the case $s=1$, i.e., $\gamma\in\mathbb{R}$ is a single parameter.
  - Z-statistic $Z=\dfrac{\hat{\gamma}-\gamma_0}{\sqrt{\sigma^2D}}\overset{H_0}{\sim}\mathcal{N}(0,1)$ if σ^2 is known; and
    we can always replace σ^2 with its unbiased estimator $\hat{\sigma}^2$ to obtain
  - T-statistic $T=\dfrac{\hat{\gamma}-\gamma_0}{\sqrt{\hat{\sigma}^2D}}\overset{H_0}{\sim}t_{n-r}$ if σ^2 is unknown, which also gives $(\hat{\gamma}-\gamma_0)^2/(\hat{\sigma}^2D)\overset{H_0}{\sim}F_{1,n-r}$.
    Then a confidence interval of confidence level $1-\alpha$ for γ is $\hat{\gamma}\pm t_{\frac{\alpha}{2},n-r}\sqrt{\hat{\sigma}^2D}$.
- The **Mahalanobis distance** of $\hat{\gamma}$ from $\mathcal{N}_s(\gamma_0,\sigma^2D)$ is defined as

  $$\|\hat{\gamma}-\gamma_0\|_{(\sigma^2D)^{-1}}=\sqrt{(\hat{\gamma}-\gamma_0)'(\sigma^2D)^{-1}(\hat{\gamma}-\gamma_0)}.$$

  Since $D^{-1/2}(\hat{\gamma}-\gamma_0)\sim\mathcal{N}_s(D^{-1/2}(\gamma-\gamma_0),\sigma^2I_s)$, it is immediate that

  $$\|\hat{\gamma}-\gamma_0\|^2_{(\sigma^2D)^{-1}}=(\hat{\gamma}-\gamma_0)'D^{-1}(\hat{\gamma}-\gamma_0)/\sigma^2\sim\chi^2_s(\lambda),$$

  where $\lambda=(\gamma-\gamma_0)'D^{-1}(\gamma-\gamma_0)/\sigma^2$. In fact, $\|\hat{\gamma}-\gamma_0\|^2_{(\sigma^2D)^{-1}}=\|\hat{\boldsymbol{Y}}-\widetilde{\boldsymbol{Y}}\|^2/\sigma^2$ has expectation $s+\lambda$. In addition,

  $$\mathbb{E}(\hat{\gamma}-\gamma_0)'D^{-1}(\hat{\gamma}-\gamma_0)/s=(1+\lambda/s)\sigma^2\geq\sigma^2,$$

  with equality holding just when $\gamma=\gamma_0$. One may reject H_0 if $(\hat{\gamma}-\gamma_0)'D^{-1}(\hat{\gamma}-\gamma_0)/(s\sigma^2)$ is large. If σ^2 is unknown, replacing σ^2 with $\hat{\sigma}^2$ yields

  $$\frac{(\hat{\gamma}-\gamma_0)'D^{-1}(\hat{\gamma}-\gamma_0)}{s\hat{\sigma}^2}=\frac{\|\hat{\boldsymbol{Y}}-\widetilde{\boldsymbol{Y}}\|^2/s}{\|\boldsymbol{Y}-\hat{\boldsymbol{Y}}\|^2/(n-r)}\sim F_{s,n-r}(\lambda),$$

  where $\lambda=0$ if and only if $H_0:\gamma=\gamma_0$ is true.

§10 Multiple Testing Techniques (2020/4/1)

The **multiple testing problem** occurs when one considers a set of statistical inferences *simultaneously*. A general method for constructing *simultaneous confidence intervals* (S.C.I.s) of level $(1-\alpha)$ is called the **Bonferroni correction**, designed to handle a *finite* number of hypotheses. If E_j's are events such that $\mathbb{P}(E_j) = 1-\alpha_j$ for $j = 1, \cdots, m$, then

$$\mathbb{P}(\cap E_j) = 1 - \mathbb{P}(\cup E_j^{\complement}) \geq 1 - \sum \alpha_j,$$

and thus $\mathbb{P}(\cap E_j) \geq 1-\alpha$ provided that $\alpha_j = \alpha/m$ for all j. For example, suppose $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}_p(\mu, \Sigma)$. The pivot $\frac{\sqrt{n}(u'\bar{X}-u'\mu)}{\sqrt{u'Su}} \sim t_{n-1}$ induces a confidence interval $u'\bar{X} \pm t_{\frac{\alpha}{2},n-1}\sqrt{\frac{u'Su}{n}}$ for $u'\mu$ of level $(1-\alpha)$ for any single $u \in \mathbb{R}^p$, where $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ is the sample mean and $S = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})'$ is the sample covariance matrix. It follows that

$$\mathbb{P}\left(\bigcap_{j=1}^m \left\{u_j'\mu \in u_j'\bar{X} \pm t_{\frac{\alpha}{2m},n-1}\sqrt{\frac{u_j'Su_j}{n}}\right\}\right) \geq 1-\alpha, \quad \forall u_1, \ldots, u_m \in \mathbb{R}^p.$$

The Bonferroni method is *conservative*: the intervals are wider than ones that would produce *exactly* the desired confidence level. The next methods discussed are more limited, but do not have this disadvantage.

For the normal linear model

$$\underset{(n\times 1)}{Y} = \underset{(n\times p)}{X}\underset{(p\times 1)}{\beta} + \underset{(n\times 1)}{\varepsilon}, \qquad \varepsilon \sim \mathcal{N}_n(0_n, \sigma^2 I_n), \quad \operatorname{rank}(X) = r,$$

Scheffé's method applies to any estimable linear transformation $\gamma = H'\beta$ of the parameters $\beta \in \mathbb{R}^p$, where $H = (h_1, \ldots, h_s) \in \mathbb{R}^{p\times s}$ has full column rank. Recall that $\hat{\gamma} = H'\hat{\beta} \sim \mathcal{N}_s(\gamma, \sigma^2 D)$ is independent of $\hat{\sigma}^2 = R_0^2/(n-r)$, where $D = H'(X'X)^- H$ and $R_0^2 = \|(I_n - P_X)Y\|^2 \sim \sigma^2\chi^2_{n-r}$. For any fixed $u \in \mathbb{R}^s$, a $(1-\alpha)$ confidence interval for $u'\gamma$ could be $u'\hat{\gamma} \pm t_{\frac{\alpha}{2},n-r}\sqrt{\hat{\sigma}^2 u'Du}$, derived from

$$\frac{u'\hat{\gamma} - u'\gamma}{\sqrt{\hat{\sigma}^2 u'Du}} \sim t_{n-r} \impliedby u'\hat{\gamma} \sim \mathcal{N}(u'\gamma, \sigma^2 u'Du).$$

Allowing $u \in \mathbb{R}^s$ to vary arbitrarily, we have

$$\sup_{u\neq 0_s} \frac{|u'\hat{\gamma} - u'\gamma|^2}{u'Du} \overset{(v=D^{1/2}u)}{=\!=\!=} \sup_{v\neq 0_s} \frac{\left|v'D^{-1/2}(\hat{\gamma}-\gamma)\right|^2}{v'v} = (\hat{\gamma}-\gamma)'D^{-1}(\hat{\gamma}-\gamma)$$

by Cauchy-Schwarz inequality. Since $(\hat{\gamma}-\gamma)'D^{-1}(\hat{\gamma}-\gamma) \sim \sigma^2\chi_s^2$, one can see that

$$\sup_{u\neq 0_s} \frac{|u'\hat{\gamma} - u'\gamma|^2}{s\hat{\sigma}^2 u'Du} = \frac{(\hat{\gamma}-\gamma)'D^{-1}(\hat{\gamma}-\gamma)}{s\hat{\sigma}^2} \sim F_{s,n-r},$$

and therefore

$$\mathbb{P}\left(\bigcap_{u\in\mathbb{R}^s}\left\{u'\gamma \in u'\hat{\gamma} \pm \sqrt{s\hat{\sigma}^2(u'Du)F_{\alpha,s,n-r}}\right\}\right) = \mathbb{P}\left\{\sup_{u\neq 0_s}\frac{|u'\hat{\gamma}-u'\gamma|^2}{s\hat{\sigma}^2u'Du} \leq F_{\alpha,s,n-r}\right\} = 1-\alpha.$$

In other words, replacing $t_{\frac{\alpha}{2},n-r} = \sqrt{F_{\alpha,1,n-r}}$ with $\sqrt{sF_{\alpha,s,n-r}}$ in the single confidence interval yields S.C.I.s, for which the Bonferroni correction will adopt $t_{\frac{\alpha}{2m},n-r}$ if $m < \infty$ linear combinations of γ are concerned. In practice, we always prefer the *shortest* intervals with the given confidence level $(1-\alpha)$ so that the *most accurate* estimation can be achieved.

Tukey proposed a **method** for simultaneously comparing means of several normal distributions, which leads naturally to Tukey's **honestly significant difference**[viii] (HSD) test. If $Z_1, \ldots, Z_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ and $R^2 \sim \chi^2_\nu$ are independent, then it's said that

$$\frac{\max_{1\leqslant i\leqslant n} Z_i - \min_{1\leqslant i\leqslant n} Z_i}{\sqrt{R^2/\nu}} \sim Q_{n,\nu},$$

the **Studentized range distribution**, which is similar to the Student's t-distribution of $\frac{\sqrt{n}\bar{Z}}{\sqrt{R^2/\nu}} \sim t_\nu$.

[viii] *Meaning no disrespect to a great statistician, John Tukey, I've never been able to think of this as anything other than Honest John's Significant Difference.* —— Christensen [C]

To illustrate how Tukey's method works, consider a balanced one-way ANOVA model:

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2), \quad j = 1, \cdots, m;\ i = 1, \cdots, k.$$

We are often interested in $\alpha_i - \alpha_{i'}$ for $1 \le i, i' \le k$, which are certainly estimable since $\mu + \alpha_i$ are estimable. Clearly

$$\hat{\mu} + \hat{\alpha}_i = \overline{Y}_{i\cdot} = \frac{1}{m}\sum_{j=1}^{m} Y_{ij} \sim \mathcal{N}(\mu + \alpha_i, \sigma^2/m), \quad \forall i = 1, \cdots, k.$$

Let $\hat{\sigma} = \sqrt{R_0^2/(n-k)}$, where $n = km$ and $R_0^2 = \sum_{i=1}^{k}\sum_{j=1}^{m}(Y_{ij} - \overline{Y}_{i\cdot})^2 \sim \sigma^2\chi^2_{n-k}$ is the residual sum of squares obtained from OLS. It's immediate that

$$\frac{\sqrt{m}}{\hat{\sigma}} \max_{1\le i,i'\le k}\left[(\overline{Y}_{i\cdot} - \overline{Y}_{i'\cdot}) - (\alpha_i - \alpha_{i'})\right] = \left[\max_{1\leqslant i\leqslant k} \tfrac{\sqrt{m}(\overline{Y}_{i\cdot} - \mu - \alpha_i)}{\sigma} - \min_{1\leqslant i\leqslant k} \tfrac{\sqrt{m}(\overline{Y}_{i\cdot} - \mu - \alpha_i)}{\sigma}\right] \Big/ \sqrt{\tfrac{R_0^2/\sigma^2}{n-k}} \sim Q_{k,n-k}.$$

Thus,

$$\mathbb{P}\left(\bigcap_{1\le i,i'\le k}\left\{(\alpha_i - \alpha_{i'}) \in \underline{(\overline{Y}_{i\cdot} - \overline{Y}_{i'\cdot}) \pm \frac{\hat{\sigma}}{\sqrt{m}} Q_{\alpha,k,n-k}}\right\}\right) = 1 - \alpha.$$

After omitting the common factor $\sqrt{2\hat{\sigma}^2/m}$, i.e., the standard error of $\hat{\alpha}_i - \hat{\alpha}_{i'}$ for $1 \le i < i' \le k$, the half lengths of $(1-\alpha)$ S.C.I.s for those $\alpha_i - \alpha_{i'}$ satisfy that

$$\underbrace{Q_{\alpha,k,n-k}/\sqrt{2}}_{\text{(Tukey)}} \le \min\Big(\underbrace{\sqrt{kF_{\alpha,k,n-k}}}_{\text{(Scheffé)}},\ \underbrace{t_{\frac{\alpha}{k(k-1)},n-k}}_{\text{(Bonferroni)}}\Big)$$

because Scheffé's method takes all linear combinations of $(\mu + \alpha_i)_{1\le i\le k}$ into account rather than merely their differences, and the Bonferroni method exploits far less information about distributions.

It's worth noting that *concentration inequalities*[ix] help multiple simultaneous statistical tests. See, e.g., Freedman, David. Wald Lecture: On the Bernstein-von Mises theorem with infinite-dimensional parameters. *Ann. Statist.* 27 (1999), no. 4, 1119–1141. https://doi.org/10.1214/aos/1017938917.

§11 Model Selection (2020/4/8)

In spite of most realistic situations that the true model is not in our candidate class, we shall try our best to make our model useful, i.e., to approximate the true model, which often involves nonparametric (infinite-dimensional) methods. It has been said that "*all models are wrong, but some are useful.*" Note that models, offering ways of inference for the population features, ultimately serve the *prediction*, from the perspective of supervised learning. If, coincidentally, the true model belongs to the candidate class, then our goal is simply to find it out, in which case we often utilize parametric (finite-dimensional) methods and pay attention to *selection consistency*.

For the purpose of model selection, basically we need some criteria that are to be minimized, which will be detailed later. Sometimes one may intuitively carry out sequential testing for nested models, and keep only the covariates whose coefficients are significant. Note that information criteria, concentrating on variable selection, are not necessarily confined to nested models.

The other layer is selection approach. For computational and statistical reasons, *best subset* selection of variables may suffer from an enormous search space. Thus, forward and backward methods are attractive alternatives. There is a serious problem that the order in which variables are added or removed affects the outcome of the selection to a considerable extent, especially when some variables are strongly correlated. This also confirms the excellence of orthogonal design. Apart from discrete cases, selection could apply to continuous indices. For example, shrinkage penalties that regularize the coefficient estimates are commonly used in high-dimensional statistics ($p \gg n$), where models are identified by a few *tuning parameters*, e.g., LASSO, SCAD, etc. Determining the optimal value for the regularization parameter is a fundamental part of ensuring that the model performs well.

[ix] cf. https://math.stackexchange.com/q/89030 and related links

Suppose that the underlying true model is

$$\boldsymbol{Y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\mu} \in \mathbb{R}^n, \ \mathbb{E}[\boldsymbol{\varepsilon}] = \mathbf{0}_n, \ \mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I}_n.$$

The candidate class consists of *stepwise* linear regression models

$$\boldsymbol{Y} = \boldsymbol{X}_p \boldsymbol{\beta}_p + \boldsymbol{\varepsilon}, \quad p = 1, \cdots, m$$

where $\boldsymbol{\beta}_p \in \mathbb{R}^p$ is the unknown parameter, and the design matrix $\boldsymbol{X}_p \in \mathbb{R}^{n \times p}$ is assumed full column rank for simplicity. Here $\boldsymbol{X}_p$'s form a nested sequence in the sense that $\mathrm{Col}(\boldsymbol{X}_1) \subset \mathrm{Col}(\boldsymbol{X}_2) \subset \cdots \subset \mathrm{Col}(\boldsymbol{X}_m)$, e.g., splines with more and more knots, or Fourier expansions of higher and higher order, where m depends on the context as n varies. The method of OLS yields $\hat{\boldsymbol{\beta}}_p = (\boldsymbol{X}_p' \boldsymbol{X}_p)^{-1} \boldsymbol{X}_p' \boldsymbol{Y}$, and the fitted values are

$$\hat{\boldsymbol{\mu}}_p := \boldsymbol{X}_p \hat{\boldsymbol{\beta}}_p = \boldsymbol{X}_p (\boldsymbol{X}_p' \boldsymbol{X}_p)^{-1} \boldsymbol{X}_p' \boldsymbol{Y} = \boldsymbol{P}_{\boldsymbol{X}_p} \boldsymbol{Y}.$$

Clearly, the *bias*

$$\mathbf{b}_p := \boldsymbol{\mu} - \mathbb{E}\hat{\boldsymbol{\mu}}_p = (\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}_p}) \boldsymbol{\mu}$$

measures the goodness of fit and its norm decreases with p. Particularly, if $\boldsymbol{\mu} = \boldsymbol{X}_{p^*} \boldsymbol{\beta}_{p^*}$ for some p^*, i.e., one of the candidate models is correct, then $\mathbf{b}_p = \mathbf{0}_n$ for all $p \geq p^*$. Besides, the *mean squared error* of $\hat{\boldsymbol{\mu}}_p$,

$$d_p := \mathbb{E} \left\| \hat{\boldsymbol{\mu}}_p - \boldsymbol{\mu} \right\|^2 = \mathbb{E} \left\| \boldsymbol{P}_{\boldsymbol{X}_p} \boldsymbol{\varepsilon} - (\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}_p}) \boldsymbol{\mu} \right\|^2 = \mathbb{E} \left\| \boldsymbol{P}_{\boldsymbol{X}_p} \boldsymbol{\varepsilon} \right\|^2 + \left\| (\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}_p}) \boldsymbol{\mu} \right\|^2 = p\sigma^2 + \left\| \mathbf{b}_p \right\|^2,$$

decomposes into its total variance $\mathrm{tr}\left(\mathrm{Var}(\hat{\boldsymbol{\mu}}_p)\right) = p\sigma^2$, which reflects the model complexity confounded with noise level, and the squared bias $\left\| \mathbf{b}_p \right\|^2$ of $\hat{\boldsymbol{\mu}}_p$. Generally, the optimal p can be estimated by

$$\hat{p} = \underset{1 \leq p \leq m}{\arg\min}\, \hat{d}_p,$$

where the estimate $\hat{d}_p$ of d_p should be specified.

Since the **sum of squared errors** for the candidate model with p covariates,

$$\mathrm{SSE}_p := \left\| \boldsymbol{Y} - \boldsymbol{X}_p \hat{\boldsymbol{\beta}}_p \right\|^2 = \left\| (\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}_p})(\boldsymbol{\mu} + \boldsymbol{\varepsilon}) \right\|^2 = \left\| \mathbf{b}_p + (\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}_p}) \boldsymbol{\varepsilon} \right\|^2,$$

which decreases with p, satisfies that

$$\mathbb{E}\mathrm{SSE}_p = \mathbb{E} \left\| (\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}_p}) \boldsymbol{\varepsilon} \right\|^2 + \left\| \mathbf{b}_p \right\|^2 = (n - p)\sigma^2 + \left\| \mathbf{b}_p \right\|^2,$$

we can define

$$\hat{d}_p := \mathrm{SSE}_p + 2p\hat{\sigma}^2,$$

where $\hat{\sigma}^2$ is unbiased for σ^2 so that $\mathbb{E}[\hat{d}_p - n\hat{\sigma}^2] = d_p$. Note that $n\hat{\sigma}^2$ is irrelevant to the selection procedure.

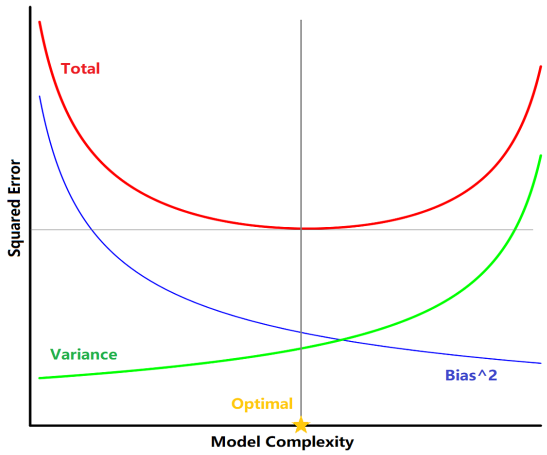

Figure 2: Tradeoff between Parsimony and Fidelity

§12 Model Selection Cont'd (2020/4/10)

Such $\hat{d}_p$ defined previously can be understood via prediction. Imagine a vector of new data $\boldsymbol{Y}^* = \boldsymbol{\mu} + \boldsymbol{\varepsilon}^*$, where $\boldsymbol{\varepsilon}^*$ is an independent copy of $\boldsymbol{\varepsilon}$. The expectation of the *sum of squared prediction errors* is

$$\mathbb{E}\mathrm{PE}_p := \mathbb{E}\left\|\boldsymbol{Y}^* - \hat{\boldsymbol{\mu}}_p\right\|^2 = \mathbb{E}\left\|\boldsymbol{\varepsilon}^*\right\|^2 + \mathbb{E}\left\|\hat{\boldsymbol{\mu}}_p - \boldsymbol{\mu}\right\|^2 = (n+p)\sigma^2 + \left\|\mathbf{b}_p\right\|^2.$$

Hence, we may take $\widehat{\mathrm{PE}}_p := \mathrm{SSE}_p + 2p\hat{\sigma}^2 = \hat{d}_p$.

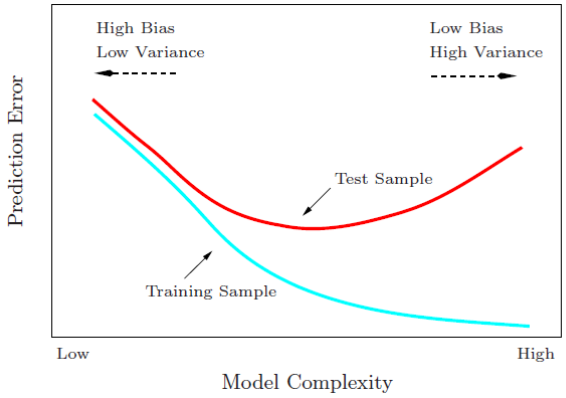

Figure 3: Prediction Error in Training and in Test[x)]

Now we are faced with the tradeoff between SSE_p and $2p\hat{\sigma}^2$. So, how to estimate σ^2, the *nuisance* parameter? Generally we may normalize squared errors using their *degree of freedom*[xi)].

♦ If we choose $\hat{\sigma}^2 = \mathrm{MSE}_p := \mathrm{SSE}_p/(n-p)$, the criterion

$$\mathrm{FPE}_p = \mathrm{SSE}_p + 2p\mathrm{MSE}_p = \tfrac{n+p}{n-p}\mathrm{SSE}_p$$

is exactly Akaike's **final prediction error**. Note that for underfitted models, MSE_p can be dramatically larger than σ^2. Thus, a large $\hat{p}$ will be encouraged for an *adequate* model.

♦ If we choose $\hat{\sigma}^2 = \mathrm{MSE}_m$ typically, using the largest model, then $\hat{d}_p = \mathrm{SSE}_p + 2p\mathrm{MSE}_m$ gives one of the most popular criteria,

$$C_p := \tfrac{1}{n}(\mathrm{SSE}_p + 2p\tfrac{\mathrm{SSE}_m}{n-m}),$$

called **Mallow's** C_p, sometimes alternatively defined as

$$C_p' := \mathrm{SSE}_p/\mathrm{MSE}_m - (n-2p) = nC_p/\mathrm{MSE}_m - n.$$

♣ (Intermezzo: nonparametric regression) In the setting of $y = m(x) + \varepsilon$ with $\varepsilon \sim (0, \sigma^2)$, another method is to use $\hat{\sigma}^2 = \sum_{i=1}^{n-1}(y_{i+1} - y_i)^2/[2(n-1)]$, where observations are arranged so that $x_1 \le \cdots \le x_n$. Besides, overfitted/under-smoothed models are often preferred to estimate σ^2 in semi-parametric framework.

Other criteria to be minimized among $1 \le p \le m$ include famous AIC and BIC.

- **Akaike information criterion**

$$\mathrm{AIC}_p = -2\log(\hat{L}_n) + 2p,$$

where $\hat{L}_n$ is the maximum value of the likelihood function of the model with p covariates. In the case of (Gaussian) linear regression, ignoring the constant terms allows us to conveniently take

$$\mathrm{AIC}_p = n\log(\mathrm{SSE}_p/n) + 2p,$$

which is asymptotically equivalent to Mallow's C_p.

[x)] from https://cosx.org/2015/08/some-basic-ideas-and-methods-of-model-selection/
[xi)] cf. https://www.zhihu.com/question/20983193/answer/784045148

- **Bayesian information criterion** (developed by Schwarz who gave a Bayesian argument)

$$\mathrm{BIC}_p = -2\log(\hat{L}_n) + \log(n)p,$$

where $\hat{L}_n$ is the maximum value of the likelihood function of the model with p covariates. In the case of (Gaussian) linear regression, ignoring the constant terms allows us to conveniently take

$$\mathrm{BIC}_p = n\log(\mathrm{SSE}_p/n) + \log(n)p.$$

When fitting models, it is possible to increase the likelihood by adding parameters, but doing so may result in overfitting. Both BIC and AIC attempt to resolve this problem by introducing a penalty term for the number of parameters in the model; the penalty term is larger in BIC than in AIC.

Let $\hat{p}_M$ be the model chosen by Mallow's C_p, which is asymptotically equivalent to AIC, and let $\hat{p}_S$ be the model chosen by Schwarz's BIC. That is,

$$\hat{p}_M = \underset{p\in\{1,\cdots,m\}}{\arg\min}\ C_p, \quad \text{and} \quad \hat{p}_S = \underset{p\in\{1,\cdots,m\}}{\arg\min}\ \mathrm{BIC}_p.$$

Their *optimalities* are stated as follows. Assume that $m \to \infty$ as $n \to \infty$.

1. If $\boldsymbol{\mu} \neq \boldsymbol{X}_p\boldsymbol{\beta}_p$ for $p = 1, \cdots, m$, i.e., the true model is not in the candidate class, then

$$\frac{\|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_{\hat{p}_M}\|}{\inf_{1\le p\le m}\|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_p\|} \xrightarrow{\mathbb{P}} 1, \quad \text{and} \quad \frac{d_{\hat{p}_M}}{\inf_{1\le p\le m} d_p} \xrightarrow{\mathbb{P}} 1.$$

2. If $\boldsymbol{\mu} = \boldsymbol{X}_{p^*}\boldsymbol{\beta}_{p^*}$ for some p^*, i.e., the true model belongs to the candidate class, then

$$\mathbb{P}\{\hat{p}_S = p^*\} \to 1.$$

Note that AIC (or Mallow's C_p) aims for *prediction*, tends to overfitting and cannot capture consistent models (under situation 2); BIC aims for model *estimation* and tends to underfitting (under situation 1). As a matter of fact,

$$\underline{\lim}\,\mathbb{P}\{\hat{p}_S < \hat{p}_M\} > 0.$$

In practice, the underlying situation depends on applications at hand.

§13 Model Assessment & Principle of Maximum Likelihood (2020/4/15)

A widely used method for assessing the performance of prediction models is **cross-validation** (CV). In K-fold cross-validation, the original sample $(x_i, y_i)_{1\le i\le n}$ is randomly partitioned into K equal-sized subsamples. Let $\kappa : \{1, \cdots, n\} \to \{1, \cdots, K\}$ be an indexing function that indicates the partition to which observations are allocated by the randomization, respectively. Denote by $\hat{f}_{-k}(\cdot)$ the fitted function that is computed with the k^{th} part of the data removed. Then the *estimate of prediction error* is

$$\mathrm{CV} = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{n/K}\sum_{\kappa(i)=k}\left[y_i - \hat{f}_{-k}(x_i)\right]^2 = \frac{1}{n}\sum_{i=1}^{n}\left[y_i - \hat{f}_{-\kappa(i)}(x_i)\right]^2.$$

In practice, one typically performs K-fold CV using $K = 5$ or $K = 10$, which is computationally efficient. The case $K = n$ is known as *leave-one-out* cross-validation, where

$$\mathrm{LOOCV} = \frac{1}{n}\sum_{i=1}^{n}\left[y_i - \hat{f}_{(-i)}(x_i)\right]^2.$$

Consider the linear regression model with p regressors, where $f(x) = x'\beta$ and $y - f(x) = \varepsilon \sim (0, \sigma^2)$. Clearly we have $\hat{f}(x) = x'\hat{\beta}$ and $\hat{f}_{(-i)}(x) = x'\hat{\beta}_{(-i)}$ by the method of least squares. Forecasting a new response y^* at x^* gives prediction error $\mathrm{PE} = \left[y^* - \hat{f}(x^*)\right]^2$. Then it holds that $\mathbb{E}\mathrm{LOOCV} = \mathbb{E}\mathrm{PE} + O(p/n^2)$.

