

# BIOS 762: Theory and Applications of Linear and Generalized Linear Models

Joseph G. Ibrahim

Department of Biostatistics  
University of North Carolina at Chapel Hill

Fall, 2019

# Chapter 1: Linear Regression and Analysis of Variance

Background: Some mathematical statistics and some exposure to linear algebra. The first 1.5 weeks of this course will be a review of linear algebra. This is covered in appendices A and B of Christensen. We will do a bit more than what Christensen does.

Reference books: (Linear Algebra)

- Halmos – Finite Dimensional Vector Spaces
- Strang – Linear Algebra and its Applications
- Schaum's outline series to linear algebra

Reference books to Linear Models:

- Guttman – Linear Models
- Searle – Linear Models
- Scheffe – The Analysis of Variance
- Seber – Linear Regression Analysis
- Graybill – Theory and Application of the Linear Model
- Rao – Linear Statistical Inference and its Applications

## Comment on Christensen's book:

Excellent book. He takes a geometric approach to linear models, unlike any other existing book on linear models.

The geometric approach has several advantages: 1) Elegant theory, 2) easy notation and formulas, 3) no crazy calculus, quadruple sums etc . . .

## Goal for the course:

This mainly a theoretical course in linear and generalized linear models. The course will involve data analysis and applications, but the main goal is to provide a solid theoretical background in linear and generalized linear models.

# Motivating Example # 1

## Motivation

### Example 1 – Simple Linear Regression

We have observations  $(x_1, y_1), \dots, (x_n, y_n)$ , where the  $x_i$ 's are known fixed values, and the  $y_i$ 's are response variables (random). We have the model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i , \quad i = 1, \dots, n$$

$y_i$  = annual melanoma mortality in state  $i$ .

$x_i$  = latitude of the center of the state (in degrees).

### Usual assumptions:

The  $\epsilon_i$ 's are *i.i.d.* from some distribution with

$$E(\epsilon_i) = 0 \text{ and } \text{Var}(\epsilon_i) = \sigma^2.$$

# Motivating Example # 1

With this model, we may want to

- ① Estimate  $\beta_0$ ,  $\beta_1$ , and  $\sigma^2$ .
- ② Test hypotheses about  $\beta_1$ , confidence limits for  $\beta_1$ .
- ③ Predict a future  $y$  at a given  $x$ .

There may be other goals:

To do exact inference, we need  $\epsilon_i$  to be normally distributed. Otherwise, we have to base inferences on large sample theory.

# Motivating Example # 2

## Example 2 – One Way ANOVA

Suppose we have observations  $y_{ij}$ ,  $j = 1, \dots, n_i$ ,  $i = 1, \dots, k$ . We have  $k$  populations with  $n_i$  observations in population  $i$ .

$$\begin{array}{cccc} pop1 & pop2 & \dots & popk \\ y_{11} & y_{21} & \dots & y_{k1} \\ \vdots & \vdots & & \vdots \\ y_{1n_1} & y_{2n_2} & \dots & y_{kn_k} \end{array} \quad (1)$$

Suppose  $E(y_{ij}) = \mu_i$  and  $\text{Var}(y_{ij}) = \sigma^2$ .

Example: We want to examine the effect of  $NO_2$  on the lungs. Consider mice which are i) not exposed ii) mildly exposed, iii) heavily exposed. ( $k = 3$ ). Response variable is percent serum fluorescence. High readings indicate damage to lung tissues.

## Motivating Example # 2

Alternatively, we can write:

$$\begin{aligned}y_{ij} &= \mu_i + (y_{ij} - \mu_i) \\&= \mu_i + \epsilon_{ij},\end{aligned}$$

$j = 1, \dots, n_i, i = 1, \dots, k.$

The  $\epsilon_{ij}$ 's are *i.i.d.* from a distribution with  $E(\epsilon_{ij}) = 0$ , and  $\text{Var}(\epsilon_{ij}) = \sigma^2$ . The model

$$y_{ij} = \mu_i + \epsilon_{ij},$$

$j = 1, \dots, n_i, i = 1, \dots, k,$  is often referred to as a means model.

## Motivating Example # 2

We may want to

- ① Estimate the  $\mu_i$ 's and  $\sigma^2$ .
- ② Test hypotheses about the  $\mu_i$ 's, e.g.

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_k$$

or

$$H_0 : c_1\mu_1 + c_2\mu_2 + \dots + c_k\mu_k = 0 ,$$

where  $\sum_{i=1}^k c_i = 0$ .

- ③ Decide which group mean is largest, multiple comparisons, etc ....