Generally, LOOCV can be easily calculated for *ridge-type regression*, where the hat matrix takes the form

$$H = X(X'X + \lambda J)^{-1}X' = (h_{ij})_{1\le i,j\le n}.$$

The goal is

$$\mathrm{LOOCV} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{y_i - \hat{f}(x_i)}{1-h_{ii}}\right]^2,$$

so it suffices to show that

$$y_i - \hat{f}_{(-i)}(x_i) = \frac{y_i - \hat{f}(x_i)}{1-h_{ii}}.$$

Proof. Let $\tilde{y} = (\tilde{y}_j)_{1\le j\le n} = \left(y_1, \ldots, y_{i-1}, \hat{f}_{(-i)}(x_i), y_{i+1}, \ldots, y_n\right)'$. Then the minimizer $\hat{f}_{(-i)}(\cdot)$ of the leave-one-out penalized sum of squares $\sum_{j\neq i}[y_j - f(x_j)]^2 + \lambda\mathcal{J}(f)$ also minimizes $\sum_{j=1}^{n}[\tilde{y}_j - f(x_j)]^2 + \lambda\mathcal{J}(f)$. It follows that $\left(\hat{f}_{(-i)}(x_j)\right)_{1\le j\le n} = H\tilde{y}$. Just combine $\hat{f}_{(-i)}(x_i) = \sum_{j=1}^{n} h_{ij}\tilde{y}_j$ with $\hat{f}(x_i) = \sum_{j=1}^{n} h_{ij}y_j$. □

It's immediate that LOOCV is not stable if[xii] $h_{ii} \approx 1$ for some i. To alleviate the tendency to under-smooth, a feasible modification is **generalized cross-validation**, given by

$$\mathrm{GCV} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{y_i - \hat{f}(x_i)}{1-\mathrm{tr}\,(H)/n}\right]^2.$$

Using the fact that $(1-p/n)^{-2} \approx (n+p)/(n-p)$ when $\mathrm{tr}\,(H) = p \ll n$, we have

$$\mathrm{GCV} = \frac{1}{n}\left(1-\frac{p}{n}\right)^{-2}\sum_{i=1}^{n}\left[y_i - \hat{f}(x_i)\right]^2 \approx \frac{1}{n}\frac{n+p}{n-p}\mathrm{SSE} = \frac{1}{n}\mathrm{FPE}.$$

Recall that the **maximum likelihood estimator** (MLE) of a parameter $\theta \in \Theta$ is defined as

$$\hat{\theta}_n \in \arg\max \ell_n(\theta),$$

where $\ell_n(\theta) = \log f(Y|\theta)$ is the *log-likelihood function*. Suppose that θ_0 is the true value of θ, i.e., $(Y_i)_{1\le i\le n}$ is sampled from $f(\bullet|\theta_0)$. To justify MLE, it's asserted that

$$\theta_0 \in \arg\max \mathbb{E}[\ell_n(\theta)].$$

Indeed, **Shannon's information inequality** states that

$$\mathbb{E}_{\theta_0}[\log f(Y|\theta)] < \mathbb{E}_{\theta_0}[\log f(Y|\theta_0)],$$

unless $f(\bullet|\theta) = f(\bullet|\theta_0)$ almost everywhere, which is impossible if θ is *identifiable*. Assume that Y_i's are i.i.d., then by the law of large numbers, $\frac{1}{n}\ell_n(\theta) \to \mathbb{E}_{\theta_0}[\log f(Y_1|\theta)]$ almost surely as $n \to \infty$. Under some mild assumptions, $\hat{\theta}_n$ is (strongly) consistent, as desired. See, e.g., overleaf.com/read/vskwkqfzbxgq

§14 Review of Likelihood Theory (2020/4/22)

Generally speaking, the existence and uniqueness of the MLE may not be established. Also, a certain regularity condition of $\ell_n(\cdot)$, the log-likelihood function, could fail. Particularly, if the density function $f(\bullet|\theta)$ has support $\{y : f(y|\theta) > 0\}$ depending on θ, then $\ell_n(\cdot)$ may not be differentiable with respect to θ, or its derivative $\dot{\ell}_n(\cdot)$ at some $\theta \in \Theta$ may not be an integrable random variable.

[xii] The diagonal elements of the hat matrix H are called **leverages** (i.e., influential points). As for OLS, $H = X(X'X)^{-1}X'$ is symmetric and idempotent, and thus $h_{ii} = \sum_{j=1}^{n} h_{ij}^2$, which also reads $h_{ii}(1-h_{ii}) = \sum_{j\neq i} h_{ij}^2$. If $h_{ii} \nearrow 1$, then $\hat{y}_i = h_{ii}y_i + \sum_{j\neq i} h_{ij}y_j$ will reflect contribution from y_i mostly. However, for prediction, we hope that y_j $(j \neq i)$ contribute to $\hat{y}_i$ as much as possible, while y_i contributes to $\hat{y}_i$ as even as possible.

Based on $\int f(y|\theta)\,\mathrm{d}y = 1,\ \forall \theta \in \Theta \subseteq \mathbb{R}^p$, successive differentiation[xiii] yields **Bartlett's identities** concerning the expectation of derivatives of $\ell_n(\theta; Y) = \log f(Y|\theta)$ with respect to θ. Taking the first order derivative gives

$$0_p = \int \frac{\partial}{\partial \theta} f(y|\theta)\,\mathrm{d}y = \int \left[\frac{\partial}{\partial \theta}\ell_n(\theta; y)\right] f(y|\theta)\,\mathrm{d}y = \mathbb{E}_\theta\left[\dot{\ell}_n(\theta)\right].$$

Next, taking the second order derivative gives

$$0_{p\times p} = \int \left\{\left[\frac{\partial^2}{\partial\theta\partial\theta'}\ell_n(\theta; y)\right] f(y|\theta) + \left[\frac{\partial}{\partial\theta}\ell_n(\theta; y)\right]\left[\frac{\partial}{\partial\theta'} f(y|\theta)\right]\right\}\mathrm{d}y = \mathbb{E}_\theta\left[\ddot{\ell}_n(\theta)\right] + \mathbb{E}_\theta\left[\dot{\ell}_n(\theta)\dot{\ell}_n(\theta)'\right].$$

Primarily, the first order derivative is used to determine the solution, and the second order derivative embodies the variability. The **Fisher information matrix** is then defined to be

$$\mathcal{I}_n(\theta) := \mathrm{Var}_\theta\Big(\dot{\ell}_n(\theta)\Big) = \mathbb{E}_\theta\Big[\dot{\ell}_n(\theta)\dot{\ell}_n(\theta)'\Big] = -\mathbb{E}_\theta\Big[\ddot{\ell}_n(\theta)\Big] \in \mathbb{R}^{p\times p}.$$

Moreover, one may take the third order derivative to obtain (exercise 7 in homework 3)

$$\mathbb{E}_\theta[\dddot{\ell}_n(\theta)] + 3\,\mathrm{Cov}_\theta\Big(\ddot{\ell}_n(\theta), \dot{\ell}_n(\theta)\Big) + \mathbb{E}_\theta\Big[\dot{\ell}_n(\theta)^3\Big] = 0, \text{ if } \theta \in \Theta \subseteq \mathbb{R}^1.$$

Denote the **score function** by $U_n(\theta) := \dot{\ell}_n(\theta)$. Suppose that the observed data Y_i's are independent but not necessarily identically distributed, each with density $f_i(\cdot)$ depending on $\theta_i = \theta_i(\theta)$ for $i = 1, 2, \cdots, n$. Note that θ_i's are often related to covariates. For example, in most regression models, θ_i is a transformation of $x_i'\beta$, where $x_i \in \mathbb{R}^p$ is the i^{th} observed covariate and $\beta \in \mathbb{R}^p$ is the regression parameter of interest. Let

$$u_i(\theta) := \tfrac{\partial}{\partial\theta}\log f_i(Y_i|\theta_i), \quad \& \quad \iota_i(\theta) := \mathrm{Var}_\theta(u_i(\theta)) = \mathbb{E}_\theta[u_i(\theta)u_i(\theta)'].$$

It is immediate that

$$U_n(\theta) = \sum_{i=1}^n u_i(\theta), \quad \& \quad \mathcal{I}_n(\theta) = \mathrm{Var}_\theta(U_n(\theta)) = \sum_{i=1}^n \iota_i(\theta).$$

To figure out the asymptotic normality of $U_n(\theta)$, i.e.,

$$\mathcal{I}_n(\theta)^{-\frac{1}{2}} U_n(\theta) \xrightarrow{d} \mathcal{N}_p(0_p, I_p),$$

it suffices to check two conditions in the Lindeberg-Feller CLT, which will be applied to

$$X_{ni} := \mathcal{I}_n(\theta)^{-\frac{1}{2}} u_i(\theta) = \mathcal{I}_n(\theta)^{-\frac{1}{2}}\iota_i(\theta)^{\frac{1}{2}}\big(\iota_i(\theta)^{-\frac{1}{2}}u_i(\theta)\big), \quad 1 \le i \le n.$$

On one hand, it's clear that $\mathbb{E}[X_{ni}] = 0_p$ and

$$\sum_{i=1}^n \mathrm{Var}(X_{ni}) = \sum_{i=1}^n \mathcal{I}_n(\theta)^{-\frac{1}{2}}\iota_i(\theta)\mathcal{I}_n(\theta)^{-\frac{1}{2}} = \mathcal{I}_n(\theta)^{-\frac{1}{2}}\mathcal{I}_n(\theta)\mathcal{I}_n(\theta)^{-\frac{1}{2}} = I_p.$$

On the other hand, we have to verify that X_{ni}'s are essentially tiny disturbances, in the sense that

$$\sum_{i=1}^n \mathbb{E}\left[\|X_{ni}\|^2\,\mathbb{1}_{\{\|X_{ni}\|>\varepsilon\}}\right] \xrightarrow{(n\to\infty)} 0, \quad \forall \varepsilon > 0.$$

Note that

$$\|X_{ni}\|^2 = \big(\iota_i(\theta)^{-\frac{1}{2}}u_i(\theta)\big)'\iota_i(\theta)^{\frac{1}{2}}\mathcal{I}_n(\theta)^{-1}\iota_i(\theta)^{\frac{1}{2}}\big(\iota_i(\theta)^{-\frac{1}{2}}u_i(\theta)\big) \le \lambda_{ni}\big\|\iota_i(\theta)^{-\frac{1}{2}}u_i(\theta)\big\|^2,$$

where λ_{ni} denotes the largest eigenvalue of $\iota_i(\theta)^{\frac{1}{2}}\mathcal{I}_n(\theta)^{-1}\iota_i(\theta)^{\frac{1}{2}}$, satisfying that

$$\sum_{i=1}^n \lambda_{ni} \le \sum_{i=1}^n \mathrm{tr}\left(\iota_i(\theta)^{\frac{1}{2}}\mathcal{I}_n(\theta)^{-1}\iota_i(\theta)^{\frac{1}{2}}\right) = \sum_{i=1}^n \mathrm{tr}\left(\mathcal{I}_n(\theta)^{-1}\iota_i(\theta)\right) = \mathrm{tr}\left(I_p\right) = p.$$

[xiii] To assure that $\frac{\partial}{\partial\theta}\int f(y|\theta)\,\mathrm{d}y = \int \frac{\partial}{\partial\theta} f(y|\theta)\,\mathrm{d}y$ at $\theta = \theta_0$ rigorously, a sufficient condition for exchangeability of differentiation and integral is the existence of a (locally) dominating function $h(y)$ such that $\left\|\frac{\partial}{\partial\theta} f(y|\theta)\right\| \le h(y)$ for every θ in a neighborhood of θ_0, and $\int h(y)\,\mathrm{d}y < \infty$. This does hold for exponential families.

It follows that

$$\sum_{i=1}^{n}\mathbb{E}\left[\|X_{ni}\|^2\,\mathbb{1}_{\{\|X_{ni}\|>\varepsilon\}}\right]\le\sum_{i=1}^{n}\lambda_{ni}\,\mathbb{E}\left[\left\|\iota_i(\theta)^{-\frac12}u_i(\theta)\right\|^2\mathbb{1}_{\{\|\iota_i(\theta)^{-\frac12}u_i(\theta)\|^2>\varepsilon^2/\lambda_{ni}\}}\right]$$
$$\le p\max_{1\leqslant i\leqslant n}\mathbb{E}\left[\left\|\iota_i(\theta)^{-\frac12}u_i(\theta)\right\|^2\mathbb{1}_{\{\|\iota_i(\theta)^{-\frac12}u_i(\theta)\|^2>\varepsilon^2/\lambda_{ni}\}}\right].$$

With

$$\lambda_n:=\max_{1\leqslant i\leqslant n}\lambda_{ni},$$

it suffices that

$$\max_{1\leqslant i\leqslant n}\mathbb{E}\left[\left\|\iota_i(\theta)^{-\frac12}u_i(\theta)\right\|^2\mathbb{1}_{\{\|\iota_i(\theta)^{-\frac12}u_i(\theta)\|^2>\varepsilon^2/\lambda_n\}}\right]\xrightarrow{(n\to\infty)}0,$$

somehow similar to uniformly integrable behavior. Note that it is usually required that $\lambda_n\to0$, which means that experiment units are evenly scattered. If Y_i's are i.i.d., then so are $u_i(\theta)$'s, in which case

$$\iota_i(\theta)=\tfrac1n\mathcal{I}_n(\theta)\implies\iota_i(\theta)^{\frac12}\mathcal{I}_n(\theta)^{-1}\iota_i(\theta)^{\frac12}=\tfrac1nI_p\implies\lambda_{ni}=\tfrac1n,$$

and thus those $Z_i:=\left\|\iota_i(\theta)^{-\frac12}u_i(\theta)\right\|$ satisfy that

$$\max_{1\leqslant i\leqslant n}\mathbb{E}\Big[Z_i^2\mathbb{1}_{\{Z_i^2>\varepsilon^2/\lambda_n\}}\Big]=\mathbb{E}\Big[Z_1^2\mathbb{1}_{\{Z_1^2>n\varepsilon^2\}}\Big]\xrightarrow{(n\to\infty)}0$$

by the dominated convergence theorem.

Going further, the asymptotic normality of the MLE $\hat\theta_n$ can be deduced using the *estimating equation*

$$U_n(\hat\theta_n)=0_p,$$

and a corollary of the continuous mapping theorem:

$$U_n(\hat\theta_n)=U_n(\theta)+\dot U_n(\theta)\big(\hat\theta_n-\theta\big)+\mathcal{I}_n(\theta)o_{\mathbb{P}}\big(\|\hat\theta_n-\theta\|\big).$$

Such an asymptotic argument exploiting Taylor's expansion is the so called *delta method*. It's direct that

$$\mathcal{I}_n(\theta)^{\frac12}\big(\hat\theta_n-\theta\big)=-\big(\mathcal{I}_n(\theta)^{-\frac12}\dot U_n(\theta)\mathcal{I}_n(\theta)^{-\frac12}\big)^{-1}\mathcal{I}_n(\theta)^{-\frac12}U_n(\theta)/[1+o_{\mathbb{P}}(1)]\xrightarrow{d}\mathcal{N}_p(0_p,I_p)$$

by Slutsky's theorem, provided that $-\mathcal{I}_n(\theta)^{-\frac12}\dot U_n(\theta)\mathcal{I}_n(\theta)^{-\frac12}\xrightarrow{\mathbb{P}}I_p$ and $\mathcal{I}_n(\theta)^{-\frac12}U_n(\theta)\xrightarrow{d}\mathcal{N}_p(0_p,I_p)$. This shows that $\hat\theta_n$ is asymptotically distributed as $\mathcal{N}\big(\theta,\mathcal{I}_n(\theta)^{-1}\big)$, enjoying the asymptotic efficiency.

§15 Likelihood Cont'd & Intro to GLMs (2020/4/24)

Let $\ell_n(\theta)=\log L_n(\theta)$ be the log-likelihood function. It can be seen that[xiv] being expanded around $\hat\theta_n$,

$$2\left(\ell_n(\hat\theta_n)-\ell_n(\theta)\right)\approx-2\left(\dot\ell_n(\hat\theta_n)'(\theta-\hat\theta_n)+\tfrac12(\theta-\hat\theta_n)'\ddot\ell_n(\hat\theta_n)(\theta-\hat\theta_n)\right)$$
$$=(\hat\theta_n-\theta)'\left(-\ddot\ell_n(\hat\theta_n)\right)(\hat\theta_n-\theta)$$
$$\approx(\hat\theta_n-\theta)'\mathcal{I}_n(\theta)(\hat\theta_n-\theta)\quad\xrightarrow{d}\chi^2_p.$$

Let's move on now to likelihood inference in the presence of nuisance parameters (arXiv:physics/0312079). Suppose that a parameter vector is partitioned as $\theta=\begin{pmatrix}\psi\\\lambda\end{pmatrix}$, where ψ is the parameter of primary interest, and λ is the nuisance parameter. For example, in linear regression, the regression coefficient β is of interest while the error variance σ^2 is a nuisance parameter. Note that β and σ^2 are so lucky to be *decoupled*, in the sense that the estimation of β has nothing to do with σ^2. Solving for θ is sometimes not an easy task. Thus, a typical approach is to solve for ψ and λ in an alternating way, e.g., **profiling maximum likelihood**:

[xiv] Moreover, for testing $H_0:\theta\in\Theta_0\leftrightarrow H_1:\theta\in\Theta_1=\Theta\setminus\Theta_0$, the **likelihood ratio statistic** $\Lambda_n:=\frac{\max_{\theta_0\in\Theta_0}L_n(\theta_0)}{\max_{\theta\in\Theta}L_n(\theta)}$ satisfies that $-2\log(\Lambda_n)\xrightarrow[H_0]{d}\chi^2_{\dim(\Theta)-\dim(\Theta_0)}$ under certain regularity conditions, which is referred to as *Wilks' theorem*. See, e.g., https://www.statlect.com/fundamentals-of-statistics/likelihood-ratio-test

1. Maximize $\ell_n(\psi, \lambda)$ with respect to ψ, treating λ as fixed;
2. Maximize $\ell_n(\psi, \lambda)$ with respect to λ, treating ψ as fixed;
3. Repeat steps 1. and 2. until convergence.

Recall again that in linear regression, there is no need to profile likelihood or least squares. Looking at the MLE, we have

$$\begin{pmatrix} \hat{\psi} \\ \hat{\lambda} \end{pmatrix} \overset{a}{\sim} \mathcal{N}\left(\begin{pmatrix} \psi \\ \lambda \end{pmatrix}, \mathcal{I}_n(\theta)^{-1} \right), \quad \& \quad \hat{\psi} \overset{a}{\sim} \mathcal{N}\left(\psi, \mathcal{I}_n(\psi)^{-1}\right).$$

By the inverse formula for a partitioned matrix[xv], it's immediate that (exercise 7 in homework 3)

$$\mathcal{I}_n(\theta) = \begin{pmatrix} i_{\psi\psi} & i_{\psi\lambda} \\ i_{\lambda\psi} & i_{\lambda\lambda} \end{pmatrix} \Longrightarrow \mathcal{I}_n(\psi) = i_{\psi\psi} - i_{\psi\lambda} i_{\lambda\lambda}^{-1} i_{\lambda\psi},$$

which is called the *Schur complement* of $i_{\lambda\lambda}$ in $\mathcal{I}_n(\theta)$. If ψ and λ can be decoupled, then $\hat{\psi}$ is free of λ, which often implies $i_{\psi\lambda} = 0$. In regression models, we usually encounter the case that $\hat{\psi}$ is free of λ, while $i_{\psi\psi}$ may involve λ.

Most data fall into one of the types below:

- numerical
  - continuous: normal-like (after transformation, e.g., Box-Cox)
  - discrete (count): somehow between continuous and categorical, perhaps heavy-tailed or skewed
- categorical
  - nominal: without natural ordering
    - binary or collapsed to binomial (in which case proportions are preferred)
    - multinomial, e.g., gender
  - ordinal: with natural ordering, sometimes treated as continuous for model parsimony, e.g., dose level

For analyzing all these different kinds of data, **generalized linear models** (GLMs) will come into play.

Responses / Covariates	Continuous	Count	Binary	Multinomial
Continuous or (Transformed) Count/Ordinal	Linear Reg	Poisson Reg, Gamma Reg	Binary Reg	Multinomial Reg
Binary	Two-Sample Comparison	Poisson Reg, Gamma Reg	Contingency Table, Binary Reg	Contingency Table, Multinomial Reg
Nominal	ANOVA	Poisson Reg, Gamma Reg	Contingency Table, Binary Reg	Contingency Table, Multinomial Reg

Table 1: Overview of Generalized Linear Models

Before investigating a general unifying theory and method to cover such models, two examples are presented:

- dilution assay — complementary log-log link (Fisher, 1922).

  Suppose that we wish to estimate the density ρ_0 of an infective organism in a given solution. Assuming for simplicity that the original solution is progressively diluted in powers of two, at the x^{th} dilution, the density of infective organisms is

  $$\rho_x = \rho_0 / 2^x, \quad x = 0, 1, 2, \cdots.$$

  After each dilution x, we "streak" m_x agar plates of known volume v, out of whom the number of infected ones is denoted by r_x. Then the observed proportion of infected plates $y_x = r_x/m_x$ may be regarded as the realization of an infection indicator Y such that

  $$\mathbb{E}[Y|x] = \mathbb{P}(Y = 1|x) = \pi_x$$

  is the probability that a plate at the x^{th} dilution is infected.

[xv] cf. https://zhuanlan.zhihu.com/p/78884647

The expected number of organisms on any plate is $\rho_x v$ and, under suitable mixing conditions, the actual number N_x of organisms follows a Poisson distribution with this parameter. Thus,

$$\pi_x = \mathbb{P}\{N_x \neq 0\} = 1 - \exp(-\rho_x v),$$

and it follows that

$$\log\big(-\log(1-\pi_x)\big) = \log(v) + \log(\rho_x) = \big(\log(v) + \log(\rho_0)\big) - x\log(2),$$

where the design covariates (x's) are associated with $-\log(2)$ as a slope.

- It is $\log\big(-\log(1-\pi_x)\big)$ rather than π_x itself that bears a linear relation with x.
- By the way, the variance of noise is clearly not constant.

- dose-response study — probit analysis (Bliss, 1935) & logit for proportions (Berkson, 1944).

  In toxicology experiments, we will focus on modeling the survival rate π_x of a certain animal at dose level x as a function of x. Let test animals be divided into n cells. The dose varies from cell to cell, but is administered uniformly within each cell. For the j^{th} cell subject to a known dose level x_j, the number y_j surviving out of the original m_j animals is recorded. Then the observed proportion y_j/m_j may be regarded as the realization of a survival indicator whose expectation is exactly π_x.

  With unknown parameters α and β to be estimated, the *probit model* is

$$\pi_x = \Phi(\alpha + \beta x) \iff \text{probit}(\pi_x) = \alpha + \beta x,$$

  where probit $= \Phi^{-1}$, and $\Phi(\cdot)$ is the cumulative distribution function of $\mathcal{N}(0,1)$. This model has the virtue that it respects the property that π_x is a probability and hence must lie between 0 and 1. For this reason alone, it is not normally sensible to contemplate linear models for probabilities.

  Alternatively, the *logistic model* (or *logit model*) is

$$\pi_x = \text{expit}(\alpha + \beta x) \iff \text{logit}(\pi_x) = \alpha + \beta x,$$

  where

$$\text{logit} : p \in (0,1) \mapsto \log\left(\tfrac{p}{1-p}\right) \in \mathbb{R},$$

  and

$$\text{expit} = \text{logit}^{-1} = \text{logistic} : \eta \in \mathbb{R} \mapsto \frac{1}{1+\exp(-\eta)} = \frac{\exp(\eta)}{\exp(\eta)+1} \in (0,1).$$

§16 Generalized Linear Models & Exponential Families (2020/4/29)

We now extend our scope from the linear model to the GLM. This extension encompasses (1) link functions of the mean equated to the linear predictor and (2) non-normal response distributions.

(1) From the systematic perspective, we may relax the linearity. In classic linear regression, $y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i$, which can be rewritten as $\mathbb{E}[y_i|\boldsymbol{x}_i] = \boldsymbol{x}_i'\boldsymbol{\beta}$. To make a difference, a differentiable (strictly) monotone function $g(\cdot)$, called the **link function**, will be introduced to give

$$g(\mathbb{E}[y_i|\boldsymbol{x}_i]) = \boldsymbol{x}_i'\boldsymbol{\beta}, \quad \text{or} \quad \mathbb{E}[y_i|\boldsymbol{x}_i] = g^{-1}(\boldsymbol{x}_i'\boldsymbol{\beta}).$$

If the error term ε_i is shown explicitly, our new model reads

$$y_i = g^{-1}(\boldsymbol{x}_i'\boldsymbol{\beta}) + \varepsilon_i, \quad i = 1, \cdots, n.$$

Note that link functions are *equivalent* up to a linear transform, since the **linear predictor** $\eta_i := \boldsymbol{x}_i'\boldsymbol{\beta}$ is a linear combination of explanatory variables, where the parameter $\boldsymbol{\beta} \in \mathbb{R}^p$ is to be estimated.

(2) As for the randomness, we may relax the distribution of the noise to be not necessarily Gaussian. The **canonical exponential families** will be considered, which cover binomial, Poisson, Gaussian, gamma and negative-binomial distributions, and many other frequently used ones. The discussion will be detailed later, where it can be seen that $\text{Var}(y_i|\boldsymbol{x}_i)$ depends on $\boldsymbol{x}_i$ (or, more precisely, is a function of $\mu_i := \mathbb{E}[y_i|\boldsymbol{x}_i]$), differing from that in the standard linear model.

For comparison, when we fit a linear regression after transforming y_i into $h(y_i)$, it is implicitly assumed that $h(y_i) = \boldsymbol{x}_i'\boldsymbol{\beta} + u_i$ takes an approximately Gaussian error $u_i \sim (0, \sigma_u^2)$, in which case $\mathbb{E}[h(y_i)|\boldsymbol{x}_i] = \boldsymbol{x}_i'\boldsymbol{\beta}$. In a generalized linear model, the mean $\mu_i := \mathbb{E}[y_i|\boldsymbol{x}_i]$ is transformed, by the link function $g(\cdot)$, instead of transforming the response y_i itself. The two methods can lead to quite different results; for example, the mean of log-transformed responses is not the same as the logarithm of the mean response. In general, transforming the mean μ_i often allows the results to be more easily interpreted, especially in that mean parameters remain on the same scale as the measured responses.

Note that the modeling framework can be further extended. Although regression models for independent responses have been unified under the class of GLMs, the issue of accounting for correlation arises when analyzing *repeated measurements*. In contrast to cross-sectional studies, where a single outcome is measured for each individual, longitudinal studies possess the defining feature that individuals are measured repeatedly through time. Because the set of observations on one subject/cluster tends to be intercorrelated, *longitudinal data* (a.k.a. *panel data* in econometrics) require special statistical methods to draw valid scientific inferences. The *mixed model* is a common and convenient tool to characterize such grouping structure, which allows for the addition of **random effects** to the linear predictor made up of *fixed effects*. The **linear mixed model**[xvi] (LMM) applies to normal responses, and the **generalized linear mixed model** (GLMM) extends it to non-normal data. Besides, the **generalized estimating equation** (GEE) is used to estimate the parameters of a GLM for clustered data, which provides a semi-parametric approach to longitudinal analysis. The very crux of GEE is instead of attempting to model the within-subject covariance structure, to treat it as a nuisance and simply model the mean response. Moreover, we can even discard the distribution assumption of the response.

Apart from a much more general treatment of the random error term, we may switch our attention to the linear predictor and try to make it more flexible. The *nonparametric* idea is to replace the usual linear combination of covariates with an unspecified smooth function. In modeling new data, one often has little knowledge of an appropriate form for the model, so the nonparametric approach is less liable to make bad mistakes. Among various nonparametric regression methods, there are **kernel regression** (KR), the **additive model** (AM), the **generalized additive model** (GAM), and the **generalized additive mixed model** (GAMM). *Semi-parametric* regression combines parametric and nonparametric models. The most popular methods are the **partially linear and generalized partially linear models** (PLM & GPLM), the **single and multiple index models** (SIM & MIM), and the **varying coefficient model** (VCM).

Let's go back to the GLM, where a monotone differentiable function $g(\cdot)$ links the mean response $\mu_i := \mathbb{E}[y_i|\boldsymbol{x}_i]$ to the linear predictor $\eta_i := \boldsymbol{x}_i'\boldsymbol{\beta}$ through the formula:

$$g(\mu_i) = \eta_i, \quad i = 1, \cdots, n.$$

The *random component* consists of independent observations $y_1, \ldots, y_n$ from the **exponential family** of distributions having probability density/mass function for y_i of the form

$$f(y; \theta, \phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right\},$$

where θ is called the **canonical parameter** and represents the location related to the regression parameter $\boldsymbol{\beta}$ of interest, while ϕ is called the **dispersion parameter** and represents the scale. Usually $a(\phi) = \phi/w$ for $\phi > 0$ and a known weight $w > 0$. Note that $b(\theta)$ appears as the *cumulant-generating function*. We now derive the mean and variance of the exponential family distributions. The log-likelihood for a single y is given by

$$\ell(\theta; \phi, y) = \log f(y; \theta, \phi) = \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi),$$

whose first two derivatives with respect to θ are

$$\frac{\partial}{\partial\theta}\ell(\theta; \phi, y) = \frac{y - \dot{b}(\theta)}{a(\phi)}, \quad \& \quad \frac{\partial^2}{\partial\theta^2}\ell(\theta; \phi, y) = -\frac{\ddot{b}(\theta)}{a(\phi)}.$$

Using Bartlett's identities $\mathbb{E}[\frac{\partial}{\partial\theta}\ell(\theta; \phi, y)] = 0$ and $\mathbb{E}[\frac{\partial^2}{\partial\theta^2}\ell(\theta; \phi, y)] + \mathrm{Var}(\frac{\partial}{\partial\theta}\ell(\theta; \phi, y)) = 0$, it can be seen that

$$\mathbb{E}[y] = \mu = \dot{b}(\theta), \quad \& \quad \mathrm{Var}(y) = a(\phi)\ddot{b}(\theta).$$

[xvi] cf. https://cosx.org/2014/04/lmm-and-me/

It's natural to assume that $\ddot{b}(\cdot) > 0$, whence $\dot{b}(\cdot)$ is invertible. Then

$$\theta = \dot{b}^{-1}(\mu),$$

where $\dot{b}^{-1}(\cdot)$ is called the **canonical link**. The **variance function** is defined to be

$$V(\cdot) = \ddot{b}(\dot{b}^{-1}(\cdot))$$

so that $V(\mu) = \ddot{b}(\theta)$ and $\mathrm{Var}(y) = a(\phi)V(\mu)$.

$y \sim$	$\mathcal{N}(\mu, \sigma^2)$	Poisson(μ)	$\frac{1}{m}$Binomial(m, π)[‡]	Gamma$(\nu, \frac{\mu}{\nu})$
$f(y;\theta,\phi)$	$\frac{1}{\sqrt{2\pi\sigma^2}}\exp\{-\frac{(y-\mu)^2}{2\sigma^2}\}$	$\frac{\exp(-\mu)\mu^y}{y!}$	$\binom{m}{my}\pi^{my}(1-\pi)^{m-my}$	$\frac{1}{\Gamma(\nu)}(\frac{\nu}{\mu})^\nu y^{\nu-1}\exp(-\frac{\nu}{\mu}y)$
Support	$(-\infty, \infty)$	$\{0, 1, 2, \cdots\}$	$\{0, \frac{1}{m}, \frac{2}{m}, \cdots, 1\}$	$(0, \infty)$
$\ell(\theta;\phi,y)$	$\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)$			
$\theta = \dot{b}^{-1}(\mu)$	μ	$\log\mu$	$\mathrm{logit}(\pi) = \log\frac{\pi}{1-\pi}$	$-\frac{1}{\mu}$
$b(\theta)$	$\frac{1}{2}\theta^2$	$\exp(\theta)$	$\log(1+\exp(\theta))$	$-\log(-\theta)$
$a(\phi)$	σ^2	1	$\frac{1}{m}$	$\frac{1}{\nu}$
$c(y,\phi)$	$-\frac{1}{2}\log(2\pi\sigma^2) - \frac{y^2}{2\sigma^2}$	$-\log(y!)$	$\log\binom{m}{my}$	$\nu\log(\nu y) - \log y - \log\Gamma(\nu)$
$\mu = \dot{b}(\theta)$	θ	$\exp(\theta)$	$\mathrm{expit}(\theta) = \frac{\exp(\theta)}{\exp(\theta)+1}$	$-\frac{1}{\theta}$
$V(\mu)$	1	μ	$\pi(1-\pi)$	μ^2

Table 2: Commonly Used Exponential Families

[‡]The mean-value parameter is denoted by π in place of μ for the scaled binomial distribution.

§17 Justification of Exponential Families & Link Functions (2020/5/13)

One may attempt to extend a fixed (standard) distribution to a family, e.g., $\mathcal{N}(0,1)$ to $\mathcal{N}(\mu, \sigma^2)$. Let a random variable X be drawn from a given probability density/mass function $f_0(\bullet)$, denoted by $X \sim f_0(\bullet)$.

- (location-scale family) Clearly $\sigma X + \mu \sim \frac{1}{\sigma} f_0(\frac{\bullet - \mu}{\sigma})$, whose support takes in an affine transformation. Hence this is not applicable to data with constraints on the support, especially when the data are discrete, say, counts.
- (exponential tilting family) Let's stick to keeping the support unchanged. Suppose that the moment generating function

  $$M_0(t) = \mathbb{E}\exp(tX) = \int \exp(tx) f_0(x)\,\mathrm{d}x$$

  takes values in $(0, \infty)$ for $t \in \mathbb{R}$ in a suitable domain (including at least an open neighborhood of 0). Write the **cumulant generating function** (c.g.f.) as

  $$K_0(t) = \log M_0(t),$$

  whose r^{th} derivative at 0 is called the r^{th} **cumulant**:

  $$\kappa_r = K_0^{(r)}(0), \quad r \in \mathbb{N},$$

  which is similar to the r^{th} moment $\mu_r = M_0^{(r)}(0)$. As is well known, the c.g.f. is linear in independent random variables. We may normalize $\exp(\theta\bullet)f_0(\bullet)$ to obtain a new probability density/mass function

  $$f_\theta(\bullet) = \exp(\theta\bullet) f_0(\bullet)/M_0(\theta) = \exp\{\theta\bullet - K_0(\theta)\} f_0(\bullet),$$

  where multiplying $\exp\{\theta\bullet - K_0(\theta)\}$ is said to be an **exponential tilting** operator.

Define $M_\theta(\bullet) = M_0(\theta+\bullet)/M_0(\theta)$ and $K_\theta(\bullet) = \log M_\theta(\bullet) = K_0(\theta+\bullet) - K_0(\theta)$. It's straightforward that

$$K_\theta^{(r)}(0) = K_0^{(r)}(\theta),$$

so $K_0(\theta)$ is essential and is often referred to as the c.g.f. of the whole family $\{f_\theta(\bullet)\}$. Note that

$$f_\theta(y) = \exp\{y\theta - K_0(\theta) + \log f_0(y)\}$$

is a special case of

$$f_{\theta,\phi}(y) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\},$$

where the dispersion parameter ϕ reflects the departure of data from the model.

For example, let $f_0 = \mathbb{1}_{[0,1]}$ be the density function of Uniform$([0,1])$. One can see that

$$M_0(t) = \begin{cases} \frac{1}{t}(\mathrm{e}^t - 1), & t > 0 \\ 1, & t = 0 \end{cases} \implies K_0(t) = \begin{cases} \log(\mathrm{e}^t - 1) - \log t, & t > 0 \\ 0, & t = 0 \end{cases}$$

which gives

$$f_\theta(\bullet) = \theta \exp(\theta\bullet)/(\mathrm{e}^\theta - 1) \cdot \mathbb{1}_{[0,1]}.$$

In the GLM, a non-linear function $g(\cdot)$ links $\mu = \mathbb{E}[y]$ to $\eta = \boldsymbol{x}'\boldsymbol{\beta}$ via $\eta = g(\mu)$, which is strictly increasing without loss of generality. Commonly used link functions include logit, probit, complementary log-log, and the *power-family* (Box-Cox type) link

$$\eta = \begin{cases} (\mu^\lambda - 1)/\lambda, & \lambda > 0; \\ \log\mu, & \lambda = 0. \end{cases}$$

At least theoretically, canonical links can be justified by the *sufficiency principle*[xvii]. Given data $(\boldsymbol{x}_i, y_i)_{1\le i\le n}$, we treat covariates $\boldsymbol{x}_i$'s as fixed, and suppose the nuisance parameter ϕ is constant while the known weights w_i's vary across observations. The density function is

$$f(y_1,\dots,y_n;\boldsymbol{\beta},\phi) = \exp\left\{\sum_{i=1}^n \frac{y_i\theta_i - b(\theta_i)}{\phi/w_i} + \sum_{i=1}^n c_i(y_i,\phi)\right\},$$

where $\theta_i = \dot b^{-1}(\mu_i)$ for $\mu_i = g^{-1}(\eta_i) = g^{-1}(\boldsymbol{x}_i'\boldsymbol{\beta})$. Pretending that ϕ is given, one may yearn for a concise sufficient statistic of $\boldsymbol{\beta}$. If we choose $g = \dot b^{-1}$, it follows that $\theta_i = \eta_i = \boldsymbol{x}_i'\boldsymbol{\beta}$, and then

$$f(y_1,\dots,y_n;\boldsymbol{\beta},\phi) = \exp\left\{\frac{1}{\phi}\sum_{i=1}^n w_i y_i \boldsymbol{x}_i'\boldsymbol{\beta} - \frac{1}{\phi}\sum_{i=1}^n w_i b(\boldsymbol{x}_i'\boldsymbol{\beta}) + \sum_{i=1}^n c_i(y_i,\phi)\right\}.$$

Therefore, $\sum_{i=1}^n w_i y_i \boldsymbol{x}_i$ is sufficient for $\boldsymbol{\beta}$ by the factorization theorem. ✍Note that the canonical link is not always the best choice. For example, in gamma regression, the canonical link producing reciprocals $\frac{1}{\mu_i}$ results in unstable computations, and thus is not preferred.

§18 Goodness of Fit & Algorithms for Model Estimation (2020/5/20)

Suppose that a sample $\{y_i\}_{1\le i\le n}$ is drawn independently from the density functions

$$f_i(y_i;\boldsymbol{\beta},\phi) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{\phi/w_i} + c_i(y_i,\phi)\right\},$$

where $\theta_i = \dot b^{-1}(\mu_i)$ for $\mu_i = g^{-1}(\eta_i) = g^{-1}(\boldsymbol{x}_i'\boldsymbol{\beta})$. It can be shown that the estimate $\hat{\boldsymbol{\beta}}$ is free of ϕ, whose details are deferred until after model checking. Let

$$\hat\mu_i = g^{-1}(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}})$$

be the **fitted values**.

[xvii] cf. https://zhuanlan.zhihu.com/p/102499608 & https://zhuanlan.zhihu.com/p/103110033

- (Pearson's chi-square) Recall that $\mathrm{Var}(y_i) = V(\mu_i)\phi/w_i$, where $V(\cdot) = \ddot{b}(\dot{b}^{-1}(\cdot))$ is the variance function. Ignoring ϕ, the standardized error

$$r_{i,P} = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)/w_i}}$$

is called the **Pearson residual**, whose sum of squares

$$\chi^2 = \sum_{i=1}^{n} r_{i,P}^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)/w_i}$$

is known as the **Pearson chi-squared statistic**, measuring globally the goodness of fit. Under suitable regularity conditions, if the model is correct, i.e., fits the data, then

$$\chi^2/\phi \overset{\circ}{\sim} \chi^2_{n-p},$$

where p is the dimension of $\boldsymbol{\beta}$. Sometimes χ^2/ϕ is said to be the *scaled Pearson's chi-square*. Now that $\mathbb{E}\chi^2/\phi \approx n - p$, we may estimate the dispersion parameter ϕ as

$$\hat{\phi} = \chi^2/(n-p),$$

which reflects how the prescribed distribution fits the data. This data-driven estimate of ϕ, instead of the default value (e.g., 1), introduces additional flexibility of the model.

For n independent observations $\boldsymbol{y} = (y_i)_{1\le i\le n}$, let

$$\ell_n(\boldsymbol{\mu}; \boldsymbol{y}) = \sum_{i=1}^{n} \frac{y_i \dot{b}^{-1}(\mu_i) - b(\dot{b}^{-1}(\mu_i))}{\phi/w_i} + \sum_{i=1}^{n} c_i(y_i, \phi)$$

be the log-likelihood function of $\boldsymbol{\mu} = (\mu_i)_{1\le i\le n}$. Considered for all possible models, the maximum achievable log-likelihood is $\ell_n(\boldsymbol{y}; \boldsymbol{y})$. This occurs for the *saturated model*, having a separate parameter for each observation and the perfect fit, which sounds good but is not helpful. The saturated model does not smooth the data or have the advantages that a simpler model has because of its parsimony, such as better estimation of the true relation. However, it often serves as a baseline for comparison with other model fits, such as for checking goodness of fit.

- (deviance) The statistic

$$D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) = 2\left[\ell_n(\boldsymbol{y}; \boldsymbol{y}) - \ell_n(\hat{\boldsymbol{\mu}}; \boldsymbol{y})\right] \cdot \phi = 2\sum_{i=1}^{n} w_i \left\{ y_i \left(\tilde{\theta}_i - \hat{\theta}_i\right) - \left[b(\tilde{\theta}_i) - b(\hat{\theta}_i)\right]\right\}$$

is called the **deviance**, where $\tilde{\theta}_i = \dot{b}^{-1}(y_i)$ and $\hat{\theta}_i = \dot{b}^{-1}(\hat{\mu}_i)$. The greater the deviance, the poorer the fit. Under suitable regularity conditions, if the fitted model is correct, the *scaled deviance*

$$D(\boldsymbol{y}, \hat{\boldsymbol{\mu}})/\phi \overset{\circ}{\sim} \chi^2_{n-p},$$

where p is the dimension of $\boldsymbol{\beta}$. One may estimate the dispersion parameter ϕ as

$$\hat{\phi} = D(\boldsymbol{y}, \hat{\boldsymbol{\mu}})/(n-p).$$

The **deviance residual** is defined to be

$$r_{i,D} = \mathrm{sign}(y_i - \hat{\mu}_i)\sqrt{d_i},$$

where

$$d_i = 2w_i\{y_i(\tilde{\theta}_i - \hat{\theta}_i) - [b(\tilde{\theta}_i) - b(\hat{\theta}_i)]\},$$

whose sum $\sum_{i=1}^{n} d_i$ equals the deviance $D(\boldsymbol{y}, \hat{\boldsymbol{\mu}})$.

Here are some remarks:

i) Pearson's chi-square is more intuitive and interpretable.

ii) The deviance, as a measure of discrepancy, is additive for nested models. Consider a full model with p_f parameters and fitted values $\hat{\boldsymbol{\mu}}_f$ and a reduced model with p_r parameters and fitted values $\hat{\boldsymbol{\mu}}_r$. If the reduced model holds, we have

$$[D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}_r) - D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}_f)]/\phi = 2\left[\ell_n(\hat{\boldsymbol{\mu}}_f; \boldsymbol{y}) - \ell_n(\hat{\boldsymbol{\mu}}_r; \boldsymbol{y})\right] \overset{\circ}{\sim} \chi^2_{p_f - p_r}.$$

iii) To conduct the nested likelihood-ratio test,

- ◇ if the dispersion is not obvious, we may pretend that $\phi = 1$, and use $D_r - D_f \sim \chi^2_{p_f - p_r}$.
- ◇ if the dispersion cannot be ignored, we may use $\frac{(D_r - D_f)/(p_f - p_r)}{D_f/(n - p_f)} \sim F_{p_f - p_r, n - p_f}$. Note that overfitting will not substantially reduce D_f, and thus $\hat{\phi}_f = D_f/(n - p_f)$ is a reasonable estimate for ϕ.

iv) We should not use χ^2 or D itself to check the goodness of fit, since the dispersion caused by ϕ cannot be ignored and the asymptotic behavior may not be guaranteed. For more robustness, we turn to check the *relative discrepancy* $|\chi^2 - D|/D$ of a considered model, which can be adapted to compare non-nested models.

v) As for model selection, a GLM consists of not only covariates, but also a link function and a family of distributions. The distribution is difficult to assign a criterion. When it comes to covariates and links, apart from commonly used AIC and BIC, we may incorporate the deviance and Pearson's chi-square.

For non-normal data y_i's, we hope to find some function $A(\cdot)$ so that the distribution of $A(y_i)$ is as normal as possible. Using $A(y_i) \approx A(\mu_i) + \dot{A}(\mu_i)(y_i - \mu_i)$, we obtain the **variance-stabilizing transformation**

$$\operatorname{Var}(A(y_i)) \approx \dot{A}(\mu_i)^2 \operatorname{Var}(y_i),$$

where $\operatorname{Var}(y_i) = V(\mu_i)\phi/w_i$, and thus $A(\mu) = \int_0^\mu V(t)^{-1/2}\,\mathrm{d}t$ seems a good choice. A *transformed residual* is given by

$$r_{i,A} = \frac{A(y_i) - A(\hat{\mu}_i)}{\sqrt{\dot{A}(\hat{\mu}_i)^2 V(\hat{\mu}_i)/w_i}},$$

which is called the **Anscombe residual** when $A(\mu) = \int_0^\mu V(t)^{-1/3}\,\mathrm{d}t$.

Next, we get down to the estimation of $\boldsymbol{\beta} \in \mathbb{R}^p$. Let $\ell(\boldsymbol{\beta})$ be the log-likelihood function, and denote its gradient (the score function) and Hessian by $\boldsymbol{u}(\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}}\ell(\boldsymbol{\beta})$ and $\boldsymbol{H}(\boldsymbol{\beta}) = \frac{\partial^2}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}'}\ell(\boldsymbol{\beta})$, respectively. Generally, the MLE $\hat{\boldsymbol{\beta}}$ is a solution to the estimating equation

$$\mathbf{0}_p = \boldsymbol{u}(\hat{\boldsymbol{\beta}}) \approx \boldsymbol{u}(\boldsymbol{\beta}^{(0)}) + \boldsymbol{H}(\boldsymbol{\beta}^{(0)})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{(0)}),$$

where $\boldsymbol{\beta}^{(0)} \in \mathbb{R}^p$ denotes the initial value. In the iterative process for updating $\boldsymbol{\beta}^{(t)}$, $t = 0, 1, 2, \ldots$, the **Newton-Raphson** method adopts

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - \boldsymbol{H}(\boldsymbol{\beta}^{(t)})^{-1}\boldsymbol{u}(\boldsymbol{\beta}^{(t)}).$$

To stabilize and accelerate the iteration, one may replace the *observed Fisher information* $-\boldsymbol{H}(\boldsymbol{\beta})$ with the expected Fisher information $\mathcal{I}(\boldsymbol{\beta}) = -\mathbb{E}\boldsymbol{H}(\boldsymbol{\beta})$. Indeed, **Fisher's scoring** method adopts

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + \mathcal{I}(\boldsymbol{\beta}^{(t)})^{-1}\boldsymbol{u}(\boldsymbol{\beta}^{(t)}).$$

In a GLM, $\ell = \sum_{i=1}^n \ell_i$, where

$$\ell_i = [y_i\theta_i - b(\theta_i)]/a_i(\phi) + c_i(y_i, \phi),$$

and a function $g(\cdot)$ links $\mu_i = \mathbb{E}[y_i] = \dot{b}(\theta_i)$ and $\eta_i = \boldsymbol{x}_i'\boldsymbol{\beta} = \sum_{r=1}^p \beta_r x_{ir}$ via $\eta_i = g(\mu_i)$. To calculate

$$u_{ir} = \frac{\partial \ell_i}{\partial \beta_r} = \frac{\partial \ell_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_r},$$

note that

- ◇ $\frac{\partial \ell_i}{\partial \theta_i} = \frac{y_i - \dot{b}(\theta_i)}{a_i(\phi)} = \frac{y_i - \mu_i}{a_i(\phi)}$
- ◇ $\frac{\partial \theta_i}{\partial \mu_i} = 1/\frac{\partial \mu_i}{\partial \theta_i} = 1/\ddot{b}(\theta_i) = 1/V(\mu_i)$
- ◇ $\frac{\partial \mu_i}{\partial \eta_i} = 1/\frac{\partial \eta_i}{\partial \mu_i} = 1/\dot{g}(\mu_i)$
- ◇ $\frac{\partial \eta_i}{\partial \beta_r} = x_{ir}$

$$\Longrightarrow \qquad u_{ir} = \frac{(y_i - \mu_i)x_{ir}}{a_i(\phi)V(\mu_i)\dot{g}(\mu_i)}.$$

Clearly $\boldsymbol{u}(\boldsymbol{\beta}) = (u_r)^{1\le r\le p}$, where $u_r = \sum_{i=1}^n u_{ir}$. It follows that the (r,s)-entry of $\mathcal{I}(\boldsymbol{\beta}) = \mathrm{Var}(\boldsymbol{u}(\boldsymbol{\beta}))$ is

$$\mathrm{Cov}(u_r, u_s) = \sum_{i=1}^n \frac{\mathrm{Var}(y_i) x_{ir} x_{is}}{a_i(\phi)^2 V(\mu_i)^2 \dot{g}(\mu_i)^2} = \sum_{i=1}^n \frac{x_{ir} x_{is}}{a_i(\phi) V(\mu_i) \dot{g}(\mu_i)^2}.$$

Let $\boldsymbol{X} = (x_{ir})_{1\le r\le p}^{1\le i\le n}$, and $\boldsymbol{W}$ be the diagonal matrix with main-diagonal elements $w_{ii} = \frac{1}{a_i(\phi)V(\mu_i)\dot{g}(\mu_i)^2}$. Then it's immediate that

$$\mathcal{I}(\boldsymbol{\beta}) = \boldsymbol{X}'\boldsymbol{W}\boldsymbol{X}.$$

Since $u_{ir} = (y_i - \mu_i) w_{ii} \dot{g}(\mu_i) x_{ir}$, we have

$$u_r = (\boldsymbol{y} - \boldsymbol{\mu})' \boldsymbol{W} \tfrac{\mathrm{d}\boldsymbol{\eta}}{\mathrm{d}\boldsymbol{\mu}} \boldsymbol{x}_{(r)},$$

where $\boldsymbol{x}_{(r)} = (x_{ir})^{1\le i\le n}$ is the r^{th} column of $\boldsymbol{X}$, and $\frac{\mathrm{d}\boldsymbol{\eta}}{\mathrm{d}\boldsymbol{\mu}}$ is the diagonal matrix with diagonal entries $\dot{g}(\mu_i)$. Thus,

$$\boldsymbol{u}(\boldsymbol{\beta}) = \boldsymbol{X}'\boldsymbol{W} \tfrac{\mathrm{d}\boldsymbol{\eta}}{\mathrm{d}\boldsymbol{\mu}} (\boldsymbol{y} - \boldsymbol{\mu}),$$

where diagonal matrices $\boldsymbol{W}$ and $\frac{\mathrm{d}\boldsymbol{\eta}}{\mathrm{d}\boldsymbol{\mu}}$ are commutative. One can see that

$$\tfrac{\partial}{\partial \boldsymbol{\beta}'}(\boldsymbol{y} - \boldsymbol{\mu}) = -\tfrac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\eta}} \tfrac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\beta}'} = -(\tfrac{\mathrm{d}\boldsymbol{\eta}}{\mathrm{d}\boldsymbol{\mu}})^{-1} \boldsymbol{X},$$

so

$$\begin{aligned} \boldsymbol{H}(\boldsymbol{\beta}) = \tfrac{\partial \boldsymbol{u}}{\partial \boldsymbol{\beta}'}(\boldsymbol{\beta}) &= \boldsymbol{X}'\boldsymbol{W} \tfrac{\mathrm{d}\boldsymbol{\eta}}{\mathrm{d}\boldsymbol{\mu}} \tfrac{\partial}{\partial \boldsymbol{\beta}'}(\boldsymbol{y} - \boldsymbol{\mu}) + \boldsymbol{X}'[\tfrac{\partial}{\partial \boldsymbol{\beta}'}(\boldsymbol{W} \tfrac{\mathrm{d}\boldsymbol{\eta}}{\mathrm{d}\boldsymbol{\mu}})](\boldsymbol{y} - \boldsymbol{\mu}) \\ &= -\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X} + \boldsymbol{X}'[\tfrac{\partial}{\partial \boldsymbol{\beta}'}(\boldsymbol{W} \tfrac{\mathrm{d}\boldsymbol{\eta}}{\mathrm{d}\boldsymbol{\mu}})](\boldsymbol{y} - \boldsymbol{\mu}), \end{aligned}$$

which certainly satisfies that

$$\mathbb{E}\boldsymbol{H}(\boldsymbol{\beta}) + \mathcal{I}(\boldsymbol{\beta}) = \boldsymbol{X}'[\tfrac{\partial}{\partial \boldsymbol{\beta}'}(\boldsymbol{W} \tfrac{\mathrm{d}\boldsymbol{\eta}}{\mathrm{d}\boldsymbol{\mu}})](\mathbb{E}\boldsymbol{y} - \boldsymbol{\mu}) = \boldsymbol{0}_{p\times p}.$$

It's partly due to the much simpler form of $\mathcal{I}(\boldsymbol{\beta})$ than $-\boldsymbol{H}(\boldsymbol{\beta})$ that we prefer Fisher's scoring algorithm to the Newton-Raphson algorithm. These two methods coincide when the canonical link is used. To see this, if $\eta_i = \dot{b}^{-1}(\mu_i) = \theta_i$, then

$$V(\mu_i) = \ddot{b}(\theta_i) = \frac{\partial \mu_i}{\partial \theta_i} = \frac{\partial \mu_i}{\partial \eta_i} \implies w_{ii} = \frac{1}{a_i(\phi) V(\mu_i) (\partial \eta_i / \partial \mu_i)^2} = \frac{1}{a_i(\phi)(\partial \eta_i / \partial \mu_i)},$$

in which case $\boldsymbol{W} \frac{\mathrm{d}\boldsymbol{\eta}}{\mathrm{d}\boldsymbol{\mu}}$ is a diagonal matrix with elements $\frac{1}{a_i(\phi)}$ not depending on $\boldsymbol{\beta}$, and thus $-\boldsymbol{H}(\boldsymbol{\beta}) = \mathcal{I}(\boldsymbol{\beta})$. It can also be seen from that

$$u_{ir} = \frac{\partial \ell_i}{\partial \theta_i} \frac{\partial \theta_i}{\cancel{\partial \eta_i}} \frac{\partial \eta_i}{\partial \beta_r} = \frac{y_i - \mu_i}{a_i(\phi)} x_{ir}.$$

There exists a relation between Fisher's scoring and weighted least squares. Note that

$$(g(\mu_i))^{1\le i\le n} = \boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}.$$

Expanding $g(y_i)$ around μ_i, define the *working response*

$$\boldsymbol{z} = \boldsymbol{\eta} + \tfrac{\mathrm{d}\boldsymbol{\eta}}{\mathrm{d}\boldsymbol{\mu}}(\boldsymbol{y} - \boldsymbol{\mu}) = (\eta_i + \dot{g}(\mu_i)(y_i - \mu_i))^{1\le i\le n}$$

as a first approximation. Note that $\mathrm{Var}(\boldsymbol{z}) = \frac{\mathrm{d}\boldsymbol{\eta}}{\mathrm{d}\boldsymbol{\mu}} \mathrm{Var}(\boldsymbol{y}) \frac{\mathrm{d}\boldsymbol{\eta}}{\mathrm{d}\boldsymbol{\mu}} = \boldsymbol{W}^{-1}$. The estimating equation

$$\boldsymbol{u}(\boldsymbol{\beta}) = \boldsymbol{X}'\boldsymbol{W}(\boldsymbol{z} - \boldsymbol{X}\boldsymbol{\beta}) = \boldsymbol{0}_p$$

is equivalent to

$$\min\ (\boldsymbol{z} - \boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{W}(\boldsymbol{z} - \boldsymbol{X}\boldsymbol{\beta}) \quad \text{w.r.t.} \quad \boldsymbol{\beta} \in \mathbb{R}^p,$$

pretending that $\boldsymbol{z}$ and $\boldsymbol{W}$ do not depend on $\boldsymbol{\beta}$. Using

$$\mathcal{I}(\boldsymbol{\beta}^{(t)}) = \boldsymbol{X}'\boldsymbol{W}^{(t)}\boldsymbol{X}, \quad \& \quad \boldsymbol{u}(\boldsymbol{\beta}^{(t)}) = \boldsymbol{X}'\boldsymbol{W}^{(t)}(\boldsymbol{z}^{(t)} - \boldsymbol{X}\boldsymbol{\beta}^{(t)}),$$

Fisher's scoring yields

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + \mathcal{I}(\boldsymbol{\beta}^{(t)})^{-1} \boldsymbol{u}(\boldsymbol{\beta}^{(t)}) = (\boldsymbol{X}'\boldsymbol{W}^{(t)}\boldsymbol{X})^{-1} \boldsymbol{X}'\boldsymbol{W}^{(t)}\boldsymbol{z}^{(t)}.$$

The process is called **iteratively re-weighted least squares** (IRLS/IWLS). The initial value is usually chosen as

$$\boldsymbol{\beta}^{(0)} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{z}^{(-1)},$$

where $\boldsymbol{z}^{(-1)}$ consists of $z_i^{(-1)} = g(y_i + \delta_i)$ with small δ_i, e.g., $y_i + \delta_i = \frac{m_i y_i + 0.5}{m_i + 1}$ for $y_i \sim \frac{1}{m_i}\mathrm{Binomial}(m_i, \pi_i)$.