If we reparameterize and write

$$\mu_i = \mu + \alpha_i,$$

then the model becomes

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij} ,$$

$$j = 1, \dots, n_i, i = 1, \dots, k.$$

## Motivating Example # 3 - General Linear Model

### Example 3 - General Linear Model

Both of the examples just presented and many others are special cases of the general linear model.

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + \epsilon_i,$$

$$i = 1, \dots, n.$$

$y_i$ 's are observations.

$x_{ij}$ 's are known (fixed) values.

$\beta_j$ 's are unknown parameters.

$\epsilon_i$ 's are unobservable random variables with

$$E(\epsilon_i) = 0, \quad \text{Cov}(\epsilon_i, \epsilon_j) = \sigma_{ij}, \quad i, j = 1, \dots, n.$$

This is the most general setup. Note: If an intercept is included in the model, then

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} + \epsilon_i.$$

## Motivating Example # 3 - General Linear Model

Often, more restrictive assumptions are made about the  $\epsilon_i$ 's.

- ① Uncorrelated:  $\text{Cov}(\epsilon_i, \epsilon_j) = 0$  for  $i \neq j$ .
- ② Equal variances:  $\text{Cov}(\epsilon_i, \epsilon_i) = \text{var}(\epsilon_i) = \sigma^2$ ,  $i = 1, \dots, n$ .
- ③ the  $\epsilon_i$ 's are normally distributed.

Example:

$y_i$  = oxygen consumption

$x_1$  = treadmill duration

$x_2$  = heart rate

$x_3$  = age

$x_4$  = height

$x_5$  = weight.

## Motivating Example # 3 - General Linear Model

We can write this model in matrix notation. Define

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}_{n \times 1} \quad X = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix}_{n \times p}$$

$$\beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}_{p \times 1}, \quad \epsilon = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}_{n \times 1}$$

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ & \sigma_{22} & & \vdots \\ & & \ddots & \vdots \\ & & & \sigma_{nn} \end{pmatrix}_{n \times n}.$$

We can now write the model as

$$Y = X\beta + \epsilon, \\ E(\epsilon) = 0, \quad \text{Cov}(\epsilon) = \Sigma,$$

## Motivating Example # 3

or often

$$\text{Cov}(\epsilon) = \sigma^2 I,$$

in the uncorrelated, equal variance errors case.

Note: In this formulation, the model says that

$$\begin{aligned}\mu &= E(Y) = X\beta \\ &= [X_1, X_2, \dots, X_p]\beta \\ &= \beta_1 X_1 + \dots + \beta_p X_p,\end{aligned}$$

where

$$X_j = \begin{pmatrix} x_{1j} \\ \vdots \\ x_{nj} \end{pmatrix}$$

is the  $j$ th column of  $X$ ,  $j = 1, \dots, p$ .

We see that  $\mu$  is a linear combination of the columns of  $X$ .

## Motivating Example # 3 - General Linear Model

In a more abstract setting,

$$\mu \in \Omega,$$

where  $\Omega$  is a linear subspace in an  $n$  dimensional vector space.

Often, the theory of linear models is done by specifying the model as

$$E(Y) = \mu, \quad \text{cov}(Y) = \Sigma$$

Without specifying a specific structure (coordinatization) on  $\mu$ , that is

$$\mu = X\beta.$$

Specifying the model by  $E(Y) = \mu$ ,  $\text{Cov}(Y) = \Sigma$  is called a coordinate-free specification. No coordinate system is specified for  $\mu$ .

The moment we write  $\mu = X\beta$ , then we have an implied coordinate system. That is, we have a coordinatization for the components of  $\mu$ . The most general treatment of the theory of linear models is done with the coordinate free approach.

Christensen does not take a coordinate free approach to linear models. The theory is essentially just as general in the non-coordinate free setup. We too will not delve into the coordinate free theory too much. Practical issues are lost when we get too theoretical.

## Motivating Example # 3 - General Linear Model

In the general linear model, we may be interested in

- ① estimating  $\beta$ ,  $\Sigma$
- ② constructing tests concerning  $\beta$ .
- ③ model selection – which columns of  $X$  are necessary.
- ④ prediction of a future  $Y$ .