§19 Fitting and Inference Cont'd & Important Examples (2020/5/22)

In summary, the iterative process of Fisher's scoring algorithm is implemented as follows[xviii]. Given $\boldsymbol{\beta}^{(t)}$,

$$\boldsymbol{\eta}^{(t)} = \boldsymbol{X}\boldsymbol{\beta}^{(t)} \ \to \ \boldsymbol{\mu}^{(t)} = g^{-1}(\boldsymbol{\eta}^{(t)}) \ \to \ \frac{\mathrm{d}\boldsymbol{\eta}}{\mathrm{d}\boldsymbol{\mu}}\Big|_{\boldsymbol{\mu}^{(t)}} = \mathrm{diag}\Big(\dot{g}(\mu_i^{(t)})\Big) \ \to$$

$$\to \ \boldsymbol{z}^{(t)} = \boldsymbol{\eta}^{(t)} + \frac{\mathrm{d}\boldsymbol{\eta}}{\mathrm{d}\boldsymbol{\mu}}\Big|_{\boldsymbol{\mu}^{(t)}}(\boldsymbol{y} - \boldsymbol{\mu}^{(t)}), \quad \boldsymbol{W}^{(t)} = \mathrm{diag}\Big(1\Big/\Big[a_i(\phi)V\big(\mu_i^{(t)}\big)\dot{g}\big(\mu_i^{(t)}\big)^2\Big]\Big) \ \to$$

$$\to \ \boldsymbol{\beta}^{(t+1)} = (\boldsymbol{X}'\boldsymbol{W}^{(t)}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}^{(t)}\boldsymbol{z}^{(t)}.$$

Particularly, the dispersion parameter ϕ doesn't affect the estimation of $\boldsymbol{\beta}$ if $a_i(\phi) = \phi/w_i$. To see this, let

$$\widetilde{\boldsymbol{W}} = \phi\boldsymbol{W} = \mathrm{diag}\big(w_i/\big[V(\mu_i)\dot{g}(\mu_i)^2\big]\big)\,,$$

then

$$\boldsymbol{\beta}^{(t+1)} = \big(\boldsymbol{X}'\widetilde{\boldsymbol{W}}^{(t)}\boldsymbol{X}\big)^{-1}\boldsymbol{X}'\widetilde{\boldsymbol{W}}^{(t)}\boldsymbol{z}^{(t)}$$

has nothing to do with ϕ.

For inference, note that the MLE $\hat{\boldsymbol{\beta}}$ satisfies that

$$\mathcal{I}(\boldsymbol{\beta})^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \overset{\circ}{\sim} \mathcal{N}_p(\mathbf{0}_p, \boldsymbol{I}_p),$$

and thus we may estimate

$$\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}) = \mathcal{I}(\hat{\boldsymbol{\beta}})^{-1} = (\boldsymbol{X}'\boldsymbol{W}(\hat{\boldsymbol{\beta}})\boldsymbol{X})^{-1} = \phi\cdot(\boldsymbol{X}'\widetilde{\boldsymbol{W}}(\hat{\boldsymbol{\beta}})\boldsymbol{X})^{-1},$$

which is evaluated at $\hat{\boldsymbol{\beta}}$. Clearly, a larger ϕ results in larger $\mathrm{Var}(\hat{\boldsymbol{\beta}})$, leading further to wider confidence intervals and less significance, which means more type II error (false negative). Suppose that we are interested in $h(\boldsymbol{\beta}) \in \mathbb{R}$. Using the delta method, one can see that

$$h(\hat{\boldsymbol{\beta}}) \overset{\circ}{\sim} \mathcal{N}(h(\boldsymbol{\beta}), \dot{h}(\boldsymbol{\beta})'\mathcal{I}(\boldsymbol{\beta})^{-1}\dot{h}(\boldsymbol{\beta})),$$

and the corresponding standard error and confidence interval are easy to derive. Moreover, if $h(\boldsymbol{\beta}) = \tilde{h}(\eta)$, where $\eta = \boldsymbol{x}'\boldsymbol{\beta}$ and $\tilde{h}(\cdot)$ is increasing, which is often encountered in survival analysis, then a confidence interval $[\hat{L}, \hat{U}]$ for η can be transformed into $[\tilde{h}(\hat{L}), \tilde{h}(\hat{U})]$ to be a confidence interval for $h(\boldsymbol{\beta})$. Note that *extrapolation* involving non-linear relation is subtle and may be problematic, so one should be careful about

$$\hat{\eta} = \boldsymbol{x}'\hat{\boldsymbol{\beta}} \overset{\circ}{\sim} \mathcal{N}(\boldsymbol{x}'\boldsymbol{\beta}, \boldsymbol{x}'\mathcal{I}(\boldsymbol{\beta})^{-1}\boldsymbol{x}).$$

The converse problem of setting confidence intervals for $\boldsymbol{x}$ that gives rise to a specified mean response μ_0 is usually accomplished using **Fieller's method**. With $g(\mu_0) = \boldsymbol{x}'\boldsymbol{\beta}$ in mind, a desired asymptotic $(1-\alpha)$ confidence set can be chosen as

$$\big\{\boldsymbol{x} : \big|\boldsymbol{x}'\hat{\boldsymbol{\beta}} - g(\mu_0)\big|/v(\boldsymbol{x}) < z_{\frac{\alpha}{2}}\big\},$$

where $v(\boldsymbol{x}) = \sqrt{\boldsymbol{x}'\mathcal{I}(\hat{\boldsymbol{\beta}})^{-1}\boldsymbol{x}}$, and $z_{\frac{\alpha}{2}} = \Phi^{-1}(1-\frac{\alpha}{2})$.

Next, we will discuss some frequently used GLMs. For practical examples and codes, please refer to Venables, W.N., & Ripley, B.D. (2002). Generalized Linear Models. In: *Modern Applied Statistics with S*. (https://doi.org/10.1007/978-0-387-21706-2_7)

- (Binomial) Link functions commonly used are
  (1) $\mathrm{logit}: \pi \mapsto \log\frac{\pi}{1-\pi}$;
  (2) $\mathrm{probit}: \pi \mapsto \Phi^{-1}(\pi)$;
  (3) complementary log-log: $\pi \mapsto \log(-\log(1-\pi))$; and
  (4) log-log: $\pi \mapsto -\log(-\log\pi)$.

  See fig. 4. Here are some reasons why the logit function is the most popular choice. First, it's the canonical link. Second, it provides an intuitive interpretation with respect to the **odds** $\frac{\pi}{1-\pi}$. Last but not least, it applies to retrospective/case-control study (see §20).

[xviii] R Example: https://stats.stackexchange.com/a/344672

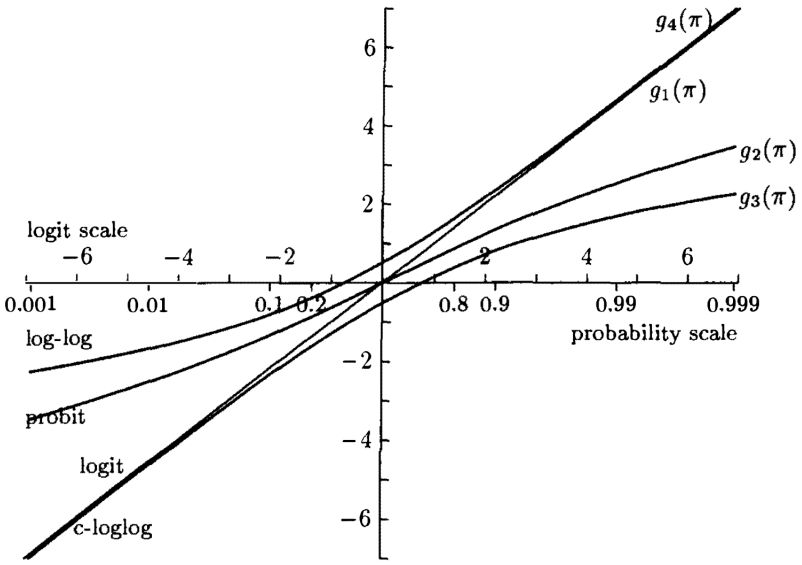

Figure 4: A Graphical Comparison of Link Functions with Logit[xix)]

Sometimes the estimated dispersion parameter $\hat{\phi}$ exceeds the default value. Thus, we may wonder a mechanism of *over-dispersion*. The simplest, and perhaps the most common mechanism, is *clustering* in the population. Consider grouped data, where the i^{th} group is characterized by a proportion y_i out of n_i subjects. Assume that $n_i y_i | p_i \sim \text{Binomial}(n_i, p_i)$, where the response probability p_i is a random variable such that $\mathbb{E}[p_i] = \pi_i$ and $\text{Var}(p_i) = \tau\pi_i(1-\pi_i)$. It follows that $\mathbb{E}[n_i y_i] = n_i \mathbb{E}[p_i] = n_i\pi_i$, and

$$\begin{aligned}\text{Var}(n_i y_i) &= \mathbb{E}[\text{Var}(n_i y_i|p_i)] + \text{Var}(\mathbb{E}[n_i y_i|p_i]) \\ &= \mathbb{E}[n_i p_i(1-p_i)] + \text{Var}(n_i p_i) \\ &= n_i\pi_i - n_i[\pi_i^2 + \tau\pi_i(1-\pi_i)] + n_i^2\tau\pi_i(1-\pi_i) \\ &= [1+(n_i-1)\tau]\cdot n_i\pi_i(1-\pi_i),\end{aligned}$$

where $1+(n_i-1)\tau$ stands for the dispersion. Note that when over-dispersion matters, the likelihood is no longer exact, and we need a more general framework, e.g., based on quasi-likelihood.

- (Poisson) The Poisson distribution is often used for *counts* of events that occur randomly over time or space at a particular rate, when outcomes in disjoint time periods or regions are independent. In Poisson regression, a central feature is that the variance function is $V(\mu) = \mu$. Note that *over-dispersion* arises quite a lot. The data might be produced by a **clustered Poisson process**, i.e., we observe

  $$Y = Z_1 + Z_2 + \cdots + Z_N,$$

  where the Z's are i.i.d. and N has a Poisson distribution independent of Z. For example,
  - (insurance claims) Z = claim amount in an accident, and N = # of accidents.
  - (ecology, line transect sampling) Z = # of animals in a cluster, and N = # of clusters.

  It can be seen that $\mathbb{E}[Y] = \mathbb{E}[N]\mathbb{E}[Z]$ and

  $$\text{Var}(Y) = \mathbb{E}[N]\,\text{Var}(Z) + \text{Var}(N)(\mathbb{E}[Z])^2 = \mathbb{E}[N]\mathbb{E}[Z^2],$$

  so there exists over-dispersion if $\mathbb{E}[Z^2] > \mathbb{E}[Z]$.

- (Gamma) Recall that the (canonical) inverse link $\theta = \dot{b}^{-1}(\mu) = -\frac{1}{\mu}$ is hardly used, and instead the power-family (Box-Cox type) link is preferred. The central feature is that the variance function is $V(\mu) = \mu^2$. Hence, other models with *constant* **coefficient of variation**[xx)] $\sqrt{\text{Var}(Y)}/\mathbb{E}[Y]$ can also be handled using gamma regression. Some examples are shown below.

[xix)] Fig. 4.1 in McCullagh & Nelder [McCN]
[xx)] cf. https://stats.stackexchange.com/q/118497

– (**multiplicative error model**)

$$Y = \mu(1+\varepsilon), \quad \mathbb{E}[\varepsilon] = 0, \ \& \ \mathrm{Var}(\varepsilon) = \sigma^2.$$

Clearly $\mathbb{E}[Y] = \mu$ and $\mathrm{Var}(Y) = \mu^2\sigma^2$, so $\mathrm{Var}(Y)/(\mathbb{E}[Y])^2$ doesn't depend on μ.

– (**log-transformed additive model**)

$$\log Y = \mu + \varepsilon, \quad \mathbb{E}[\varepsilon] = 0, \ \& \ \mathrm{Var}(\varepsilon) = \sigma^2.$$

Clearly $Y = \mathrm{e}^{\mu}\mathrm{e}^{\varepsilon}$ has $\mathbb{E}[Y] = \mathrm{e}^{\mu}\mathbb{E}[\mathrm{e}^{\varepsilon}]$ and $\mathrm{Var}(Y) = \mathrm{e}^{2\mu}\,\mathrm{Var}(\mathrm{e}^{\varepsilon})$, so

$$\mathrm{Var}(Y)/(\mathbb{E}[Y])^2 = \mathrm{Var}(\mathrm{e}^{\varepsilon})/(\mathbb{E}[\mathrm{e}^{\varepsilon}])^2 \approx \mathrm{Var}(1+\varepsilon)/(\mathbb{E}[1+\varepsilon])^2 = \sigma^2,$$

provided that σ^2 is small. The case $\varepsilon \sim \mathcal{N}(0,\sigma^2)$ is called *log-normal model*.
As an alternative, to fit directly the log-transformed model (and other models with *constant* coefficient of variation), we can expand

$$\log Y = \log(\mathbb{E}[Y]) + \frac{1}{\mathbb{E}[Y]}(Y - \mathbb{E}[Y]) - \frac{1}{2}\frac{1}{(\mathbb{E}[Y])^2}(Y-\mathbb{E}[Y])^2 + \cdots.$$

It follows that

$$\mathbb{E}[\log Y] \approx \log(\mathbb{E}[Y]) - \frac{1}{2}\frac{\mathrm{Var}(Y)}{(\mathbb{E}[Y])^2}, \quad \& \quad \mathrm{Var}(\log Y) \approx \frac{\mathrm{Var}(Y)}{(\mathbb{E}[Y])^2}.$$

In conclusion, log-transform stabilizes variance for data with constant coefficient of variation, but produces an offset in $\mathbb{E}[\log Y]$. Practically, we often fit a linear model (OLS) to the data after log-transform, where the intercept should be biased by an offset $-\frac{1}{2}\widehat{\mathrm{Var}}(\log Y)$.
✍ Notice that the Poisson regression needs no offset, since there the variance and the mean are of the same order.

§20 Applications of GLMs (2020/5/27)

One would go through Subsection 6.3.2 *A study of wave damage to cargo ships* in McCullagh & Nelder [McCN] at first, where some lessons are worth noting.

1. We have to consider the practical limitation when modeling data.
2. If there exists significant interaction, main effects might not have a clear interpretation.
3. When analyzing data, outliers are often subtle.
4. The degree of freedom could be reduced because of degenerate distributions.

Next, we discuss the dose-response experiment. A latent variable model called the **threshold model** provides motivation for GLMs. Suppose that each individual has a threshold X_i, $i = 1, 2, \cdots, n$, which are i.i.d. random variables. When we apply dose of level d, the observed outcome will be $\mathbb{1}_{\{d \geq X_i\}}$, indicating whether there is a response or not. The probability of response at dose level d is

$$\pi = \mathbb{P}(X_i \leq d) = F^*(\log d),$$

where $F^*(\cdot)$ is the distribution function of $\log X_i$. It's usually assumed that $F^*(\cdot)$ comes from a location-scale family with parameters μ and σ, i.e., $(\log X_i - \mu)/\sigma \overset{\text{i.i.d.}}{\sim} F$ for a standard distribution function $F(\cdot)$. Then

$$\pi = \mathbb{P}(\tfrac{\log X_i - \mu}{\sigma} \leq \eta) = F(\eta), \ \text{where } \eta = \tfrac{\log d - \mu}{\sigma}.$$

This gives the GLM

$$F^{-1}(\pi) = \eta = -\tfrac{\mu}{\sigma} + \tfrac{1}{\sigma}\log d = \beta_0 + \beta_1 \log d,$$

where the link function $F^{-1}(\cdot)$ has a physical interpretation. We are interested in the **effective dose level** $ED\alpha$ such that

$$\mathbb{P}(X_i \leq ED\alpha) = \tilde{\alpha} = \alpha/100,$$

and so

$$ED\alpha = \exp\{\frac{F^{-1}(\tilde{\alpha}) - \beta_0}{\beta_1}\}.$$

The methods of constructing confidence intervals for $ED\alpha$ have been shown before. Here we want to compare $ED\alpha$ of two treatments A and B. Let

$$ED_A\alpha = \exp\{\frac{F_A^{-1}(\tilde{\alpha}) - \beta_{0A}}{\beta_{1A}}\}, \quad \& \quad ED_B\alpha = \exp\{\frac{F_B^{-1}(\tilde{\alpha}) - \beta_{0B}}{\beta_{1B}}\}.$$

Define the **response quotient** as

$$q(\tilde{\alpha}) = ED_B\alpha / ED_A\alpha,$$

which stands for the strength of B relative to A to achieve the same effect size α, e.g., one can see that B is weaker than A if $q(\tilde{\alpha}) > 1$. For fit assessment, we often take $F_A = F_B$, i.e., two samples are drawn from the same population. In such a case, we have

$$q(\tilde{\alpha}) = \exp\{\frac{F^{-1}(\tilde{\alpha}) - \beta_{0B}}{\beta_{1B}} - \frac{F^{-1}(\tilde{\alpha}) - \beta_{0A}}{\beta_{1A}}\}.$$

If $\beta_{1A} = \beta_{1B}$, it can be seen that

$$q(\tilde{\alpha}) = \exp\{\frac{\beta_{0A} - \beta_{0B}}{\beta_1}\}$$

is constant for all α, which implies the so-called *log-parallel assays*. This happens only if the same substance is diluted, and $q(\tilde{\alpha}) \equiv \zeta$ is called the **dilution factor**, which entails

$$\log d_B = \log d_A + \log \zeta,$$

where $d = ED\alpha$. Note that

$$F^{-1}(\pi) = \eta = \beta_{0A} + \beta_{1A} \log d_A = \beta_{0B} + \beta_{1B} \log d_B,$$

which can also be written as

$$\eta = \beta_{0A} + (\beta_{0B} - \beta_{0A})\mathbb{1}_B + \beta_{1A} \log d + (\beta_{1B} - \beta_{1A})\mathbb{1}_B \log d.$$

To test for log-parallelism ($\beta_{1A} = \beta_{1B}$), we can combine data from two treatment groups and consider the reparametrized model

$$F^{-1}(\pi) = \gamma_0 + \gamma_1 \mathbb{1}_B + \gamma_2 \log d + \gamma_3 \mathbb{1}_B \log d,$$

then it's equivalent to test for no interaction. If we retain $H_0 : \gamma_3 = 0$, refitting the main effect model will lead to appropriate inference about $\zeta = \exp\{-\frac{\gamma_1}{\gamma_2}\}$.

Now let's consider a new topic. Many epidemiological studies have the goal of comparing distinct groups, e.g., assessing risk factors for some disease. Denote

$$D = \{\text{disease}\}, \quad \& \quad E = \{\text{exposure}\} = \{\text{risk factor}\}.$$

The frequency counts of outcomes for a sample can be summarized in a 2×2 **contingency table**:

	D	$\overline{D}$	Total
E	a	b	$a+b$
$\overline{E}$	c	d	$c+d$
Total	$a+c$	$b+d$	$a+b+c+d$

Table 3: Cross-Classification of Disease and Exposure

There are two types of sampling designs, prospective and retrospective. The former is exemplified by *cohort studies* and the latter by *case-control studies*. Prospective studies usually condition on the (row) totals for categories of the predictor (E,$\overline{E}$). Retrospective studies usually condition on the (column) totals for categories of the response (D,$\overline{D}$). In cohort studies, disease-free subjects make their own choice about whether to expose or not, and the study observes in future time who develops the disease, e.g., Framingham

Heart Study[xxi]. In case-control studies, researchers first choose subjects with or without the disease, and then compare their exposure to the risk factor. Let

$$p_e = \mathbb{P}(D|E), \quad \& \quad p_u = \mathbb{P}(D|\overline{E}).$$

The **relative risk** (exposed vs. unexposed) is defined to be

$$\rho = p_e/p_u.$$

Also, we are interested in the **odds ratio**

$$\psi = \frac{p_e/(1-p_e)}{p_u/(1-p_u)}.$$

Only in prospective/cohort studies may it be assumed that

$$a \sim \text{Binomial}(a+b, p_e) \perp\!\!\!\perp c \sim \text{Binomial}(c+d, p_u),$$

and then we can estimate

$$\hat{p}_e = \frac{a}{a+b}, \ \hat{p}_u = \frac{c}{c+d}, \ \hat{\rho} = \frac{\hat{p}_e}{\hat{p}_u}, \ \& \ \hat{\psi} = \frac{\hat{p}_e/(1-\hat{p}_e)}{\hat{p}_u/(1-\hat{p}_u)} = \frac{ad}{bc}.$$

Generally, contingency tables are too simple and we hope to incorporate more covariates, for which the *logistic regression* works. In cohort study, the indicator or the proportion of disease cases, denoted by y at covariate $\boldsymbol{x}$, can be modeled by

$$\text{logit}(\mathbb{E}[y|\boldsymbol{x}]) = \eta = \boldsymbol{x}'\boldsymbol{\beta},$$

where $\boldsymbol{\beta}$ is the unknown parameter. For example, if $\boldsymbol{x} = (1, \mathbb{1}_E)'$ and $\boldsymbol{\beta} = (\beta_0, \beta_1)'$, then

$$p_e = \frac{\exp(\beta_0+\beta_1)}{1+\exp(\beta_0+\beta_1)}, \quad \& \quad p_u = \frac{\exp(\beta_0)}{1+\exp(\beta_0)},$$

so the odds ratio

$$\psi = \frac{\exp(\beta_0+\beta_1)}{\exp(\beta_0)} = \exp(\beta_1).$$

In case-control study, one cannot estimate the risk/odds directly, because it involves unestimable sampling probability of subjects. Let S be the event that an individual is selected. To avoid sampling bias, it is assumed that

$$\pi_1 = \mathbb{P}(S|D, \boldsymbol{x}) = \mathbb{P}(S|D)$$

and

$$\pi_2 = \mathbb{P}(S|\overline{D}, \boldsymbol{x}) = \mathbb{P}(S|\overline{D})$$

don't depend on the individual features $\boldsymbol{x}$. Clearly our goal is $p(\boldsymbol{x}) = \mathbb{P}(D|\boldsymbol{x})$, but the only thing we can estimate is

$$\tilde{p}(\boldsymbol{x}) = \mathbb{P}(D|S, \boldsymbol{x}) = \frac{\mathbb{P}(S, D|\boldsymbol{x})}{\mathbb{P}(S|\boldsymbol{x})} = \frac{\pi_1 p(\boldsymbol{x})}{\pi_1 p(\boldsymbol{x}) + \pi_2(1-p(\boldsymbol{x}))},$$

which can be modeled by

$$\text{logit}(\tilde{p}(\boldsymbol{x})) = \eta = \boldsymbol{x}'\boldsymbol{\beta} = \beta_0 + \sum_{j=1}^{k} \beta_j x_j,$$

where $\boldsymbol{\beta}$ is the unknown parameter. It can be seen that

$$\text{logit}(\tilde{p}(\boldsymbol{x})) - \log(\pi_1/\pi_2) = \text{logit}(p(\boldsymbol{x})),$$

where π_1 and π_2 are free of $\boldsymbol{x}$ but not estimable. Fortunately,

$$\text{logit}(\tilde{p}(\boldsymbol{x}_1)) - \text{logit}(\tilde{p}(\boldsymbol{x}_0)) = \text{logit}(p(\boldsymbol{x}_1)) - \text{logit}(p(\boldsymbol{x}_0)),$$

i.e.,

$$\frac{\tilde{p}(\boldsymbol{x}_1)/(1-\tilde{p}(\boldsymbol{x}_1))}{\tilde{p}(\boldsymbol{x}_0)/(1-\tilde{p}(\boldsymbol{x}_0))} = \frac{p(\boldsymbol{x}_1)/(1-p(\boldsymbol{x}_1))}{p(\boldsymbol{x}_0)/(1-p(\boldsymbol{x}_0))},$$

and thus we can make inferences about the odds ratio. For example, we have $\hat{\psi} = \frac{ad}{bc}$ using table 3.

[xxi] cf. https://zhuanlan.zhihu.com/p/79121196

§21 Diagnostics in GLMs & Quasi-Likelihood Methods (2020/6/3)

We introduce some tools for detecting violations of the assumptions in a GLM. Recall that a GLM can be treated locally as a linear regression model with working responses z_i and working weights w_{ii}, which results in

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}\boldsymbol{z}$$

upon convergence. When computing the corresponding linear predictor $\hat{\eta}_i = \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}$, the contribution z_i makes, up to standardization, is the i^{th} **leverage** h_{ii}, which turns out to be the diagonal element of the **hat matrix**

$$\boldsymbol{H} = \boldsymbol{W}^{1/2}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}^{1/2},$$

since $\mathrm{Var}(\boldsymbol{z}) = \boldsymbol{W}^{-1}$ and

$$\boldsymbol{W}^{1/2}\hat{\boldsymbol{\eta}} = \boldsymbol{W}^{1/2}\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{H}\boldsymbol{W}^{1/2}\boldsymbol{z}.$$

Similar to the linear model, the **residual plots** of the GLM also provide graphical diagnostics:

- scatter plot of residuals vs. fitted values $\hat{\mu}_i = g^{-1}(\hat{\eta}_i)$ for lack of fit;
- scatter plot of residuals vs. explanatory variables $\boldsymbol{x}_i$ for trend in $\boldsymbol{x}_i$;
- scatter plot of residuals vs. indexes i for auto-regressive pattern.

In the context of GLMs, the residuals could be

- Pearson residuals $r_{i,P}$;
- deviance residuals $r_{i,D}$;
- response residuals $y_i - \hat{\mu}_i$;
- working residuals $z_i - \hat{\eta}_i$.

Note that residuals are complicated by the link function. Intuitively, if the link function is suitable and the model fits the data properly, there will be no obvious trend in residual plots. In other words, a trend in residual plots could be caused by missing terms in some $\boldsymbol{x}$ or by inappropriate choice of the link. An analytic method for choosing the link function is to utilize the **embedding technique**. Consider a family of functions $g(\bullet;\alpha)$ that yield models

$$g(\mu;\alpha) = \eta = \boldsymbol{x}'\boldsymbol{\beta},$$

where the suitable link corresponds to some α_0. Here are two examples.

- For the response $y \sim \frac{1}{m}\mathrm{Binomial}(m,\pi)$, the Aranda-Ordaz family consists of

$$g(\pi;\alpha) = \log\left\{\frac{(1-\pi)^{-\alpha}-1}{\alpha}\right\}, \quad \alpha > 0,$$

  which satisfies that $g(\pi;1) = \mathrm{logit}(\pi) = \log\frac{\pi}{1-\pi}$ and $g(\pi;0) = \lim_{\alpha\downarrow 0} g(\pi;\alpha) = \log(-\log(1-\pi))$.

- For the response $y \sim \mathrm{Poisson}(\mu)$, the Box-Cox family consists of

$$g(\mu;\alpha) = \begin{cases} (\mu^\alpha - 1)/\alpha, & \alpha > 0; \\ \log\mu, & \alpha = 0. \end{cases}$$

One can perform estimation of α based on data, e.g., profile the pseudo-likelihood function of $(\alpha,\boldsymbol{\beta})$ and maximize it iteratively (see §15). Alternatively, one can test

$$H_0 : \alpha = \alpha_0 \quad \text{vs.} \quad H_1 : \alpha \neq \alpha_0$$

for an assumed α_0. Expanding

$$\eta = g(\mu;\alpha) \approx g(\mu;\alpha_0) + (\alpha-\alpha_0)\frac{\partial g}{\partial\alpha}(\mu;\alpha_0),$$

we may proceed as follows. First, fit the GLM of y onto $\boldsymbol{x}$ with the link $g(\bullet;\alpha_0)$, where the data-driven dispersion parameter is obtained. Next, calculate

$$\gamma = \frac{\partial g}{\partial \alpha}(\hat{\mu};\alpha_0)$$

using the fitted values within each observation. Finally, refit the GLM of y onto $\boldsymbol{x}$ and γ with the link function $g(\bullet;\alpha_0)$, and the null hypothesis $H_0:\alpha=\alpha_0$ becomes that the coefficient of γ vanishes, for which it's natural to invoke the nested likelihood-ratio test (see §18).

✍ The embedding technique can also apply to the variance function, which may be parameterized into a family, e.g., $V(\mu;\zeta)=\begin{cases}\mu^\zeta, & \zeta>0\\ \log\mu, & \zeta=0\end{cases}$ for count data. We can obtain the maximum quasi-likelihood (explained soon) for each ζ, and then choose $\hat{\zeta}$ to be the maximizer of this function.

For a GLM, the inference depends on the assumed distribution for y_i only through the mean μ_i and the variance function $V(\cdot)$. A moment-based approach, **quasi-likelihood estimation** (QLE), generalizes the exponential dispersion family. In addition to the link function and the variance function, a GLM specifies the independence of observations, which is not indispensable in the quasi-likelihood approach. Note that if y_i is drawn from the p.d.f. $f_i(y_i;\theta_i,\phi)=\exp\{\frac{y_i\theta_i-b(\theta_i)}{\phi/w_i}+c_i(y_i,\phi)\}$, then $\mu_i=\dot{b}(\theta_i)$ and we have

$$\frac{\partial\log f_i}{\partial\mu_i}=\frac{\partial\log f_i}{\partial\theta_i}\Big/\frac{\partial\mu_i}{\partial\theta_i}=\frac{y_i-\dot{b}(\theta_i)}{\phi/w_i}\Big/\ddot{b}(\theta_i)=\frac{y_i-\mu_i}{\phi V(\mu_i)/w_i}.$$

Suppose now the response vector $\boldsymbol{y}=(y_i)^{1\le i\le n}$ has the covariance matrix

$$\mathrm{Var}(\boldsymbol{y})=\sigma^2\mathbf{V},$$

where σ^2 represents the dispersion and $\mathbf{V}=\mathbf{V}(\boldsymbol{\mu})$ is a known matrix-valued function of $\boldsymbol{\mu}=(\mu_i)^{1\le i\le n}$. It is assumed that the parameter of interest, $\boldsymbol{\beta}\in\mathbb{R}^p$, relates to the dependence of $\boldsymbol{\mu}$ on covariates $\boldsymbol{x}$, i.e., $g(\mu_i)=\boldsymbol{x}_i'\boldsymbol{\beta}$ for a link function $g(\cdot)$. The **quasi-score function** is then defined as

$$\boldsymbol{U}(\boldsymbol{\beta})=\boldsymbol{D}'\mathbf{V}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})/\sigma^2,$$

where

$$\boldsymbol{D}=\partial\boldsymbol{\mu}/\partial\boldsymbol{\beta}'=(\partial\mu_i/\partial\beta_r)_{1\le r\le p}^{1\le i\le n}=(\tfrac{\mathrm{d}\boldsymbol{\eta}}{\mathrm{d}\boldsymbol{\mu}})^{-1}\boldsymbol{X}=(x_{ir}/\dot{g}(\mu_i))_{1\le r\le p}^{1\le i\le n}.$$

One can see that $\mathbb{E}[\boldsymbol{U}(\boldsymbol{\beta})]=\mathbf{0}_p$, $\mathrm{Var}(\boldsymbol{U}(\boldsymbol{\beta}))=\boldsymbol{D}'\mathbf{V}^{-1}\boldsymbol{D}/\sigma^2$, and

$$\tfrac{\partial}{\partial\boldsymbol{\beta}'}\boldsymbol{U}(\boldsymbol{\beta})=\big\{\tfrac{\partial}{\partial\boldsymbol{\beta}'}\big(\boldsymbol{D}'\mathbf{V}^{-1}\big)(\boldsymbol{y}-\boldsymbol{\mu})-\boldsymbol{D}'\mathbf{V}^{-1}\boldsymbol{D}\big\}/\sigma^2\implies\mathbb{E}[\tfrac{\partial}{\partial\boldsymbol{\beta}'}\boldsymbol{U}(\boldsymbol{\beta})]=-\boldsymbol{D}'\mathbf{V}^{-1}\boldsymbol{D}/\sigma^2.$$

Thus,

$$\mathcal{I}(\boldsymbol{\beta})=\mathrm{Var}(\boldsymbol{U}(\boldsymbol{\beta}))=-\mathbb{E}[\tfrac{\partial}{\partial\boldsymbol{\beta}'}\boldsymbol{U}(\boldsymbol{\beta})]=\boldsymbol{D}'\mathbf{V}^{-1}\boldsymbol{D}/\sigma^2$$

serves as the **quasi-information matrix**. If the line integral,

$$Q(\boldsymbol{\mu};\boldsymbol{y})=\sigma^{-2}\int_{s_0}^{s_1}(\boldsymbol{y}-\boldsymbol{\gamma})'\mathbf{V}(\boldsymbol{\gamma})^{-1}\,\mathrm{d}\boldsymbol{\gamma}(s)$$

along $\boldsymbol{\gamma}:[s_0,s_1]\to\mathbb{R}^n$ from $\boldsymbol{\gamma}(s_0)=\boldsymbol{y}$ to $\boldsymbol{\gamma}(s_1)=\boldsymbol{\mu}$, is path-independent, it makes sense to use this function as a **quasi-log-likelihood**, which satisfies that

$$\tfrac{\partial}{\partial\boldsymbol{\beta}}Q(\boldsymbol{\mu};\boldsymbol{y})=\boldsymbol{U}(\boldsymbol{\beta}).$$

When the responses y_i are independent so that we can write $\mathbf{V}(\boldsymbol{\mu})=\mathrm{diag}(V_1(\mu_1),\dots,V_n(\mu_n))$,

$$Q(\boldsymbol{\mu};\boldsymbol{y})=\sum_{i=1}^n Q_i(\mu_i;y_i)=\sum_{i=1}^n\int_{y_i}^{\mu_i}\frac{y_i-t}{\sigma^2V_i(t)}\,\mathrm{d}t\le 0,$$

and the **quasi-deviance** is defined to be

$$D(\boldsymbol{y},\boldsymbol{\mu})=2\Big[\underbrace{Q(\boldsymbol{y};\boldsymbol{y})}_{=0}-Q(\boldsymbol{\mu};\boldsymbol{y})\Big]\cdot\sigma^2=2\sum_{i=1}^n\int_{\mu_i}^{y_i}\frac{y_i-t}{V_i(t)}\,\mathrm{d}t\ge 0.$$

$V(\mu)$	$\sigma^2 Q(\mu;y)+C(y)=\int_y^\mu \frac{y-t}{V(t)}\,\mathrm{d}t+C(y)$	Distribution	Range Restrictions
1	$-(y-\mu)^2/2$	Normal	—
μ	$y\log\mu-\mu$	Poisson	$\mu>0,\ y\geq 0$
μ^2	$-y/\mu-\log\mu$	Gamma	$\mu>0,\ y\geq 0$
μ^3	$-y/(2\mu^2)+1/\mu$	Inverse Gaussian	$\mu>0,\ y\geq 0$
μ^ζ	$\mu^{-\zeta}(\frac{\mu y}{1-\zeta}-\frac{\mu^2}{2-\zeta})$	—	$\mu>0,\ \zeta\neq 0,1,2$
$\mu(1-\mu)$	$y\log(\frac{\mu}{1-\mu})+\log(1-\mu)$	Scaled Binomial	$0<\mu<1,\ 0\leq y\leq 1$
$\mu^2(1-\mu)^2$	$(2y-1)\log(\frac{\mu}{1-\mu})-\frac{y}{\mu}-\frac{1-y}{1-\mu}$	—	$0<\mu<1,\ 0\leq y\leq 1$
$\mu+\mu^2/k$	$y\log(\frac{\mu}{k+\mu})+k\log(\frac{k}{k+\mu})$	Negative Binomial	$\mu>0,\ y\geq 0$

Table 4: Quasi-Likelihoods Associated with Some Simple Variance Functions

The quasi-likelihood estimator $\hat{\boldsymbol{\beta}}$ is obtained by solving the **generalized estimating equation** (GEE)

$$\mathbf{0}_p=\sigma^2\boldsymbol{U}(\boldsymbol{\beta})=\boldsymbol{D}'\mathbf{V}^{-1}(\boldsymbol{y}-\boldsymbol{\mu}).$$

Step t in the iterative process of Fisher's scoring algorithm is given by

$$\begin{aligned}\boldsymbol{\beta}^{(t+1)}&=\boldsymbol{\beta}^{(t)}+\mathcal{I}(\boldsymbol{\beta}^{(t)})^{-1}\boldsymbol{U}(\boldsymbol{\beta}^{(t)})\\&=\boldsymbol{\beta}^{(t)}+\{\boldsymbol{D}(\boldsymbol{\beta}^{(t)})'\mathbf{V}(\boldsymbol{\beta}^{(t)})^{-1}\boldsymbol{D}(\boldsymbol{\beta}^{(t)})\}^{-1}\boldsymbol{D}(\boldsymbol{\beta}^{(t)})'\mathbf{V}(\boldsymbol{\beta}^{(t)})^{-1}(\boldsymbol{y}-\boldsymbol{\mu}(\boldsymbol{\beta}^{(t)}))\\&=\boldsymbol{\beta}^{(t)}+\big\{\boldsymbol{X}'(\tfrac{\mathrm{d}\boldsymbol{\eta}}{\mathrm{d}\boldsymbol{\mu}}\big|_{\boldsymbol{\mu}(\boldsymbol{\beta}^{(t)})})^{-1}\mathbf{V}(\boldsymbol{\beta}^{(t)})^{-1}(\tfrac{\mathrm{d}\boldsymbol{\eta}}{\mathrm{d}\boldsymbol{\mu}}\big|_{\boldsymbol{\mu}(\boldsymbol{\beta}^{(t)})})^{-1}\boldsymbol{X}\big\}^{-1}\boldsymbol{X}'(\tfrac{\mathrm{d}\boldsymbol{\eta}}{\mathrm{d}\boldsymbol{\mu}}\big|_{\boldsymbol{\mu}(\boldsymbol{\beta}^{(t)})})^{-1}\mathbf{V}(\boldsymbol{\beta}^{(t)})^{-1}(\boldsymbol{y}-\boldsymbol{\mu}(\boldsymbol{\beta}^{(t)}))\\&=\big\{\boldsymbol{X}'(\tfrac{\mathrm{d}\boldsymbol{\eta}}{\mathrm{d}\boldsymbol{\mu}}\big|_{\boldsymbol{\mu}(\boldsymbol{\beta}^{(t)})})^{-1}\mathbf{V}(\boldsymbol{\beta}^{(t)})^{-1}(\tfrac{\mathrm{d}\boldsymbol{\eta}}{\mathrm{d}\boldsymbol{\mu}}\big|_{\boldsymbol{\mu}(\boldsymbol{\beta}^{(t)})})^{-1}\boldsymbol{X}\big\}^{-1}\boldsymbol{X}'(\tfrac{\mathrm{d}\boldsymbol{\eta}}{\mathrm{d}\boldsymbol{\mu}}\big|_{\boldsymbol{\mu}(\boldsymbol{\beta}^{(t)})})^{-1}\mathbf{V}(\boldsymbol{\beta}^{(t)})^{-1}(\tfrac{\mathrm{d}\boldsymbol{\eta}}{\mathrm{d}\boldsymbol{\mu}}\big|_{\boldsymbol{\mu}(\boldsymbol{\beta}^{(t)})})^{-1}\boldsymbol{z}^{(t)},\end{aligned}$$

where

$$\boldsymbol{z}^{(t)}=\boldsymbol{X}\boldsymbol{\beta}^{(t)}+\tfrac{\mathrm{d}\boldsymbol{\eta}}{\mathrm{d}\boldsymbol{\mu}}\big|_{\boldsymbol{\mu}(\boldsymbol{\beta}^{(t)})}(\boldsymbol{y}-\boldsymbol{\mu}(\boldsymbol{\beta}^{(t)})).$$

Note that the working weighting matrix

$$\boldsymbol{W}^{(t)}=(\tfrac{\mathrm{d}\boldsymbol{\eta}}{\mathrm{d}\boldsymbol{\mu}}\big|_{\boldsymbol{\mu}(\boldsymbol{\beta}^{(t)})})^{-1}\mathbf{V}(\boldsymbol{\beta}^{(t)})^{-1}(\tfrac{\mathrm{d}\boldsymbol{\eta}}{\mathrm{d}\boldsymbol{\mu}}\big|_{\boldsymbol{\mu}(\boldsymbol{\beta}^{(t)})})^{-1}$$

might be non-diagonal. Under suitable regularity conditions, $\hat{\boldsymbol{\beta}}$ is approximately unbiased for $\boldsymbol{\beta}$ and asymptotically normally distributed with limiting variance-covariance matrix

$$\mathrm{Var}(\hat{\boldsymbol{\beta}})\approx\mathcal{I}(\boldsymbol{\beta})^{-1}=\sigma^2(\boldsymbol{D}'\mathbf{V}^{-1}\boldsymbol{D})^{-1}.$$

The conventional estimate of σ^2 is based on Pearson's chi-square or the deviance, namely

$$\hat{\sigma}^2=\frac{1}{n-p}\sum_{i=1}^n\frac{\big(y_i-\mu_i(\hat{\boldsymbol{\beta}})\big)^2}{V_i\big(\mu_i(\hat{\boldsymbol{\beta}})\big)}\quad\text{or}\quad\frac{1}{n-p}D\big(\boldsymbol{y},\boldsymbol{\mu}(\hat{\boldsymbol{\beta}})\big).$$

In all of the above respects, the quasi-likelihood behaves just like an ordinary likelihood.

The quasi-score function is a rather special case of what is known as an estimating function. A function $\boldsymbol{G}(\boldsymbol{\beta};\boldsymbol{y})$ of the parameter $\boldsymbol{\beta}$ and the data $\boldsymbol{y}$ is said to be an **estimating function** if it has zero mean for all parameter values. Provided that there are as many equations as parameters, estimates $\widetilde{\boldsymbol{\beta}}$ are obtained as the solution of the equation $\boldsymbol{G}(\boldsymbol{\beta};\boldsymbol{y})=\mathbf{0}$. Consider the class of linear estimating functions

$$\mathbf{h}(\boldsymbol{\beta})=\mathbf{H}'(\boldsymbol{y}-\boldsymbol{\mu}(\boldsymbol{\beta})),$$

where $\mathbf{H}$ may depend on $\boldsymbol{\beta}$ but not on $\boldsymbol{y}$. We will show that $\mathbf{V}^{-1}\boldsymbol{D}$ is the optimal $\mathbf{H}$. Let $\widetilde{\boldsymbol{\beta}}$ be the root of $\mathbf{h}(\bullet)$, whose first order Taylor expansion around the true parameter yields

$$\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}\approx(\mathbf{H}'\boldsymbol{D})^{-1}\mathbf{h}(\boldsymbol{\beta})\implies\mathbb{E}[\widetilde{\boldsymbol{\beta}}]\approx\boldsymbol{\beta},\ \&\ \mathrm{Var}(\widetilde{\boldsymbol{\beta}})\approx\sigma^2(\mathbf{H}'\boldsymbol{D})^{-1}\mathbf{H}'\mathbf{V}\mathbf{H}(\boldsymbol{D}'\mathbf{H})^{-1}.$$

Therefore,

$$\mathrm{Var}(\hat{\boldsymbol{\beta}})^{-1}-\mathrm{Var}(\widetilde{\boldsymbol{\beta}})^{-1}\approx\sigma^{-2}\boldsymbol{D}'\mathbf{V}^{-1/2}(\boldsymbol{I}-\boldsymbol{P}_{\mathbf{V}^{1/2}\mathbf{H}})\mathbf{V}^{-1/2}\boldsymbol{D}\succeq 0,$$

where $\boldsymbol{P}_{\mathbf{V}^{1/2}\mathbf{H}}=\mathrm{proj}_{\mathrm{Col}(\mathbf{V}^{1/2}\mathbf{H})}$ (see §2), and it follows that $\mathrm{Var}(\widetilde{\boldsymbol{\beta}})-\mathrm{Var}(\hat{\boldsymbol{\beta}})\succeq 0$ asymptotically.

§22 Nonparametric Regression & Additional Material (2020/6/5)

We will focus on nonparametric regression function estimation. To put the problem into a statistical framework, it is convenient to think of each dataset as a realization of a random sample from a certain population, consisting of i.i.d. observations $(X_i, Y_i)_{1\le i\le n}$. Let (X, Y) be a generic member of the sample. The conditional mean and variance of the response Y on the covariate X are denoted respectively by

$$m(x) = \mathbb{E}[Y|X = x] \quad \& \quad \sigma^2(x) = \mathrm{Var}(Y|X = x).$$

Many applications involve estimation of the **regression function** $m(x)$ or its ν^{th} derivative $m^{(\nu)}(x)$. To aid understanding, regard the data as being generated from the location-scale model

$$Y = m(X) + \sigma(X)\varepsilon,$$

where $\mathbb{E}[\varepsilon] = 0$, $\mathrm{Var}(\varepsilon) = 1$, and X and ε are independent. Here are some remarks[xxii].

✍ Generally speaking, nonparametric regression estimators are not defined with the random or fixed design specifically in mind, i.e., there is no real distinction made here. We may always concentrate on the *conditional* design. As a caveat, some wavelet-based estimators do assume evenly spaced fixed inputs, such as $X_i \equiv i/n$ in the univariate case.

✍ Importantly, in nonparametric regression we don't assume a particular parametric form for $m(\cdot)$. This doesn't mean, however, that we can't estimate $m(\cdot)$ using, say, a linear combination of spline basis functions. A common question: the coefficients on the spline basis functions are parameters, so how can this be nonparametric? Again, the point is that we don't assume a parametric form for $m(\cdot)$, i.e., we don't assume that $m(\cdot)$ itself is an exact linear combination of splines basis functions.

The case of one-dimensional explanatory variables lays out the foundation and provides insights. Suppose that the regression function $m(\cdot)$ can be approximated by

$$m(\cdot) \approx \sum_{j=0}^{p} \frac{m^{(j)}(x)}{j!}(\cdot - x)^j = \sum_{j=0}^{p} \beta_j(\cdot - x)^j,$$

using Taylor's expansion. In (global) polynomial regression, we may estimate $m^{(\nu)}(x)$ by $\hat{m}_\nu(x) = \nu!\hat{\beta}_\nu$, where

$$\left(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p\right) = \underset{(\beta_0,\beta_1,\ldots,\beta_p)}{\arg\min} \sum_{i=1}^{n} \left[Y_i - \sum_{j=0}^{p} \beta_j(X_i - x)^j\right]^2.$$

While this approach has been widely used, it suffers from a few drawbacks. One is that polynomial functions are not very flexible because of smoothness. Another is that individual observations can have a large influence on remote parts of the curve. Besides, the polynomial degree cannot be controlled continuously. Hence it is natural to consider *local regression methods* and *regularization methods*.

Without assuming a specific form of the regression function, data points far from x carries little information about the value of $m(x)$. Thus, an intuitive estimator for $m(x)$ is the running local average, e.g., the k-NN estimator in §5. An improved version of this is the *locally weighted average*. We now start the exploration of the method of **local polynomial regression**. Let $K(\cdot)$ be a symmetric density function, called the **kernel**, and $h > 0$, called the **bandwidth**. Note that three common kernels are

- the Gaussian kernel $K(u) = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-u^2/2}$,
- the boxcar kernel $K(u) = \mathbb{1}_{[|u|\le 1]}/2$, and
- the Epanechnikov kernel $K(u) = \frac{3}{4}(1 - u^2)\mathbb{1}_{[|u|\le 1]}$, which is optimal to some extent.

Then, we may estimate $m^{(\nu)}(x)$ by

$$\hat{m}_\nu(x) = \nu!\hat{\beta}_\nu,$$

where

$$\left(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p\right) = \underset{(\beta_0,\beta_1,\ldots,\beta_p)}{\arg\min} \sum_{i=1}^{n} K_h(X_i - x)\left[Y_i - \sum_{j=0}^{p} \beta_j(X_i - x)^j\right]^2,$$

[xxii] from http://www.stat.cmu.edu/~ryantibs/statml/lectures/nonpar.pdf

where $K_h(\cdot) = \frac{1}{h}K(\frac{\cdot}{h})$. The locality is determined by the rescaled kernel $K_h(\cdot)$ which downweighs all the X_i's that are not close to x. Working with matrix notation, the solution $\hat{\boldsymbol{\beta}} = (\hat{\beta}_j)^{0\le j\le p}$ to the weighted least squares problem

$$\min_{\boldsymbol{\beta}} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{W}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}),$$

where $\boldsymbol{X} = ((X_i - x)^j)_{0\le j\le p}^{1\le i\le n}$, $\boldsymbol{Y} = (Y_i)^{1\le i\le n}$, and $\boldsymbol{W} = \mathrm{diag}(K_h(X_i - x))$, is given by

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}\boldsymbol{Y}.$$

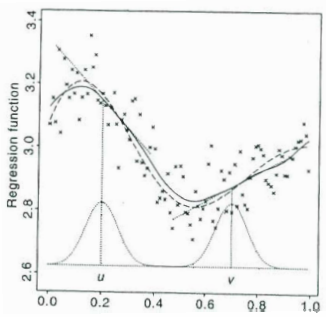

Figure 5.1. *Local linear kernel estimate (solid curve) of the regression function m given by (5.1) based on 100 simulated observations (represented by crosses). The dashed curve is the true function m. The dotted curves are the kernel weights and linear fits at the points u and v.*

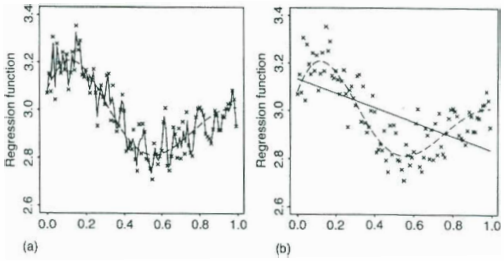

Figure 5.2. *Local linear kernel estimates based on the same data as used in Figure 5.1, but with (a) a very small bandwidth (b) a very large bandwidth.*

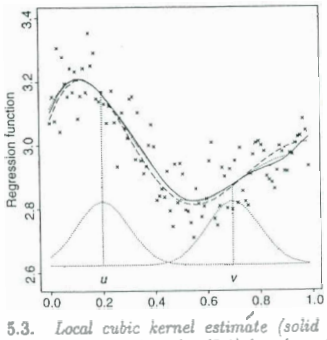

Figure 5.3. *Local cubic kernel estimate (solid curve) of the regression function m given by (5.1) based on 100 simulated observations (represented by crosses). The dashed curve is the true function m. The dotted curves are the kernel weights and cubic fits at the points u and v.*

Figure 5: Local Polynomial Kernel Estimates[xxiii)]

Setting $p = 0$ gives the (Nadaraya-Watson) **kernel regression** estimator

$$\hat{m}^{\mathrm{NW}}(x) = \frac{\sum_{i=1}^n K_h(X_i - x)Y_i}{\sum_{i=1}^n K_h(X_i - x)}.$$

The case where $p = 1$ is called **local linear regression** and is recommended as a default choice for estimating $m(\cdot)$. It's suggested that for estimating $m^{(\nu)}(x)$, one should choose p so that $p - \nu$ is an *odd*

[xxiii)] from Wand, M.P., & Jones, M.C. (1995). *Kernel Smoothing.* (http://matt-wand.utsacademics.info/webWJbook/)

number. Usually, $p = \nu + 1$. Not surprisingly, the most important tuning parameter that governs the complexity of the model is h, the bandwidth. As stated in Fan & Gijbels [FG], Thm. 3.1, asymptotically we have

$$\mathrm{Var}(\hat{m}_\nu(x)) \propto \frac{\sigma^2(x)}{f_X(x)} \frac{1}{nh^{2\nu+1}},$$

where $f_X(x)$ is the density for the distribution of X at x, and

$$\mathrm{Bias}(\hat{m}_\nu(x)) = \mathbb{E}\hat{m}_\nu(x) - m^{(\nu)}(x) \propto m^{(p+1)}(x) h^{p+1-\nu}$$

for $p - \nu$ odd. The performance of $\hat{m}_\nu(x)$ is conveniently assessed via its mean squared error

$$\mathrm{MSE}(\hat{m}_\nu(x)) = \mathbb{E}\big[\hat{m}_\nu(x) - m^{(\nu)}(x)\big]^2 = \mathrm{Var}(\hat{m}_\nu(x)) + \mathrm{Bias}^2(\hat{m}_\nu(x)) \approx \frac{A}{nh^{2\nu+1}} + Bh^{2p+2-2\nu}.$$

Clearly $\mathrm{MSE}(\hat{m}_\nu(x))$ increases with ν, so it's harder to estimate high-order derivatives of $m(\cdot)$. Note that we should require that the effective sample size nh grows to infinity while the bandwidth $h \to 0$. When dealing with the bandwidth selection problem, a key issue is the bias-variance tradeoff. A theoretically ideal choice of a bandwidth for estimating $m^{(\nu)}(x)$ can be approximated by the asymptotically optimal local bandwidth which minimizes the asymptotic MSE. This leads to $h_{\mathrm{opt}} \propto n^{-1/(2p+3)}$, together with the optimal rate $n^{-(2p+2-2\nu)/(2p+3)}$ of convergence. One cannot choose an arbitrarily high p, as it reflects the underlying smoothness of $m(\cdot)$. The slower rate $n^{-(2p+2-2\nu)/(2p+3)}$ than n^{-1} in most parametric models is the price of using nonparametric methods. In practice, we cannot use the asymptotically optimal bandwidth depending on unknown quantities. Various techniques for selecting tuning parameters have been proposed, among which cross-validation (see §13) and rules of thumb are commonly used.

Spline smoothing is motivated from a different perspective, where estimators are constructed globally. A spline is a special piecewise polynomial. Given a strictly increasing **knots** sequence $\xi_1 < \cdots < \xi_K$ contained in some finite interval (a, b), a function $f(\cdot)$ is called a **spline** of **order** M if $f(\cdot)$ is a polynomial of *degree* no more than $M - 1$ over $[a, \xi_1], [\xi_1, \xi_2], \ldots, [\xi_K, b]$, and $f(\cdot)$ and its derivatives up to order $M - 2$ are continuous on $[a, b]$. For example, a (piecewise) cubic spline is of order $M = 4$. Compared with a global polynomial of degree M that has the continuous $(M-1)^{\text{th}}$ derivative, $f^{(M-1)}(\cdot)$ is a step function with jumps at $\xi_1, \ldots, \xi_K$. It can be seen that the total number of free parameters is $M + K$. The spline can be represented by a set of basis functions.

- The general form of the **truncated power basis** is

$$\begin{aligned} h_j(x) &= x^{j-1}, && j = 1, \ldots, M; \\ h_{M+l}(x) &= (x - \xi_l)_+^{M-1}, && l = 1, \ldots, K. \end{aligned}$$

  Here x_+ denotes the positive part of x, i.e., $x_+ = \max(x, 0) = x\mathbb{1}_{[x>0]}$.

- However, numerically, power of large numbers may result in severe rounding problems. It is preferred to use the **B-spline basis** that is nearly orthogonal, for efficient and stable computation. Before we get started, we need to augment the knots sequence. Let $\xi_0 = a$ and $\xi_{K+1} = b$ be two boundary knots, which typically define the domain over which we wish to evaluate the spline. The augmented knots $\tau_1, \ldots, \tau_{K+2M}$ are determined such that

  - $\tau_1 \le \tau_2 \le \cdots \le \tau_M \le \xi_0$;
  - $\tau_{j+M} = \xi_j,\ j = 1, \ldots, K$;
  - $\xi_{K+1} \le \tau_{K+M+1} \le \tau_{K+M+2} \le \cdots \le \tau_{K+2M}$.

  The actual values of these additional knots beyond the boundary are arbitrary, and it is customary to make them all the same and equal to ξ_0 and ξ_{K+1}, respectively. We define the B-spline basis functions recursively in terms of divided differences as follows. First, $B_{j,1}(x) = \mathbb{1}_{[\tau_j, \tau_{j+1})}(x)$ for $j = 1, \ldots, K + 2M - 1$, a.k.a. Haar basis functions. Next, for $m = 2, \ldots, M$,

$$B_{j,m}(x) = \frac{x - \tau_j}{\tau_{j+m-1} - \tau_j} B_{j,m-1}(x) + \frac{\tau_{j+m} - x}{\tau_{j+m} - \tau_{j+1}} B_{j+1,m-1}(x), \quad j = 1, \ldots, K + 2M - m.$$

  It is understood that if the denominator is 0, then the function is defined to be 0. Finally, $B_{j,M}(\cdot),\ j = 1, \ldots, K + M$, are the B-spline basis functions.

- There is a boundary problem that matters. The splines beyond boundary knots behave even more wildly than the corresponding global polynomials in that region. It can be remedied by controlling the change rate of the piecewise polynomial in the boundary regions $[a, \xi_1]$ and $[\xi_K, b]$. A spline $f(\cdot)$ of order $M = 2r$ is called a **natural spline**, if $f(\cdot)$ is a polynomial of degree no more than $r-1$ outside $[\xi_1, \xi_K]$, which implies that the number of free parameters is $(M+K)-2[(M-1)-(r-1)] = K$. For example, a natural cubic spline corresponds to $r = 2$ and is linear outside $[\xi_1, \xi_K]$. Note that there is a variant of the truncated power basis and a variant of the B-spline basis for natural splines.

In what follows, we are concerned with estimating $f(x) = \mathbb{E}[Y|X = x]$ by using the spline method. The fitted **regression spline** is

$$\hat{f}(\cdot) = \sum_{j=1}^{M+K} \hat{\beta}_j b_j(\cdot), \quad \text{where} \quad (\hat{\beta}_1, \ldots, \hat{\beta}_{M+K}) = \underset{(\beta_1,\ldots,\beta_{M+K})}{\arg\min} \sum_{i=1}^{n} \left[Y_i - \sum_{j=1}^{M+K} \beta_j b_j(X_i) \right]^2,$$

provided that $b_1(\cdot), \ldots, b_{M+K}(\cdot)$ are spline basis. In practice, knots are specified by equal distance (for evenly distributed observations) or equal quantile (for highly clustered designs), and the number K of knots is usually selected via generalized cross-validation (see §13). A **smoothing spline** estimator is defined by minimizing

$$\sum_{i=1}^{n} \left[Y_i - f(X_i) \right]^2 + \lambda \int_a^b \left[f^{(r)}(x) \right]^2 \mathrm{d}x$$

over the class of functions $f(\cdot)$ that have an absolutely continuous $(r-1)^{\text{th}}$ derivative and a square-integrable r^{th} derivative, where $\lambda \geq 0$ is the tuning parameter that assigns weight to the roughness penalty. When $r = 2$, it is the so-called cubic smoothing spline. It has been shown[xxiv] that, if $n > r$, there exists a unique minimizer that is a natural spline of order $2r$ with knots $X_1, \ldots, X_n$. One can easily see that the solution is in the form of a ridge regression estimator. Denote the design matrix by $\boldsymbol{X_b} = (b_j(X_i))$ and the penalty matrix (which can be generalized) by $\boldsymbol{\Omega} = (\int b_j^{(r)} b_k^{(r)})$, where $\boldsymbol{b} = \{b_j(\cdot)\}$ is a basis for the splines. Note that the natural spline basis functions placing knots at all inputs circumvent the problem of knot selection. Then, the **penalized spline** smoother is $\hat{f}(\cdot) = \sum_j \hat{\beta}_j b_j(\cdot)$, where $\hat{\boldsymbol{\beta}} = (\hat{\beta}_j)$ is given by

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \{ \|\boldsymbol{Y} - \boldsymbol{X_b}\boldsymbol{\beta}\|^2 + \lambda \boldsymbol{\beta}' \boldsymbol{\Omega} \boldsymbol{\beta} \} = (\boldsymbol{X_b}' \boldsymbol{X_b} + \lambda \boldsymbol{\Omega})^{-1} \boldsymbol{X_b}' \boldsymbol{Y}.$$

The term $\lambda\boldsymbol{\Omega}$ shrinks the regression coefficients and gives rise to a smoother fit. In analogy to standard linear models, the trace of the hat matrix $\boldsymbol{H}_\lambda = \boldsymbol{X_b}(\boldsymbol{X_b}'\boldsymbol{X_b} + \lambda\boldsymbol{\Omega})^{-1}\boldsymbol{X_b}'$ is said to be the **effective degree of freedom**. We often choose the smoothing parameter λ by generalized cross-validation (see §13).

Just as in §6, direct multivariate kernel/spline estimators suffer from the curse of dimensionality. Interpreting and visualizing a high-dimensional fit is difficult. As the number of covariates increases, the computational burden becomes prohibitive. In the regression context, a variety of alternative approaches has been proposed to overcome this problem. One of the simplest is the additive modeling methodology discussed in detail by Hastie & Tibshirani [HT]. The **generalized additive model** (GAM) replaces the parametric terms in the GLM with smooth terms in the explanatory variables. Suppose that we have scalar responses y_i and p-dimensional covariates $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})'$. The mean μ_i of y_i conditional on $\boldsymbol{x}_i$ is modeled by

$$g(\mu_i) = \eta_i = \alpha + \sum_{j=1}^{p} f_j(x_{ij}),$$

where $g(\cdot)$ is the link function, α is the intercept, and $f_j(\cdot)$ are smooth functions without a predetermined form. Note that the additive model will do poorly when strong interactions exist, in which case we might consider adding terms like $f_{jk}(x_{ij}x_{ik})$ or even $f_{jk}(x_{ij}, x_{ik})$ if there is sufficient data. The additive model is clearly not as general as fitting $f(x_{i1}, \ldots, x_{ip})$ but it is much simpler to compute and so it is often a good starting point. Iterative procedures known as **backfitting**[xxv], which involve residuals from current fits, are used for estimation. Although the GAM has no analytic form and is hard to interpret, it is flexible and data-driven due to its nonparametric nature. In data analysis, one can explore patterns of the regression using generalized additive models (if suitable).

[xxiv] see also Exercise 5.7 in https://web.stanford.edu/~hastie/ElemStatLearn/printings/ESLII_print12.pdf
[xxv] cf. http://rafalab.github.io/pages/649/section-10.pdf

The following are examples of nonparametric regression, where the code can be implemented in R.

9–44 *Chapter 9. Regression and Smoothing*

```
                          poly(E, degree = 4)3
    poly(E, degree = 4)1
    poly(E, degree = 4)2
    poly(E, degree = 4)3
    poly(E, degree = 4)4  0
    > poly.transform(poly(E,4), coef(ethanol.poly))
          x^0       x^1      x^2       x^3     x^4
     174.3601 -872.2071 1576.735 -1211.219 335.356
```

The summary clearly shows the significance of the 4th order term.

9.12 Smoothing

Polynomial regression can be useful in many situations. However, the choice of terms is not always obvious, and small effects can be greatly magnified or lost completely by the wrong choice.

Another approach to analyzing nonlinear data, attractive because it relies on the data to specify the form of the model, is to fit a curve to the data points *locally*, so that at any point the curve at that point depends only on the observations at that point and some specified neighboring points. Because such a fit produces an estimate of the response that is less variable than the original observed response, the result is called a *smooth*, and procedures for producing such fits are called *scatterplot smoothers*.

S-PLUS offers a variety of scatterplot smoothers:

`loess.smooth`	a *locally weighted regression* smoother.
`smooth.spline`	a cubic smoothing spline, with local behavior similar to that of kernel-type smoothers.
`ksmooth`	a kernel-type scatterplot smoother.
`supsmu`	a very fast variable span bivariate smoother.

9.12. Smoothing 9–45

Halfway between the global parametrization of a polynomial fit and the local, nonparametric fit provided by smoothers are the parametric fits provided by *regression splines*. Regression splines fit a continuous curve to the data by piecing together polynomials fit to different portions of the data. Thus, like smoothers, they are *local* fits. Like polynomials, they provide a parametric fit. In S-PLUS, regression splines can be used to specify the form of a predictor in a linear or more general model, but are not intended for top-level use.

9.12.1 Locally Weighted Regression Smoothing

In locally weighted regression smoothing, we build the smooth function $s(x)$ pointwise as follows:

1. Take a point, say x_0. Find the k nearest neighbors of x_0, which constitute a neighborhood $N(x_0)$. The number of neighbors k is specified as a percentage of the total number of points. This percentage is called the *span*.
2. Calculate the largest distance between x_0 and another point in the neighborhood:
$$\Delta(x_0) = \max_{N(x_0)}|x_0 - x_i|$$
3. Assign weights to each point in $N(x_0)$ using the tri-cube weight function:
$$W\left(\frac{|x_0 - x_i|}{\Delta(x_0)}\right)$$
where
$$W(u) = \begin{cases} (1-u^3)^3, & \text{for } 0 \le u < 1 \\ 0 & \text{otherwise} \end{cases}$$
4. Calculate the weighted least squares fit of y on the neighborhood $N(x_0)$. Take the fitted value $\hat{y}_0 = s(x_0)$.
5. Repeat for each predictor value.

Use the `loess.smooth` function to calculate a locally weighted regression smooth. For example, suppose we want to smooth the **ethanol** data. The following expressions produce the plot shown in figure 9.14:

```
> plot(E,NOx)
> lines(loess.smooth(E,NOx))
```

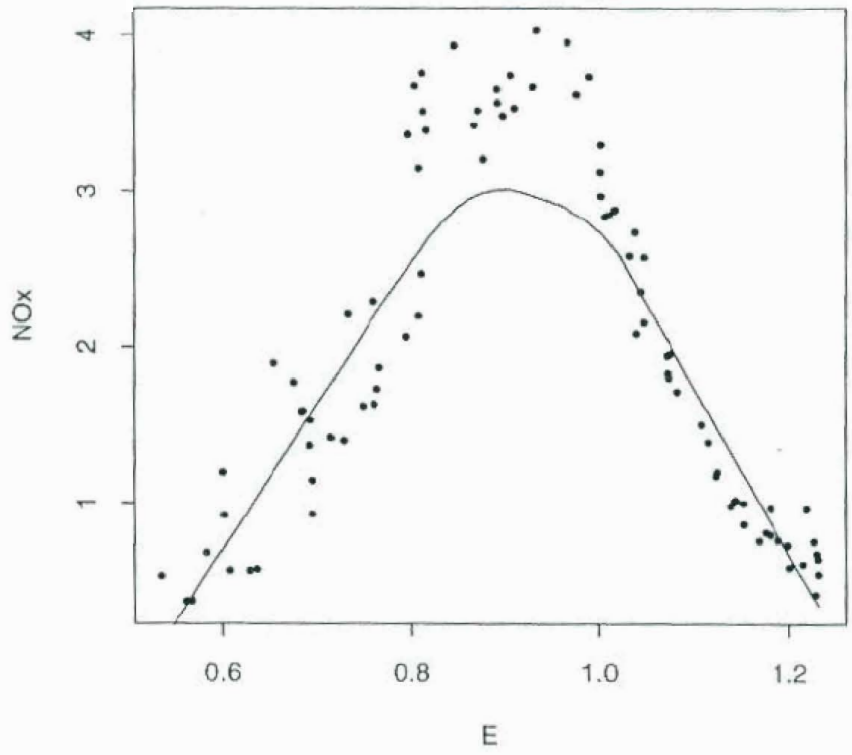

Figure 9.14: Loess-smoothed ethanol data.

The plot shown in figure 9.14 shows the default smoothing, which uses a span of 2/3. For most uses, you will want to specify a smaller span, typically in the range of 0.3 to 0.5.

9.12. Smoothing 9–47

9.12.2 Using the Supersmoother

With `loess`, the span is constant over the entire range of predictor values. However, a constant value will not be optimal if either the error variance or the curvature of the underlying function f varies over the range of x. An increase in the error variance requires an increase in the span whereas an increase in the curvature of f requires a decrease. Local cross-validation avoids this problem by choosing a span for the predictor values x_j based on only the leave-one-out residuals whose predictor values x_i are in the neighborhood of x_j. The supersmoother, `supsmu`, uses local cross-validation to choose the span. Thus, for one-predictor data, it can be a useful adjunct to `loess`.

For example, figure 9.15 shows the result of supersmoothing the response `NOx` as a function of `E` in the `ethanol` data (dotted line) superimposed on a `loess` smooth. To create the figure, use the following commands:

```
> scatter.smooth(E,NOx, span=1/4)
> lines(supsmu(E,NOx), lty=2)
```

Local Cross-Validation

Let $s(x|k)$ denote the linear smoother value at x when span k is used. We wish to choose $k = k(X)$ so as to minimize the mean squared error

$$e^2(k) = E_X E_{Y|X}[Y - s(X|k)]^2$$

where we are considering the joint random variable model for (X, Y). Since

$$E_X E_{Y|X}[Y - s(X|k)]^2 = E_X E_{Y|X}[Y - s(X|k)]^2$$

we would like to choose $k = k(x)$ to minimize

$$\begin{aligned} e_x^2(k) &= E_{Y|X=x}[Y - s(X|k)]^2 \\ &= E_{Y|X=x}[Y - s(x|k)]^2. \end{aligned}$$

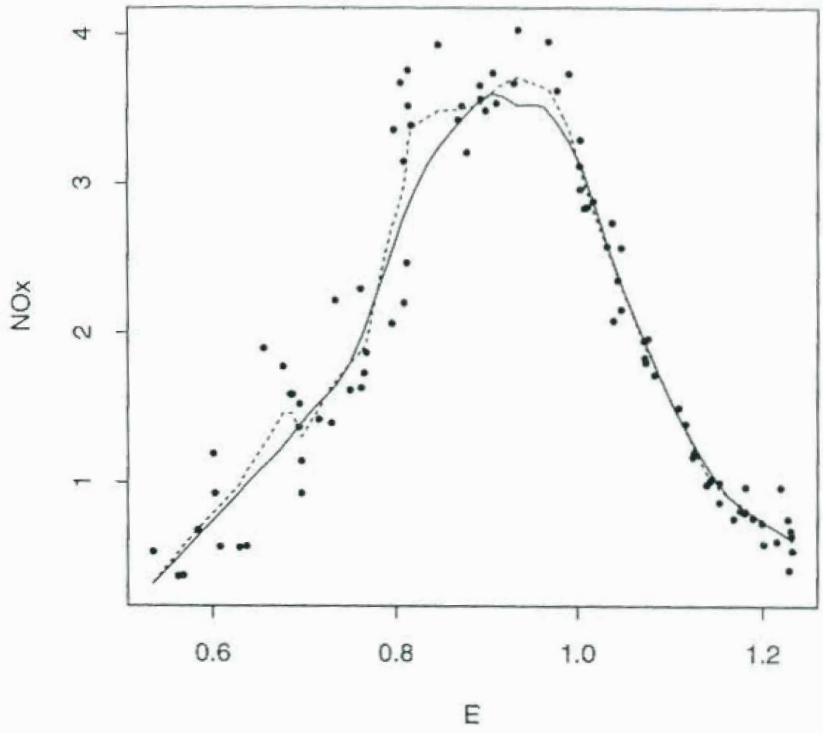

Figure 9.15: Supersmoothed ethanol data (dotted line).

However, we have only the data (x_i, y_i), $i = 1, \ldots, n$, and not the true conditional distribution needed to compute $E_Y|X = x$, and so we cannot calculate $e_x^2(k)$. Thus we resort to cross-validation and try to minimize the cross-validation estimate of $e_x^2(k)$:

$$\hat{e}_{CV}^2(k) = \sum_{i=1}^{n} [y_i - s_{(i)}(x_i|k)]^2.$$

Here $s_{(i)}(x_i|k)$ is the "leave-one-out" smooth at x_i, that is, $s_{(i)}(x_i|k)$ is constructed using all the data (x_j, y_j), $j = 1, \ldots, n$, *except* for (x_i, y_i), and then the resultant local least squares line is evaluated at x_i thereby giving $s_{(i)}(x|k)$. The *leave-one-out residuals*

$$r_{(i)}(k) = y_i - s_{(i)}(x_i|k)$$

are easily obtained from the ordinary residuals

$$r_i(k) = y_i - s(x_i|k)$$

using the standard regression model relation

$$r_{(i)}(k) = \frac{r_i(k)}{h_{ii}}.$$

Here h_{ii}, $i = 1, \ldots, n$, are the diagonals of the so-called "hat" matrix, $H = X(X^TX)^{-1}X^T$, where, for the case at hand of local straight-line regression, X is a 2-columned matrix.

9.12.3 Using the Kernel Smoother

A kernel-type smoother is a type of local average smoother that, for each *target point* x_i in predictor space, calculates a weighted average $\hat{y}_i$ of the observations in a neighborhood of the target point:

$$\hat{y}_i = \sum_{j=1}^{n} w_{ij} y_j \tag{9.6}$$

where

$$w_{ij} = \tilde{K}\left(\frac{x_i - x_j}{b}\right) = \frac{K\left(\frac{x_i - x_j}{b}\right)}{\sum_{k=1}^{n} K\left(\frac{x_i - x_l}{b}\right)}.$$

are weights which sum to one:

$$\sum_{j=1}^{n} w_{ij} = 1.$$

The function K used to calculate the weights is called a *kernel* function, which typically has the following properties:

(a) $K(t) \geq 0$ for all t

(b) $\int_{-\infty}^{\infty} K(t)\,dt = 1$

(c) $K(-t) = K(t)$ for all t (symmetry)

Note that properties (a) and (b) are those of a probability density function.

The parameter b is the *bandwidth* parameter, which determines how large a neighborhood of the target point is used to calculate the local average. A large bandwidth generates a smoother curve, while a small bandwidth generates a wigglier curve. Hastie and Tibshirani [HT90] point out that the choice of bandwidth is much more important than the choice of kernel.

To perform kernel smoothing in S-PLUS, use the `ksmooth` function. The kernels available in `ksmooth` are shown in table 9.2.

Table 9.2: Kernels available for `ksmooth`.

Kernel	Explicit form
`"box"`	$K_{box}(t) = \begin{cases} 1, & \lvert t\rvert \le .5 \\ 0, & \lvert t\rvert > .5 \end{cases}$
`"triangle"`[1]	$K_{tri}(t) = \begin{cases} 1-(\lvert t\rvert/C), & \lvert t\rvert \le \frac{1}{C} \\ 0, & \lvert t\rvert > \frac{1}{C} \end{cases}$
`"parzen"`[2]	$K_{par}(t) = \begin{cases} (k_1 - t^2)/k_2, & \lvert t\rvert \le C_1 \\ (t^2/k_3) - k_4\lvert t\rvert + k_5, & C_1 < \lvert t\rvert \le C_2 \\ 0, & C_2 < \lvert t\rvert \end{cases}$
`"normal"`	$K_{nor}(t) = (1/\sqrt{2\pi}k_6)\exp(-t^2/2{k_6}^2)$

[1] In convolution form, $K_{tri}(t) = K_{box} * K_{box}(t)$
[2] In convolution form, $K_{par}(t) = K_{tri} * K_{box}(t)$

The constants C, C_1, C_2, k_1, ..., k_6 shown in the explicit forms above are used to scale the resulting kernel so that the upper and lower quartiles occur at $\pm.25$. Also, the bandwidth is taken to be 1 and the dependence of the kernel on the bandwidth is suppressed.

Of the available kernels, the default `"box"` kernel gives the crudest smooth. For most data, the other three kernels yield virtually identical smooths. We recommend `"triangle"` because it is the simplest and fastest to calculate.

9.12. Smoothing 9–51

The intuitive sense of the kernel estimate $\hat{y}_i$ is clear: Values of y_j such that x_j is close to x_i get relatively heavy weights, while values of y_j such that x_j is far from x_i get small or zero weight. The bandwidth parameter b determines the width of $K(t/b)$, and hence controls the size of the region around x_i for which y_j receives relatively large weights.

Since bias increases and variance decreases with increasing bandwidth b, selection of b is a compromise between bias and variance in order to achieve small mean squared error. In practice this is usually done by trial and error.

For example, we can compute a kernel smooth for the ethanol data as follows:

```
> plot(E,NOx)
> lines(ksmooth(E,NOx, kernel="triangle", bandwidth=.2))
> lines(ksmooth(E,NOx, kernel="triangle", bandwidth=.1),
+         lty=2)
> legend(.54,4.1,c("bandwidth=.2", "bandwidth=.1"),
+         lty=c(1,2))
```

The resulting plot is shown in figure 9.16.

9.12.4 Smoothing Splines

A *cubic smoothing spline* behaves approximately like a kernel smoother, but it arises as the function $\hat{f}$ that minimizes the *penalized residual sum of squares* given by

$$PRSS = \sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda \int (f''(t))^2\, dt$$

over all functions with continuous first and integrable second derivatives. The parameter λ is the smoothing parameter, corresponding to the span in `loess` or `supsmu` or the bandwidth in `ksmooth`.

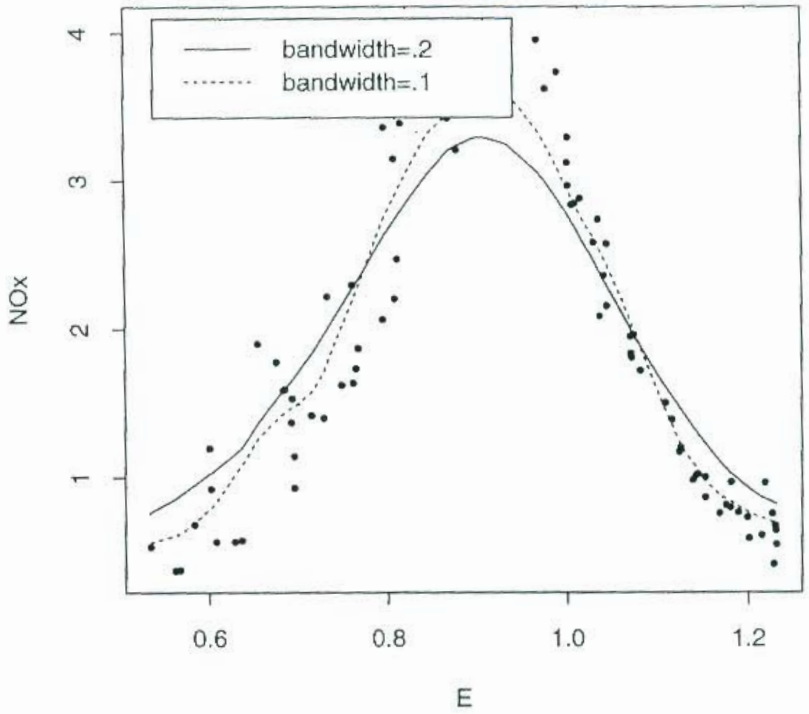

Figure 9.16: Kernel smooth of ethanol data for two bandwidths.

To generate a cubic smoothing spline in S-PLUS, use the function `smooth.spline`, which fits a cubic B-spline smooth to the input data:

```
> plot(E,NOx)
> lines(smooth.spline(E,NOx))
```

You can specify a different λ using the `spar` argument, although it is not intuitively obvious what a "good" choice of λ might be. In general, you should either let S-PLUS choose the smoothing parameter, using either ordinary or generalized cross-validation, or supply an alternative argument, `df`, which specifies the *degrees of freedom* for the smooth. For example,

9.12. Smoothing 9-53

to add a smooth with approximately 5 degrees of freedom to our previous plot, use the following:

```
> lines(smooth.spline(E,NOx, df=5), lty=2)
```

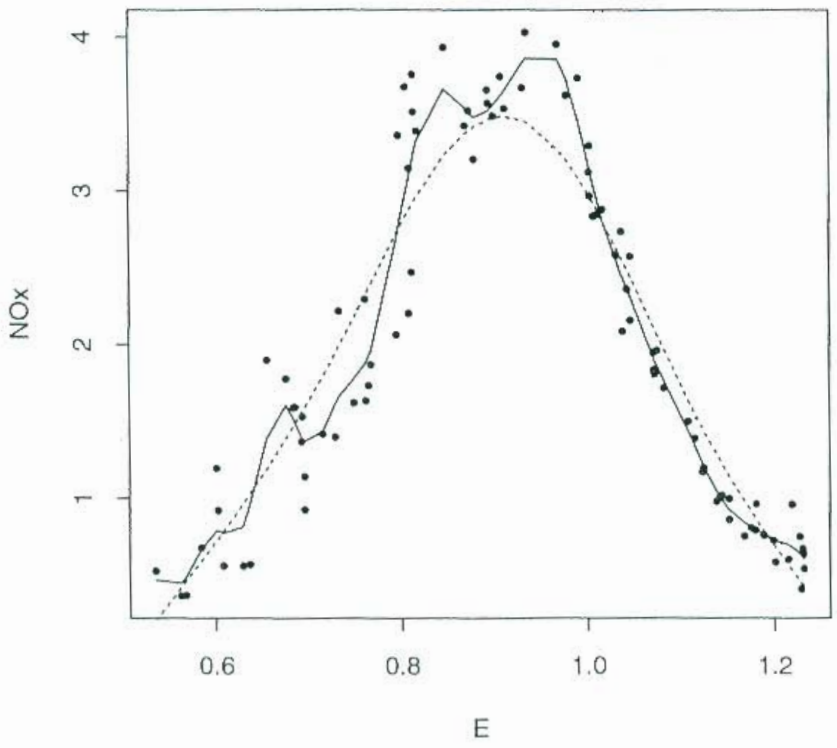

Figure 9.17: Smoothing spline of ethanol data with cross-validation (solid line) and pre-specified degrees of freedom.

The resulting plot is shown in figure 9.17.

9.12.5 Comparing Smoothers

The choice of a smoother is somewhat subjective. All the smoothers discussed in this section can generate reasonably good smooths; you might select one or another based on theoretical considerations or the ease with which one or another of the smoothing criteria can be applied.

For a direct comparision of these smoothers, consider the artificial data constructed as follows:

```
> set.seed(14) # set the seed to reproduce this example
> e <- rnorm(200)
> x <- runif(200)
> y <- sin(2*pi*(1-x)^2)+x*e
```

A "perfect" smooth would recapture the original signal, $f(x) = \sin(2\pi(1-x)^2$, exactly. The following commands sort the input and calculate the *exact* smooth:

```
> sx <- sort(x)
> fx <- sin(2*pi*(1-sx)^2)
```

The following commands create a scatter plot of the original data, then superimpose the exact smooth and smooths calculated using each of the smoothers described in this chapter:

```
> plot(x,y)
> lines(sx,fx)
> lines(supsmu(x,y),lty=2)
> lines(ksmooth(x,y),lty=3)
> lines(smooth.spline(x,y),lty=4)
> lines(loess.smooth(x,y),lty=5)
> legend(0,2,c("perfect", "supsmu", "ksmooth",
+          "smooth.spline", "loess"), lty=1:5)
```


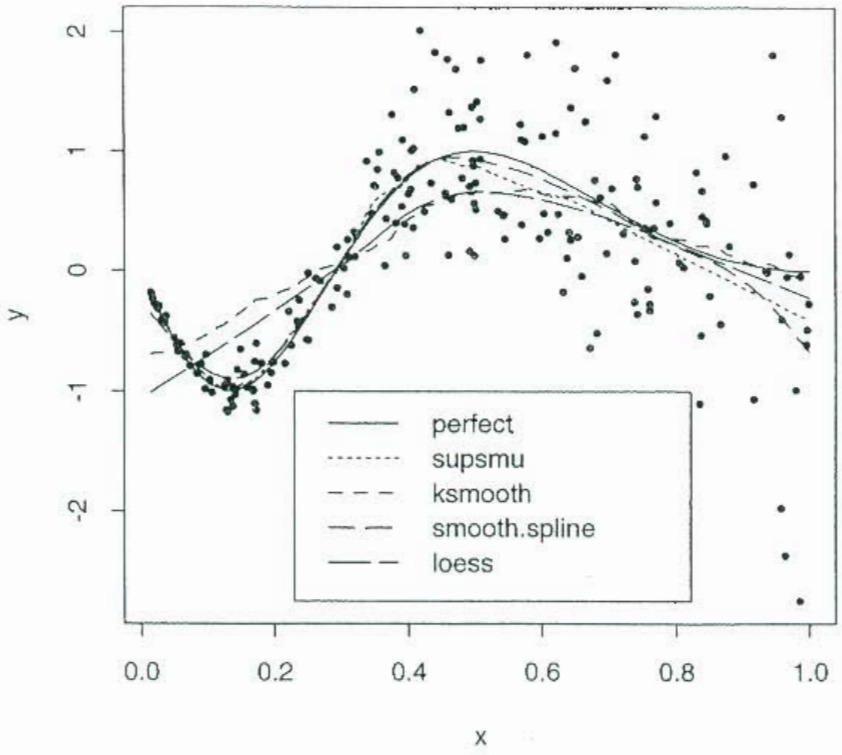

Figure 9.18: Comparison of S-PLUS smoothers.

The resulting plot is shown in figure 9.18. This comparison is crude, at best, because by default each of the smoothers does a different amount of smoothing. A fairer comparison would adjust the smoothing parameters to be roughly equivalent.

9.13 Additive Models

An *additive* model extends the notion of a linear model by allowing some or all linear functions of the predictors to be replaced by arbitrary smooth

The following are examples of the GAM, where the code can be implemented in R.

10–2 *Chapter 10. Generalizing the Linear Model*

particular, we focus on logistic and Poisson regressions and also include a brief discussion of the fitting of models when you can't specify an exact likelihood, using the quasi-likelihood method.

To include gam library in R, go to "packages" "install packages", choose "gam" After installed gam package. load package "gam" or type library(gam) "set CRAN mirror", pick Canada(ON), then

10.1 Logistic Regression

To fit a logistic regression model, use either the `glm` function or the gam function with a formula to specify the model and the `family` argument set to `binomial`. In this case the response variable is necessarily binary or two-valued. As an example, consider the built-in data frame `kyphosis`. A summary of the data frame produces the following:

library(rpart) in R (need to install/update packages if necessary) already built-in in S+ to load the data kyphosis

```
> summary(kyphosis)
```

Kyphosis	Age	Number	Start
absent :64	Min. : 1.00	Min. : 2.000	Min. : 1.00
present:17	1st Qu.: 26.00	1st Qu.: 3.000	1st Qu.: 9.00
	Median : 87.00	Median : 4.000	Median :13.00
	Mean : 83.65	Mean : 4.049	Mean :11.49
	3rd Qu.:130.00	3rd Qu.: 5.000	3rd Qu.:16.00
	Max. :206.00	Max. :10.000	Max. :18.00

The four variables in `kyphosis` are defined as follows:

`Kyphosis`	A binary variable indicating the presence/absence of a postoperative spinal deformity called *Kyphosis*.
`Age`	The age of the child in months.
`Number`	The number of vertebrae involved in the spinal operation.
`Start`	The beginning of the range of the vertebrae involved in the operation.

```
> kyph.gam.all <-
+       gam(Kyphosis ~ s(Age) + s(Number) + s(Start),
+       family = binomial, data = kyphosis)
```

Including each variable as an argument to the `s` function instructs gam to estimate the "smoothed" relationships with each predictor by using cubic B-splines. Alternatively we could have used the `lo` function for local regression smoothing (`loess`). A summary of the fit is:

```
> summary(kyph.gam.all)

Call: gam(formula = Kyphosis ~ s(Age) + s(Number) + s(Start),
family = binomial, data = kyphosis)
Deviance Residuals:
       Min         1Q     Median         3Q      Max
 -1.351358 -0.4439636 -0.1666238 -0.01061843 2.10851

(Dispersion Parameter for Binomial family taken to be 1 )

    Null Deviance: 83.23447 on 80 degrees of freedom

Residual Deviance: 40.75732 on 68.1913 degrees of freedom

Number of Local Scoring Iterations: 7

DF for Terms and Chi-squares for Nonpar          ic Effects
                Df Npar Df Npar Chisq    P(Chi)
(Intercept)      1
     s(Age)      1      2.9    5.782245 0.1161106
  s(Number)      1      3.0    5.649706 0.1289318
   s(Start)      1      2.9    5.802950 0.1139286
```

The summary of a gam fit is similar to the summary of a glm fit. One noticeable difference is the analysis of deviance table. For the additive

fit the tests correspond to *approximate* partial tests for the importance of the smooth for each term in the model. These tests are typically used for screening variables for inclusion in the model. The approximate nature of these tests is discussed in detail in Hastie and Tibshirani [HT90]. For a single variable in the model, this is equivalent to testing for a difference between a linear fit and a smooth fit which includes a linear term along with the smooth term.

Now let's fit two additional models, adding a smooth of each of Age and Number to the base model which has a smooth of Start. We produce the following analysis of deviance tables:

```
> anova(kyph.gam.start, kyph.gam.start.age, test="Chi")
Analysis of Deviance Table

Response: Kyphosis

               Terms Resid. Df Resid. Dev
1           s(Start)  76.24543   59.11262
2 s(Start) + s(Age)  72.09458   48.41713
     Test         Df  Deviance   Pr(Chi)
1
2 +s(Age)  4.150842  10.69548  0.0336071

> anova(kyph.gam.start, kyph.gam.start.number, test="Chi")
Analysis of Deviance Table

Response: Kyphosis

                 Terms Res.Df Res.Dev
1           s(Start) 76.245 59.1126
2 s(Start)+s(Number) 72.180 54.1790
         Test       Df Deviance   Pr(Chi)
1
2 +s(Number) 4.064954 4.933668 0.3023856
```

The indication is that Age is important in the model even with Start included whereas Number isn't important under the same conditions.

You can plot the fit with a smooth on Age and Start adding partial residuals while maintaining all figures on the same scale as follows:

```
> par(mfrow = c(2,2))
> plot(kyph.gam.start.age, resid = T, scale = 8)
```

Or you can simply plot the fit adding pointwise confidence intervals for the fit.

```
> plot(kyph.gam.start.age, se = T, scale = 10)
```

Figure 10.4 displays the resulting plots produced by plot.gam. Notice the vertical axes labels now. They reflect the smoothing operation included in the modeling.

The summary of the additive fit with smooths of Age and Start included appears as follows:

```
> summary(kyph.gam.start.age)

Call: gam(formula = Kyphosis ~ s(Start) + s(Age),
family = binomial, data = kyphosis)
Deviance Residuals:
       Min         1Q     Median         3Q      Max
 -1.694389 -0.4212112 -0.1930565 -0.02753535 2.087434

(Dispersion Parameter for Binomial family taken to be 1 )

    Null Deviance: 83.23447 on 80 degrees of freedom
```


```
Residual Deviance: 48.41713 on 72.09458 degrees of freedom

Number of Local Scoring Iterations: 6

DF for Terms and Chi-squares for Nonparametric Effects

            Df Npar Df Npar Chisq    P(Chi)
(Intercept)  1
   s(Start)  1     2.9   7.729677 0.0497712
     s(Age)  1     3.0   6.100143 0.1039656
```

10.1.3 Returning to the Linear Model

The plots of the fits of the additive model displayed in figure 10.4 suggest a quadratic relationship for `Age` and a piecewise linear relationship for `Start`. It is useful to fit these suggested relationships as a linear model if the model is further simplified without losing too much precision in predicting the response.

For `Age` we fit a second degree polynomial. For `Start`, recall that its values indicate the beginning of the range of the vertebrae involved in the operation. Values less than or equal to 12 correspond to the thoracic region of the spine and values greater than 12 correspond to the lumbar region. Since the relationship for `Start` is family flat for values of `Start` approximately less than or equal to 12 and then drops off linearly for values greater than 12, we will try fitting a linear model with the term `I((Start - 12) * (Start > 12))`. The `I` function is used here to prevent the "*" from being used for factor expansion in the formula sense.

Figure 10.5 displays the resulting fit along with the partial residuals as well as the fit along with two standard errors bands.

10–14 *Chapter 10. Generalizing the Linear Model*

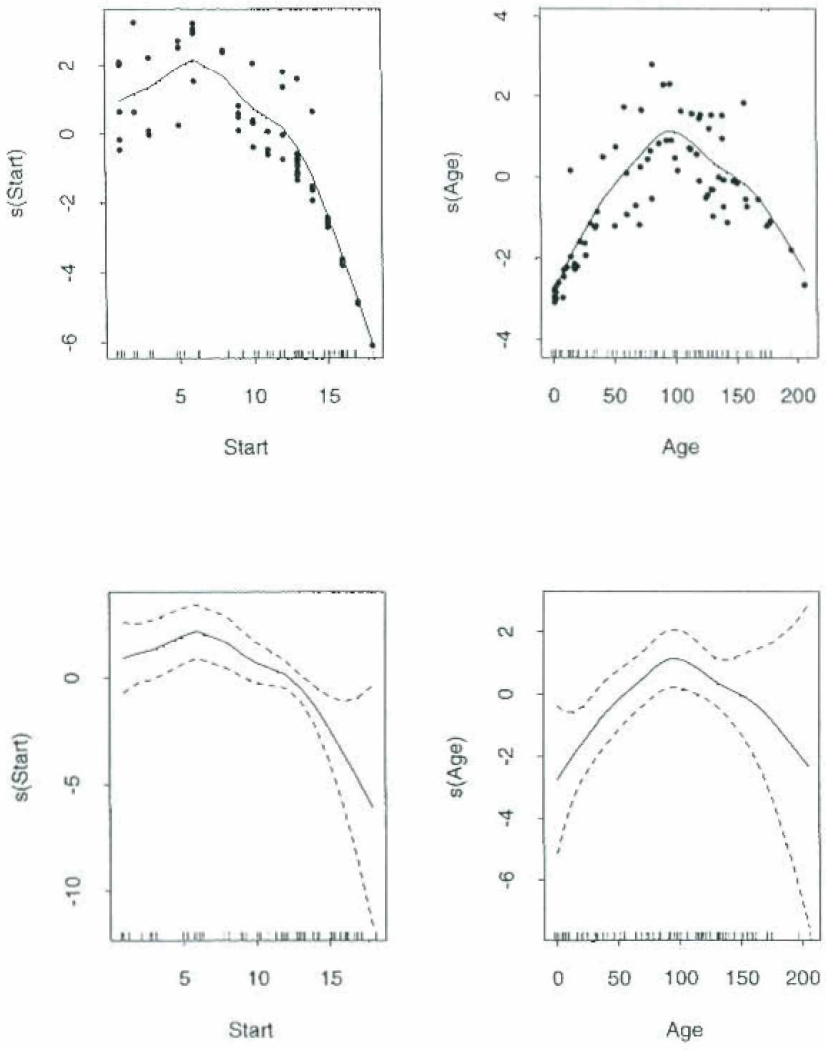

Figure 10.4: The partial fits for the generalized additive logistic regression model of `Kyphosis` with `Age` and `Start` as predictors.

10.1. Logistic Regression 10–15

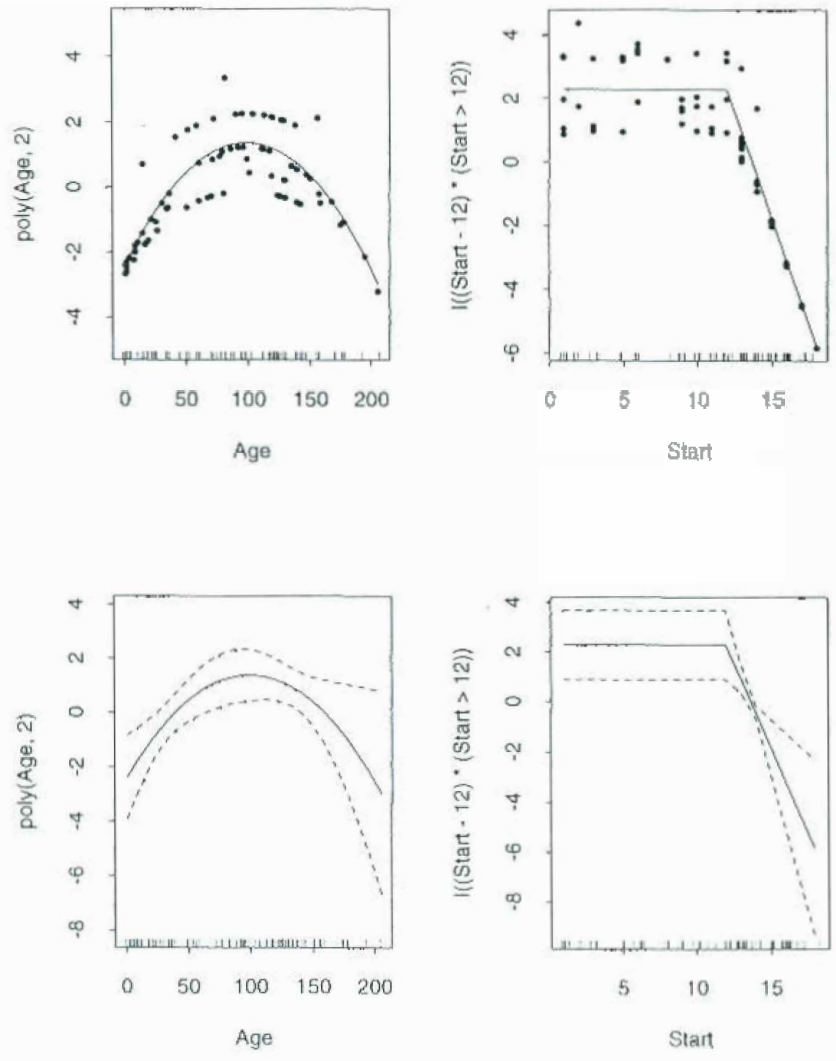

Figure 10.5: The partial fits for the generalized linear logistic regression model of Kyphosis with quadratic fit for Age and piecewise linear fit for Start.

The summary of the fit follows:

```
> summary(kyph.glm.istart.age2)

Call: glm(formula = Kyphosis ~ poly(Age, 2) +
   I((Start - 12) * (Start > 12)), family = binomial,
   data = kyphosis)
Deviance Residuals:
      Min         1Q     Median         3Q      Max
 -1.42301 -0.5014355 -0.1328078 -0.01416602 2.116452

Coefficients:
                                Value Std. Error   t value
             (Intercept)  -0.6849607  0.4570976 -1.498500
          poly(Age, 2)1   5.7719269  4.1315471  1.397038
          poly(Age, 2)2 -10.3247767  4.9540479 -2.084109
I((Start-12)*(Start>12))  -1.3510122  0.5072018 -2.663658

(Dispersion Parameter for Binomial family taken to be 1 )

    Null Deviance: 83.23447 on 80 degrees of freedom

Residual Deviance: 51.95327 on 77 degrees of freedom

Number of Fisher Scoring Iterations: 6

Correlation of Coefficients:
                          (Intercept) poly(Age,2)1 poly(Age,2)2
          poly(Age, 2)1 -0.1133772
          poly(Age, 2)2  0.5625194    0.0130579
I((Start-12)*(Start>12)) -0.3261937   -0.1507199   -0.0325155
```

Contrasting the summary of this linear fit (`kyph.glm.istart.age2`) with the additive fit with smooths of `Age` and `Start` (`kyph.gam.age.start`) we can see the following important details:

10.1. Logistic Regression 10–17

1. The linear fit is more parsimonious; the effective number of parameters being estimated is approximately 5 less than for the additive model with smooths.

2. The residual deviance has increased by only about 3.6 even with a decrease in the effective number of parameters in fitting the linear model by about five. We use the anova function to verify that there is no difference between these models. how nested models .

```
> anova(kyph.glm.istart.age2, kyph.gam.start.age,
+         test="Chi")
Analysis of Deviance Table

Response: Kyphosis

                                          Terms  Res. Df Res. Dev
1 poly(Age,2)+I((Start-12)*(Start>12)) 77.00000 51.95327
2                   s(Start) + s(Age) 72.09458 48.41713

     Test       Df Deviance   Pr(Chi)
1
2 1 vs. 2 4.905415 3.536134 0.6050618
```

3. Having fit a linear model, we can produce an *analytical* expression for the model, which we can't do for an additive model with smooth fits. This is because for a linear model, coefficients are estimated for a *parametric* relationship whereas for an additive model with smooth fits, the smooths are nonparametric estimates of the relationship. In general, these nonparametric estimates have no analytical form and are based on an iterative computer algorithm. This is an important distinction between linear models and additive models with smooth terms. allow for more interpretability .

§23 Mixed Effects & Generalized Estimating Equations

population-averaged vs. subject-specific models

If individual observations are normally distributed, and if we can specify the form of the mean vector through a regression model and the form of the covariance matrix of the data vector, either through outright assumption or using a mixed effects structure, then we can fully specify the particular multivariate normal distribution that will be used as the basis for inference. Because of this, it was quite straightforward to contemplate models for longitudinal normally distributed data. In addition, because we had a full probability model, we could write down the joint probability distribution of the data and use methods of maximum likelihood or REML to fit the model and make inferences.

Slide 1

population-averaged vs. subject-specific models (cont.)

We hope we could do something similar when the elements of a data vector $\mathbf{Y}_i$ are other types of responses, such as count or binary. That is, it would be desirable if there were extensions of Poisson and Bernoulli (and other) distributions that could be fully specified by adding assumptions about correlation to the individual observation assumptions on the mean and variance. Unfortunately, this is not the case. Multivariate extensions of these probability models are not straightforward. There are additional concerns.

There are two perspectives on longitudinal modeling, which we have not routinely discussed as of yet:

1. **population-averaged** (or marginal)
2. **subject-specific**

Slide 2

population-averaged vs. subject-specific models (cont.)

In terms of continuous normally distributed response data:

- with the **population-averaged** (or **marginal**) perspective, we model the mean response of the elements of a data vector by some function of time and possibly other covariates. This function is linear in the parameters, e.g., $E(Y_{i,j}) = \beta_0 + \beta_1 t_{i,j}$. We then model the covariance matrix $var(\mathbf{Y}_i)$ of the data vector explicitly. Here, with $var(\mathbf{Y}_i)$, we would hope to account for all sources of variability, both between and within subjects, simultaneously.
- with the **subject-specific** perspective, using a two-stage process (either directly or indirectly), we model the individual trajectory of the elements by some function of time. This function is linear in the subject-specific parameters, e.g., $Y_{i,j} = \beta_{0,i} + \beta_{1,i} t_{i,j} + \varepsilon_{i,j}$. Then, the subject-specific parameters $\beta_{0,i}$ and $\beta_{1,i}$ are in turn modeled as linear functions of a fixed parameter $\boldsymbol{\beta}$ and random effects $\mathbf{b}_i$, e.g., $\boldsymbol{\beta}_i = \boldsymbol{\beta} + \mathbf{b}_i$, that characterized respectively the "typical" values of the elements of $\boldsymbol{\beta}_i$, and how the individual values deviate from these typical values.

Slide 3

population-averaged vs. subject-specific models (cont.)

- In the subject-specific approach, the end result is a model for the mean response averaged across subjects that is a linear function of time:

  $$E(Y_{i,j}|\mathbf{b}_i) = (\beta_0 + b_{0,i}) + (\beta_1 + b_{1,i}) t_{i,j} \Longrightarrow E(Y_{i,j}) = \beta_0 + \beta_1 t_{i,j}$$

  Here, the covariance matrix $var(\mathbf{Y}_i)$ comes from the combination of assumptions about $\mathbf{b}_i$ and $\boldsymbol{\epsilon}_i$, thus naturally taking into account variation from between and within subjects separately.

  Nevertheless, for the normal linear model for longitudinal data, the result is a model for the mean response $E(Y_{i,j})$ as a linear function of a fixed parameter $\boldsymbol{\beta}$ of interest. The most important feature in this discussion in that *we end up with the exact same linear mean model from either approach, assuming the* $var(\mathbf{Y}_i)$ *is the same.*

Slide 4

population-averaged vs. subject-specific models (cont.)

Turning to generalized linear longitudinal responses, we will see that when such models are non-linear for the mean of $\mathbf{Y}_i$ (e.g., with Poisson or Bernoulli data), it is no longer the case that population-averaged and subject-specific approaches can lead to the same mean model. As a result, the *interpretation of the different types of models are no longer both valid at the same time*.

Slide 5

Generalized linear mixed effect models

We now introduce subject-specific models for generalized longitudinal response data. The most popular terms used for such models include **generalized linear mixed models** or **generalized linear mixed effect models** or simply **GLMM's**.

Where $\mathbf{b}_i$ are random effects and the distribution of the responses given these random effects is in the exponential family, the form of a GLMM is as follows:

- $Y_{i,j}|\mathbf{b}_i \overset{\perp}{\sim} f_{Y_{i,j}|\mathbf{b}_i}(y_{i,j}|\mathbf{b}_i)$
- $f_{Y_{i,j}|\mathbf{b}_i}(y_{i,j}|\mathbf{b}_i) = exp\{\frac{y_{i,j}\gamma_{i,j} - d(\gamma_{i,j})}{\phi} - c(y_{i,j}, \phi)\}$ [14.1]

Slide 6

Generalized linear mixed effect models (cont.)

- $E(Y_{i,j}|\mathbf{b}_i) = \mu_{i,j} = d'(\gamma_{i,j})$
- $g(\mu_{i,j}) = \mathbf{X}_{i,j}^t\boldsymbol{\beta} + \mathbf{Z}_{i,j}^t\mathbf{b}_i$, where $g(\cdot)$ is a known link function linking the conditional mean of $Y_{i,j}$ and the linear form of the predictors (and random effects).
- $var(Y_{i,j}|\mathbf{b}_i) = \phi V(\mu_{i,j}) = \phi V[E(Y_{i,j}|\mathbf{b}_i)] = \phi d''(\gamma_{i,j})$
- Typically, we assume $\mathbf{b}_i \sim N(0, \mathbf{D})$. To be more flexible, we could say $\mathbf{b}_i \sim f(\mathbf{b}_i)$.

Slide 7

Generalized linear mixed effect models (cont.)

Unconditional mean of $Y_{i,j}$

$E(Y_{i,j}) = E[E(Y_{i,j}|\mathbf{b}_i)] = E[\mu_{i,j}] = E[h(\mathbf{X}_{i,j}^t\boldsymbol{\beta} + \mathbf{Z}_{i,j}^t\mathbf{b}_i)]$, where $h(\cdot) = g^{-1}(\cdot)$. This cannot, in general, be simplified further due to g^{-1} being nonlinear.

However, suppose we have a log link, so that $g(\mu) = log(\mu)$ and $g^{-1}(\cdot) = exp(\cdot)$, and assume $b_i \sim N(0, \sigma_b^2)$, where b_i is the sole random effect, a random intercept. Then, it can be shown that $E(Y_{i,j}) = exp(\mathbf{X}_{i,j}^t\boldsymbol{\beta})exp(\sigma_b^2/2)$ or $log[E(Y_{i,j})] = \mathbf{X}_{i,j}^t\boldsymbol{\beta} + \sigma_b^2/2$.

Slide 8

Generalized linear mixed effect models (cont.)

Unconditional variance of $Y_{i,j}$

$$\begin{aligned} var(Y_{i,j}) &= var[E(Y_{i,j}|\mathbf{b}_i)] + E[var(Y_{i,j}|\mathbf{b}_i)] \\ &= var(\mu_{i,j}) + E(\phi V[\mu_{i,j}]) \\ &= var(h[\mathbf{X}_{i,j}^t\boldsymbol{\beta} + \mathbf{Z}_{i,j}^t\mathbf{b}_i]) + E[\phi V(h[\mathbf{X}_{i,j}^t\boldsymbol{\beta} + \mathbf{Z}_{i,j}^t\mathbf{b}_i])] \end{aligned}$$

This cannot be simplified further without specific assumptions about the form of $g(\cdot)$ and/or the conditional distribution of $Y_{i,j}$. One example of explicitly obtaining $var(Y_{i,j})$ is to continue the example from the previous page, where $g(\mu) = log(\mu)$ and $g^{-1}(\cdot) = exp(\cdot)$. Now, further assume that $Y_{i,j}|\mathbf{b}_i \overset{\perp}{\sim} Pois(\mu_{i,j})$.
Here, $E(Y_{i,j}|\mathbf{b}_i) = var(Y_{i,j}|\mathbf{b}_i) = \mu_{i,j}$.
Then, $var(Y_{i,j}) = var(\mu_{i,j}) + E(\mu_{i,j})$.

Slide 9

Generalized linear mixed effect models (cont.)

Unconditional variance of $Y_{i,j}$ (cont.)

We also continue to assume $b_i \sim N(0, \sigma_b^2)$, where b_i is a random intercept. Then, it can be shown that

$$\begin{aligned} var(Y_{i,j}) &= exp\{\mathbf{X}_{i,j}^t\boldsymbol{\beta} + \sigma_b^2/2\} * \\ &\quad (exp\{\mathbf{X}_{i,j}^t\boldsymbol{\beta}\}[exp\{3\sigma_b^2/2\} - exp\{\sigma_b^2/2\}] - 1) \\ &= E(Y_{i,j})K_{i,j} \ . \end{aligned}$$

Note that $K_{i,j} > 1$, so that the variance is larger than the mean. Therefore, although $f(y_{i,j}|\mathbf{b}_i)$ is Poisson, the marginal (or unconditional) distribution of $Y_{i,j}$ cannot be. Under these assumptions, $f(y_{i,j})$ will always be **overdispersed** compared to the Poisson distribution. In this sense, we can think of random effects as a way to model or attribute overdispersion to a particular source (e.g., an individual).

Slide 10

Estimation of GLMM's

Generalized linear mixed effects models pose special challenges beyond standard LME models because of the high-dimensional integration required to evaluate and hence maximize the complicated likelihood. Recalling [14.1], the likelihood for this model is

$$L = \int \prod_{i,j} f_{Y_{i,j}|\mathbf{b}_i}(y_{i,j}|\mathbf{b}_i) f_{\mathbf{b}_i}(\mathbf{b}_i) d\mathbf{b}_i \ .$$

We want to maximize this likelihood to find estimates for $\boldsymbol{\beta}$, ϕ (if necessary), and $\mathbf{D}$.

With only a single random effect (e.g., a random intercept), the likelihood is the product of one-dimensional integrals. Still, with the typical complexity of even these one-dimensional integrals, numerical integration is needed, as the integrals cannot be solved in closed form.

Slide 11

Estimation of GLMM's (cont.)

Approach #1

Numerical quadrature, for example, Gauss-Hermite quadrature. These involve solving integrals of the form $\int_{-\infty}^{\infty} H(v)exp(-v^2)dv$, which can be approximated by a weighted sum $\sum_{k=1}^{r} H(z_k)w_k$, where z_k are evaluation points, w_k are weights, and r = # of quadrature points.

The z_k and w_k are reported in tables, or algorithms are used. Sometimes there is inefficiency with this method, as some quadrature points will be laid out in part of the normal distribution that is not supported by $H(\cdot)$.

Hence, a method was created called **adaptive Gaussian quadrature** that centers and scales the quadrature points as if $H(v)exp(-v^2)$ were a normal distribution. The mean of this normal distribution is the mode $\hat{v}$ of $log[H(v)exp(-v^2)H(v)]$, and the variance will equal
$U = [-\frac{\partial^2}{\partial v^2} log[H(v)exp(-v^2)]|_{v=\hat{v}}]^{-1}$

Slide 12

Estimation of GLMM's (cont.)

Approach #1 (cont.)

New adaptive quadrature points and weignts are based on $\hat{v}$ and U.

Quadrature methods are feasible for low-dimensional integrals (e.g., one or two random effects, possibly nested), but are not amenable to high-dimensional integrals or for crossed random factors or higher levels of nesting. Also, this method is limited by requiring normally distributed random effects or transformations thereof. Finally, the higher the order of r, the better approximation will result (of the N integrals in the likelihood). This comes at a cost of computational efficiency for GQ, and though AGQ needs fewer quadrature points than GQ, it is more computationally intensive.

For more information on these methods, see Crouch and Speigelman (JASA, 1990), Press et al. (text, 1992), McCulloch (JASA, 1994), Liu and Pierce (Biometrika, 1994).

One option in *lmer* function in R, and used by *Proc NLMIXED* in SAS.

Slide 13

Estimation of GLMM's (cont.)

Approach #2

E-M algorithm

1. Choose starting values $\boldsymbol{\beta}^{(m)}$, $\phi^{(m)}$, $\mathbf{D}^{(m)}$, setting $m = 0$.
2. Next, calculate (with expectations evaluated under current values)
   (a) $\boldsymbol{\beta}^{(m+1)}$ and $\phi^{(m+1)}$ to maximize $E[log f_{Y_{i,j}|\mathbf{b}_i}(y_{i,j}|\mathbf{b}_i, \boldsymbol{\beta}, \phi)|y_{i,j}]$
   (b) $\mathbf{D}^{(m+1)}$ to maximize $E[log f_{\mathbf{b}_i}(\mathbf{b}_i\ \mathbf{D})|y_{i,j}]$
   (c) If convergence is achieved, the current values are the MLE's; if not, return to step (a) and continue process until convergence.

In general, expectations in neither steps 2(a) nor 2(b) can be computed in closed form, because the conditional distribution of $\mathbf{b}_i|y_{i,j}$ involves $f_{Y_{i,j}}$, i.e., the likelihood we are trying to avoid calculating directly. Monte Carlo approaches help avoid this problem. See McCulloch (JASA, 1997). Many feel these Monte Carlo methods are the best choice for GLMM's. One problem: available software.

Slide 14

Estimation of GLMM's (cont.)

Approach #3

Bayesian approaches

Flat or diffuse priors are often recommended (e.g., Zeger and Karim (JASA, 1991)), but this can cause problems (Hobert and Casella (JASA, 1996)). For more on these approaches, see Gamerman (Statistics & Computing, 1997), Robert and Casella (text, 1999), McCulloch (JASA, 1997), and Booth and Hobert (JRSS-B, 1999).

Slide 15

Estimation of GLMM's (cont.)

Approach #4

Laplace method

- similar to adaptive Gaussian quadrature with only one quadrature point: approximating integrals of the form $J = \int exp(Q(\mathbf{v}))d\mathbf{v}$
- as such, less computationally intensive than standard quadrature approaches
- based on second-order Taylor series expansion of $Q(\mathbf{v})$ about the value $\hat{\mathbf{v}}$ for which $Q(\cdot)$ is maximized
- for more on this approach, see Gelman et al. (text, 1995) and Raudensbush, Yang, and Yosef (2000), the latter which suggests higher-order terms in the Taylor expansion (up to order 6) which makes improvements close to Gaussian quadrature approaches
- an option in *lmer* in R

Slide 16

Estimation of GLMM's (cont.)

Approach #5

Quasi-Likelihood methods

- based on second-order Taylor series expansion of likelihood about zero (marginal QL) or the random effects (penalized QL)
- has been popular, as it is generally the least difficult computationally
- has bias problems with conditional distributions not close to being normally distributed (e.g., binary longitudinal responses)
- for introduction to method, see Breslow and Clayton (JASA, 1993)
- for discussion of method's limitations, see Breslow and Lin (Biometrika, 1995), and Lin and Breslow (JASA, 1996).

Slide 17

Estimation of GLMM's (cont.)

Approach #5 (cont.)

- SAS used to have a macro (*GLIMMIX*) that utilized this method, but has since developed *Proc NLMIXED*, likely in recognition of some of this method's limitations. *Proc GLIMMIX* may be an improvement over the older macro.
- R has *glmmPQL* in the MASS library, and *lmer* also will do PQL as an option

Approach #6

Stochastic approximation algorithms

Seen as a competitor to one of the Monte Carlo methods, called Monte-Carlo Newton-Raphson (MCNR). For more on MCNR, see McCulloch (JASA, 1997). For more on stochastic approximation algorithms, see Gu and Kong (PNAS, 1998).

Slide 18

Prediction of random effects in GLMM's

Like the linear mixed effects model, obtaining prediction of the random effects $\mathbf{b}_i$ from GLMM's is based on their posterior distribution:

$$f(\mathbf{b}_i|\mathbf{y}_i,\boldsymbol{\beta},\mathbf{D},\phi) = \frac{f(\mathbf{y}_i|\boldsymbol{\beta},\mathbf{D},\phi)f(\mathbf{b}_i|\mathbf{D})}{\int f(\mathbf{y}_i|\boldsymbol{\beta},\mathbf{D},\phi)f(\mathbf{b}_i|\mathbf{D})d\mathbf{b}_i}$$

However, unlike the linear case, this, in general, is not a normal distribution. Hence, the posterior mode, not the posterior mean, is typically used as the prediction for $\mathbf{b}_i$, with $\hat{\mathbf{b}_i}$ the mode of this posterior distribution with estimates for $\boldsymbol{\beta}$, $\mathbf{D}$, and ϕ substituted.

Slide 19

Testing of Hypotheses for GLMM's

- Can use LRT's (with care) for comparing nested models.
  - can use these for testing variance components being equal to 0 with the same caveat as mentioned in the past
  - likelihood cannot be evaluated analytically, in general, so the approximated methods just discussed are utilized; these approximations can be used for AIC (for non-nested models) as well
- Asymptotic variances for the estimated parameters for GLMM's often come from the observed Fisher information (i.e., the negative of the second derivative matrix of the log-likelihood), but this also requires numerical methods; expected Fisher information can be used as well. SAS (and likely *lmer*) uses observed Fisher information.
- Wald tests can be performed for testing on $\boldsymbol{\beta}$ components. First, need to obtain the asymptotic variance of $\boldsymbol{\beta}$ using either observed or expected Fisher information. Use *contrast* argument in *lmer*.

Slide 20

marginal models for generalized longitudinal responses

The **marginal or population-averaged** approach is focused on modeling the mean response across the population of units at each time point as a function of time and possibly other covariates.

In this approach, in general, we will forget about trying to model the entire multivariate probability distribution of a data vector $\mathbf{Y}_i$. Instead, we will just model the mean response and covariance matrix of $\mathbf{Y}_i$. So, *for non-normal responses, when modeling this way, standard likelihood approaches are not possible* (why not?).

We will focus on a model fitting approach, called **generalized estimating equations** or **GEE**, which requires only the mean response and covariance matrix (of $\mathbf{Y}_i$).

Slide 1

marginal models for generalized longitudinal responses (cont.)

We use generalized linear models (GLM's) as the starting point.

- We write the mean response model as $\mu_{i,j} = E(\mathbf{Y}_{i,j}) = h(\mathbf{X}_{i,j}^t\boldsymbol{\beta})$, where h again is the inverse of the link function g (i.e., $g(\mu_{i,j}) = \mathbf{X}_{i,j}^t\boldsymbol{\beta}$).
- The variance of $\mathbf{Y}_{i,j}$ is then modeled by some function of the mean response, $var(\mathbf{Y}_{i,j}) = \phi V(\mu_{i,j})$, where ϕ is a dispersion parameter, sometimes referred to as the overdispersion parameter in the marginal setting (any guesses why?). As an example, in the Bernoulli case, $E(\mathbf{Y}_{i,j}) = \mu_{i,j}$ and $var(\mathbf{Y}_{i,j}) = \phi\mu_{i,j}(1-\mu_{i,j})$.
- The standard deviation of $\mathbf{Y}_{i,j}$ is $\{\phi V(\mu_{i,j})\}^{1/2}$.

Slide 2

marginal models for generalized longitudinal responses (cont.)

- Define the $(n_i\ x\ n_i)$ standard deviation matrix $\mathbf{T}_i$ for subject i:

$$\mathbf{T}_i^{1/2} = \begin{pmatrix} \{V(\mu_{i1})\}^{1/2} & 0 & \cdots & 0 \\ & \{V(\mu_{i2})\}^{1/2} & \cdots & 0 \\ & & \ddots & \vdots \\ & & & \{V(\mu_{in_i})\}^{1/2} \end{pmatrix} .$$

- Let $\boldsymbol{\Gamma}_i$ be the $(n_i\ x\ n_i)$ correlation matrix for subject i. Then, we may write the covariance matrix $\boldsymbol{\Sigma}_i$ for $\mathbf{Y}_i$ as

$$\boldsymbol{\Sigma}_i = \phi\mathbf{T}_i^{1/2}\boldsymbol{\Gamma}_i\mathbf{T}_i^{1/2} .$$

Slide 3

marginal models for generalized longitudinal responses (cont.)

We assume we don't know $\boldsymbol{\Gamma}_i$ ahead of time, but make a reasonable guess for it, $\mathbf{G}_i$. We call $\mathbf{G}_i$ the **working correlation matrix**, as there is uncertainty in how we define the correlation structure. The correlation we consider in these marginal models represents all sources of variation relevant for such models:

1. correlation due to units coming from same subject/cluster (this typically exists)
2. serial correlation (this may exist)

We may then call $\mathbf{S}_i = \phi\mathbf{T}_i^{1/2}\mathbf{G}_i\mathbf{T}_i^{1/2}$ the **working covariance matrix**.

Slide 4

marginal models for generalized longitudinal responses (cont.)

Possible choices of $\mathbf{G}_i$ include:

(a) Unstructered correlation

$$\mathbf{G}_i = \begin{pmatrix} 1 & \rho_{1,2} & \cdots & \rho_{1,n_i} \\ \rho_{2,1} & 1 & \cdots & \rho_{2,n_i} \\ \vdots & & \ddots & \vdots \\ \rho_{n_i,1} & \rho_{n_i,2} & \cdots & 1 \end{pmatrix}$$

where

$\rho_{j,k} = \rho_{k,j}$ for all j, k, $\rho_{j,k} = 1$ if $j = k$, and $-1 \leq \rho_{j,k} \leq 1$ if $j \neq k$.

This correlation matrix depends on $\frac{n_i(n_i+1)}{2}$ distinct correlation parameters, placing no restriction on the nature of the associations among distinct elements of $\mathbf{Y}_i$. Can lead to an overparameterization problem.

Slide 5

marginal models for generalized longitudinal responses (cont.)

(b) Compound symmetry (or exchangeable) correlation

$$\mathbf{G}_i = \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix}$$

This correlation matrix depends on only one distinct correlation parameter. Not as popular for marginal modeling for longitudinal data as it is for other types of clustering, such as family or neighborhood clustering.

Slide 6

marginal models for generalized longitudinal responses (cont.)

(c) One-dependence correlation

$$\mathbf{G}_i = \begin{pmatrix} 1 & \rho & 0 & \cdots & 0 \\ \rho & 1 & \rho & \cdots & 0 \\ \vdots & & \ddots & & \vdots \\ \vdots & & & \ddots & \rho \\ 0 & \cdots & \cdots & \rho & 1 \end{pmatrix}$$

Also depends on only one distinct correlation parameter. Only observations adjacent in time are assumed correlated.

Slide 7

marginal models for generalized longitudinal responses (cont.)

(d) AR(1) correlation

$$\mathbf{G}_i = \begin{pmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{n_i-1} \\ \rho & 1 & \rho & \cdots & \vdots \\ \vdots & & \ddots & & \rho^2 \\ \vdots & & & \ddots & \rho \\ \rho^{n_i-1} & \cdots & \rho^2 & \rho & 1 \end{pmatrix}$$

Again, just one correlation parameter. Here, the correlation parameter tails off as the observations get further separated in time.

Slide 8

marginal models for generalized longitudinal responses (cont.)

(e) Spatial power structure (or CAR(1))

$$\mathbf{G}_i = \begin{pmatrix} 1 & \rho^{|t_1-t_2|} & \rho^{|t_1-t_3|} & \cdots & \rho^{|t_1-t_{n_i}|} \\ \rho^{|t_2-t_1|} & 1 & \rho^{|t_2-t_3|} & \cdots & \vdots \\ \vdots & & \ddots & & \vdots \\ \vdots & & & \ddots & \rho^{|t_{n_i-1}-t_{n_i}|} \\ \rho^{|t_{n_i}-t_1|} & \cdots & \cdots & \rho^{|t_{n_i}-t_{n_i-1}|} & 1 \end{pmatrix}$$

The continuous time analog to AR(1).

Slide 9

marginal models for generalized longitudinal responses (cont.)

(f) Independence

$$\mathbf{G}_i = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & \cdots & 1 \end{pmatrix}$$

No correlation parameters. Typically unrealistic (why?), though it is used occasionally.

Unbalanced data, especially data measured at different time points for different individuals/clusters, may make some of the above correlation structures (e.g., exchangeable or spatial) more reasonable than others (e.g., one-dependence or AR(1)).

Slide 10

marginal models for generalized longitudinal responses (cont.)

In summary:

- mean response of $\mathbf{Y}_{i,j}$ is modeled by a suitable function h (inverse link) of a linear predictor $\mathbf{X}_{i,j}^t\boldsymbol{\beta}$
- variance is then modeled as some function of the mean response; it is composed of a standard deviation, leading to the s.d. matrix $\mathbf{T}_i^{1/2}$, and an overdispersion parameter ϕ
- correlation is modeled by a working correlation assumption $\mathbf{G}_i$; let ω be the vector of unknown correlation parameters that fully characterizes $\mathbf{G}_i$

Slide 11

marginal models for generalized longitudinal responses (cont.)

In summary (cont.):

- $E(\mathbf{Y}_i) = \begin{pmatrix} h(\mathbf{X_{i,1}}^t\boldsymbol{\beta}) \\ h(\mathbf{X_{i,2}}^t\boldsymbol{\beta}) \\ \vdots \\ h(\mathbf{X_{i,n_i}}^t\boldsymbol{\beta}) \end{pmatrix}$
- $cov(\mathbf{Y}_i) = \phi\mathbf{T}_i^{1/2}\boldsymbol{\Gamma}_i\mathbf{T}_i^{1/2} = \boldsymbol{\Sigma}_i$; $\mathbf{S}_i = \phi\mathbf{T}_i^{1/2}\mathbf{G}_i\mathbf{T}_i^{1/2}$
- assume $\mathbf{Y}_i$ are independent across individual subjects/clusters

Goal: Find estimate of $\boldsymbol{\beta}$ knowing we cannot use maximum likelihood techniques; also, obtain var($\hat{\boldsymbol{\beta}}$) and estimates of ϕ, ω, and then $\boldsymbol{\Sigma}_i$.

Slide 12

generalized estimating equations

Our approach for obtaining parameter estimates for a marginal model is to solve an estimating equation consisting of p equations for $\boldsymbol{\beta}_{(p\ x\ 1)}$ that:
(i) is a linear function of deviations $(\mathbf{Y}_i - h_i(\boldsymbol{\beta}))$
and
(ii) weights these deviations using the inverse of the assumed covariance matrix $\mathbf{S}_i$ of $\mathbf{Y}_i$, with an estimator for the unknown parameters ω in $\mathbf{G}_i$ plugged in.

Recall when we discussed generalized linear models, we presented the following estimating equation for the estimation of $\boldsymbol{\beta}$, which was based on maximum likelihood, via iteratively reweighted least squares (IRLS):

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{N}(\frac{\partial\mu_i}{\partial\boldsymbol{\beta}})^t v_i^{-1}\{y_i - \mu_i(\boldsymbol{\beta})\} = 0\ ,$$

where $v_i = var(Y_i)$.

Slide 13

generalized estimating equations (cont.)

The IRLS approach for the independent case leads to a similarly formed estimating equation to solve for $\boldsymbol{\beta}$ in the longitudinal setting for generalized response data:

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{N}\boldsymbol{\Delta}_i\boldsymbol{\Sigma}_i^{-1}\{y_i - h_i(\boldsymbol{\beta})\} = 0\ , \qquad [15.1]$$

where $\boldsymbol{\Delta}_i = (\frac{\partial h_i(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}})_{(n_i\ x\ p)}$, $\boldsymbol{\Sigma}_i = cov(\mathbf{Y}_i)$, $h_i(\boldsymbol{\beta}) = \mu_i$ (i.e., $\mu_i(\boldsymbol{\beta})$).

We will solve for $\boldsymbol{\beta}$, plugging in estimates of $\boldsymbol{\Sigma}_i$, which uses estimates of ω and ϕ. The solution for $\boldsymbol{\beta}$ yields $\hat{\boldsymbol{\beta}}$. Equation [15.1] is called a **generalized estimating equation**.

Slide 14

generalized estimating equations (cont.)

We will discuss only a few details in finding $\hat{\phi}$ and $\hat{\omega}$.

Think of completely independent case, where all observations both within and between subjects are independent, and fitting a mean model in this case. Let

$$r_{i,j} = \frac{\mathbf{Y}_{i,j} - h(\mathbf{X}_{i,j}^t\hat{\boldsymbol{\beta}})}{[V\{h(\mathbf{X}_{i,j}^t\hat{\boldsymbol{\beta}})\}]^{1/2}}\ .$$

Then,

$$\hat{\phi} = \frac{1}{n-p}\sum_{i=1}^{N}\sum_{j=1}^{n_i}\frac{\{\mathbf{Y}_{i,j} - h(\mathbf{X}_{i,j}^t\hat{\boldsymbol{\beta}})\}^2}{V\{h(\mathbf{X}_{i,j}^t\hat{\boldsymbol{\beta}})\}} = \frac{1}{n-p}\sum_{i=1}^{N}\sum_{j=1}^{n_i}r_{i,j}^2\ ,$$

where $n = \sum_{i=1}^{N} n_i$. This is similar to a chi-square statistic in ordinary GLM's, divided by its df, now taken across all deviations on all subjects/clusters.

Slide 15

generalized estimating equations (cont.)

Regarding ω, if $\mathbf{G}_i$ corresponds to unstructured correlation assumption, then

$$\hat{\rho}_{j,k} = \frac{1}{N}\frac{1}{\hat{\phi}}\sum_{i=1}^{N}r_{i,j}r_{i,k}\ .$$

If $\mathbf{G}_i$ corresponds to exchangeable correlation assumption, then

$$\hat{\rho} = \frac{1}{N}\frac{1}{\hat{\phi}}\sum_{i=1}^{N}\frac{1}{n_i-1}\sum_{j=1}^{n_i-1}r_{i,j}r_{i,j+1}\ ,$$

where consideration here is only taken for adjacent pairs.

Slide 16

generalized estimating equations (cont.)

Heuristically, we take the following steps:

1. Obtain an initial estimate for $\boldsymbol{\beta}$ by assuming all observations across all subjects are independent; this may be carried out by method of IRLS for ordinary GLM's.
2. Using initial estimator of $\boldsymbol{\beta}$ from (1), estimate ϕ and then ω as appropriate, for the assumed working correlation matrix.
3. Use estimators for $\boldsymbol{\beta}$, ϕ, and ω to form estimate of $\boldsymbol{\Sigma}_i$. Now, treat $\hat{\boldsymbol{\Sigma}}_i$ as fixed in GEE [15.1]. The resulting equation can be solved by numerical techniques (essentially an extended version of IRLS). This yields a new estimator $\boldsymbol{\beta}$.
4. Return to step (2) and repeat the process, until convergence.

The above is essentially what is followed in *gee* function (from gee package) in R (and in *Proc GENMOD* in SAS).

Slide 17

generalized estimating equations (cont.)

For N large, the GEE estimator $\hat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ satisfies:

$$\hat{\boldsymbol{\beta}} \dot{\sim} N_p(\boldsymbol{\beta}, V_{\hat{\boldsymbol{\beta}}}) \ ,$$

where $V_{\hat{\boldsymbol{\beta}}} = var(\hat{\boldsymbol{\beta}}) = \phi(\sum_{i=1}^{N} \boldsymbol{\Delta}_i^t \mathbf{S}_i^{-1} \boldsymbol{\Delta}_i)^{-1}$

This consistency of $\hat{\boldsymbol{\beta}}$ occurs even if our guess for $\boldsymbol{\Sigma}_i$, i.e., $\mathbf{S}_i$, is misspecified.

Note when $\boldsymbol{\Sigma}_i = \mathbf{S}_i$, we get $V_{\hat{\boldsymbol{\beta}}} = var(\hat{\boldsymbol{\beta}}) = \phi(\sum_{i=1}^{N} \boldsymbol{\Delta}_i^t \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\Delta}_i)^{-1}$, which is the usual full likelihood based variance.

Slide 18

generalized estimating equations (cont.)

For practical purposes, we do not know ϕ, ω, or $\boldsymbol{\beta}$, so we use their estimates to obtain $\hat{V}_{\hat{\boldsymbol{\beta}}} = \hat{var}(\hat{\boldsymbol{\beta}}) = \hat{\phi}(\sum_{i=1}^{N} \hat{\boldsymbol{\Delta}}_i{}^t \hat{\mathbf{S}}_i{}^{-1} \hat{\boldsymbol{\Delta}}_i)^{-1}$.

Standard errors for elements of $\hat{\boldsymbol{\beta}}$ may be obtained as square roots of diagonal elements of $V_{\hat{\boldsymbol{\beta}}}$ (or $\hat{V}_{\hat{\boldsymbol{\beta}}}$).

Wald tests can follow as we have implemented in the past.

Slide 19

generalized estimating equations (cont.)

Robust estimate of covariance of $\hat{\boldsymbol{\beta}}$

We would like to modify the estimated covariance matrix $\hat{V}_{\hat{\boldsymbol{\beta}}}$ to allow for better protection against the possibility that the choice of the correlation structure $\mathbf{G}_i$ was misspecified. This modified version, otherwise known as the **robust estimate of covariance** of $\hat{\boldsymbol{\beta}}$, is:

$$\hat{V}^R_{\hat{\boldsymbol{\beta}}} =$$
$$\hat{\phi}(\sum_{i=1}^{N} \hat{\boldsymbol{\Delta}}_i{}^t \hat{\mathbf{S}}_i{}^{-1} \hat{\boldsymbol{\Delta}}_i)^{-1} (\sum_{i=1}^{N} \hat{\boldsymbol{\Delta}}_i{}^t \hat{\mathbf{S}}_i{}^{-1} \hat{C}_i \hat{\mathbf{S}}_i{}^{-1} \hat{\boldsymbol{\Delta}}_i)(\sum_{i=1}^{N} \hat{\boldsymbol{\Delta}}_i{}^t \hat{\mathbf{S}}_i{}^{-1} \hat{\boldsymbol{\Delta}}_i)^{-1} \ ,$$
where $\hat{C}_i = \{\mathbf{Y}_i - h_i(\hat{\boldsymbol{\beta}})\}\{\mathbf{Y}_i - h_i(\hat{\boldsymbol{\beta}})\}^t$.

Slide 20

generalized estimating equations (cont.)

Notes regarding $\hat{V}^R_{\hat{\boldsymbol{\beta}}}$

- If $\hat{C}_i = \hat{\mathbf{S}}_i$, then $\hat{V}^R_{\hat{\boldsymbol{\beta}}} = \hat{V}_{\hat{\boldsymbol{\beta}}}$, noting that $var(\mathbf{Y}_i) = \boldsymbol{\Sigma}_i = E[\{\mathbf{Y}_i - h_i(\boldsymbol{\beta})\}\{\mathbf{Y}_i - h_i(\boldsymbol{\beta})\}^t]$.
- In the model, we have chosen $\mathbf{S}_i$ through choosing $\mathbf{G}_i$ as the working correlation structure. By including the middle term (i.e., the **sandwich** term), we hope to balance out an alternative "guess" for $\mathbf{S}_i$ vs. the current assumed model $\mathbf{S}_i$.
- For large N, $\hat{V}^R_{\hat{\boldsymbol{\beta}}}$ will provide a reliable (unbiased) estimate of the true sampling covariance matrix of $\hat{\boldsymbol{\beta}}$ even if the chosen model of $\mathbf{S}_i$ (through $\mathbf{G}_i$) is incorrect. In contrast, if the model is incorrect, $\hat{V}_{\hat{\boldsymbol{\beta}}}$ will not provide a reliable estimate.
- $\hat{V}^R_{\hat{\boldsymbol{\beta}}}$ is called the **robust estimate** of the covariance of $\hat{\boldsymbol{\beta}}$; also called the **empirical estimate**; $\hat{V}_{\hat{\boldsymbol{\beta}}}$ is called the **model-based estimate**.

Slide 21

generalized estimating equations (cont.)

Notes regarding $\hat{V}^R_{\hat{\boldsymbol{\beta}}}$ (cont.)

- In finite samples, choosing between $\hat{V}_{\hat{\boldsymbol{\beta}}}$ and $\hat{V}^R_{\hat{\boldsymbol{\beta}}}$ is not as clear as in "large" samples (where $\hat{V}^R_{\hat{\boldsymbol{\beta}}}$ is generally the better choice).
  - If $\hat{V}_{\hat{\boldsymbol{\beta}}}$ is very different from $\hat{V}^R_{\hat{\boldsymbol{\beta}}}$, some say that is an indication that the original assumption for $\mathbf{S}_i$ (via $\mathbf{G}_i$) is wrong.
  - On the other hand, if one or more of observed $\mathbf{Y}_i$ vectors contains unusual values, which are unlikely to be seen, this may be enough to cause problems for $\hat{V}^R_{\hat{\boldsymbol{\beta}}}$.
  - $\hat{V}^R_{\hat{\boldsymbol{\beta}}}$ may also be sub-optimal in very unbalanced designs or those where there are few replications, if any, of $\mathbf{Y}_i$. associated with each distinct set of covariate values (e.g., in observational studies with many covariate combinations).
  - So, deciding between $\hat{V}^R_{\hat{\boldsymbol{\beta}}}$ and $\hat{V}_{\hat{\boldsymbol{\beta}}}$ is problem-dependent.
- $\hat{\boldsymbol{\beta}}$ is generally robust to working correlation misspecification. From above, $\hat{V}^R_{\hat{\boldsymbol{\beta}}}$ is, in general, more reliable than $\hat{V}_{\hat{\boldsymbol{\beta}}}$ under such misspecification.

Slide 22

subject-specific vs. population-averaged (marginal) models

Recall: Use GLMM's for subject-specific models.
Use marginal modeling / GEE for population-averaged models.

With GLMM:

- recall $E(\mathbf{Y}_{i,j}|\mathbf{b}_i) = \mu_{i,j},\ g(\mu_{i,j}) = \mathbf{X}_{i,j}^t\boldsymbol{\beta} + \mathbf{Z}_{i,j}^t\mathbf{b}_i$, $var(\mathbf{Y}_{i,j}|\mathbf{b}_i) = \phi V(\mu_{i,j})$, and $\mathbf{b}_i \sim N_q(0, D)$.
- the goal of the analysis is to estimate $\boldsymbol{\beta}$, the variance components of $\mathbf{D}$, maybe $\mathbf{b}_i$, and possibly ϕ, when $\phi \neq 1$.
- here, the parameter $\boldsymbol{\beta}$ describes on average how an individual's response, via the link function (e.g., log odds of the individual response for binary data), changes with $\mathbf{X}$.

Slide 23

subject-specific vs. population-averaged (marginal) models (cont.)

With marginal model:

- recall $E(\mathbf{Y}_{i,j}) = \mu_{i,j},\ g(\mu_{i,j}) = \mathbf{X}_{i,j}^t\boldsymbol{\beta}$, $var(\mathbf{Y}_i) = \phi\mathbf{T}_i^{1/2}\boldsymbol{\Gamma}_i\mathbf{T}_i^{1/2} = \boldsymbol{\Sigma}_i$.
- the goal of the analysis is to estimate $\boldsymbol{\beta}$, ϕ, and $\boldsymbol{\Sigma}_i$ (based on assumption on form of $\boldsymbol{\Gamma}_i$, via the working correlation matrix $\mathbf{G}_i$).
- here, the parameter $\boldsymbol{\beta}$ describes on average how the population response, via the link function (e.g., log odds of the population response for binary data), changes with $\mathbf{X}$.

Slide 24

subject-specific vs. population-averaged (marginal) models (cont.)

Comments/Recommendations

(1) SS models are desirable when the response for an individual rather than for the population is the focus, e.g., studies of growth curves.

PA models are most effectively used in population studies, e.g., large-scale epidemiological studies of population over time.

(2) As an example of interpretation, if $\mathbf{X}_{i,j}$ indicates whether subject i smokes at time j and $\mathbf{Y}_{i,j}$ is the presence/absence of respiratory infection, the SS model estimates the expected change in individual's probability of infection given a change in smoking status, whereas the PA model estimates the difference in infections rates between smokers and non-smokers.

Slide 25

subject-specific vs. population-averaged (marginal) models (cont.)

Comments/Recommendations (cont.)

(3) In SS models, ultimate inference for $\boldsymbol{\beta}$ may depend on choice of g, as well as on assumption for distribution of $\mathbf{b}_i$.
In PA models, ultimate inference for $\boldsymbol{\beta}$ may depend on choice of g, and, usually to a lesser extent, on the working correlation assumption.

(4) *$\boldsymbol{\beta}$ has same interpretation in SS model and PA model for identity link*, i.e., the normal linear model. The random effects in the linear mixed effects model do not alter the marginal expectation of $\mathbf{Y}$, only the marginal covariance matrix.

(5) Better to have more obs per subject for SS analysis.
Better to have more subjects for PA analysis.

(6) PA analyses are computationally less complex than SS analyses for generalized longitudinal response data.

Slide 26

The End