The one-way ANOVA model introduced earlier:

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij},$$

$j = 1, \dots, n_i$ ,  $i = 1, \dots, k$ , can now be written in matrix notation as

$$\begin{aligned} Y &= X\beta + \epsilon, \\ \epsilon &\sim N_N(0, \sigma^2 I), \end{aligned}$$

where

## Motivating Example # 3 - General Linear Model

$$Y = \begin{pmatrix} y_{11} \\ \vdots \\ y_{1n_1} \\ y_{21} \\ \vdots \\ y_{kn_k} \end{pmatrix}_{N \times 1}, X = \begin{bmatrix} 1 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ \vdots & 1 & 0 & \dots & \vdots \\ \vdots & 0 & 1 & \dots & \vdots \\ \vdots & \vdots & \vdots & & \vdots \\ \vdots & \vdots & 1 & \dots & 0 \\ \vdots & \vdots & 0 & \dots & 1 \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & 0 & 0 & \dots & 1 \end{bmatrix}_{N \times (k+1)}$$

and  $N = \sum_{i=1}^k n_i$ .

## Motivating Example # 3 - General Linear Model

$$\beta = \begin{pmatrix} \mu \\ \alpha_1 \\ \vdots \\ \alpha_k \end{pmatrix}_{(k+1) \times 1} \quad \epsilon = \begin{pmatrix} \epsilon_{11} \\ \vdots \\ \epsilon_{1n_1} \\ \epsilon_{21} \\ \vdots \\ \epsilon_{2n_2} \\ \vdots \\ \epsilon_{kn_k} \end{pmatrix}_{N \times 1}$$

## Motivating Example # 3 - General Linear Model

For example if

$$k = 3, n_1 = 3, n_2 = 1, n_3 = 2, N = 3 + 1 + 2 = 6$$

$$Y = \begin{pmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{31} \\ y_{32} \end{pmatrix}_{6 \times 1} \quad X = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix}_{6 \times 4}$$

$$\beta = \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix}_{4 \times 1} \quad \epsilon = \begin{pmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{13} \\ \epsilon_{21} \\ \epsilon_{31} \\ \epsilon_{32} \end{pmatrix}_{6 \times 1}$$

## Motivating Example # 4

Example of a linear model with correlated error structure

**Example 4** – Split-plot design (See Chapter XI of Christensen).

A split-plot design is a generalization of a randomized complete block design.

As part of a program to standardize measurement of the blood level of phenytoin, an anti-epileptic drug, samples with known amounts of active ingredients are sent to 3 commercial laboratories (Factor A) for analysis. Each lab employs 4 technicians (Factor B) who conduct an assay. We wish to study the effect of these two factors on the blood level measurements.

Each replicate of a factorial experiment requires 12 assays, and the experimenter decided to run three replicates, doing 1 replicate per day. The days are thus considered as blocks.

On a given day, the experiment is conducted as follows. Drug samples are divided and sent to the 3 laboratories. Then the drug samples within a laboratory are divided between 4 technicians, who each conduct 1 assay per day. Thus, the measurements for a given lab and technician within a lab are correlated.

## Motivating Example # 4

Each block in the design is divided into three plots called whole plots, and the laboratories are called the whole plot treatments. Each whole plot is divided into four parts called subplots (or split-plots), and one technician is assigned to each. The technician is the subplot treatment.

The linear model for the split-plot design is

$$y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ij}^{(1)} + \gamma_k + (\alpha\gamma)_{ik} + \epsilon_{ijk}^{(2)},$$
$$i = 1, \dots, a, j = 1, \dots, b, \text{ and } k = 1, \dots, c.$$

$\alpha_i$  = *i*th whole plot treatment effect.

$\beta_j$  = *j*th block effect.

$\gamma_k$  = *k*th subplot treatment effect.

$(\alpha\gamma)_{ik}$  = whole plot treatment and subplot treatment interaction.

## Motivating Example # 4

For our example,  $a = 3$ ,  $b = 3$ ,  $c = 4$ . We have two error terms – one for each stage of the randomization. A split-plot experiment involves a two stage randomization. We usually assume

$$\epsilon_{ij}^{(1)} \sim N(0, \sigma_1^2) \quad i.i.d.$$

$$\epsilon_{ijk}^{(2)} \sim N(0, \sigma_2^2) \quad i.i.d.$$

$\epsilon_{ij}^{(1)}$  indep of  $\epsilon_{ijk}^{(2)}$ .

$$\epsilon_{ijk} = \epsilon_{ij}^{(1)} + \epsilon_{ijk}^{(2)}.$$

We can write the model as  $Y = X\beta + \epsilon$  ,  
where

$$\epsilon = \begin{pmatrix} \epsilon_{111} \\ \epsilon_{112} \\ \vdots \\ \epsilon_{abc} \end{pmatrix} = \begin{pmatrix} \epsilon_{11}^{(1)} + \epsilon_{111}^{(2)} \\ \epsilon_{11}^{(1)} + \epsilon_{112}^{(2)} \\ \vdots \\ \epsilon_{ab}^{(1)} + \epsilon_{abc}^{(2)} \end{pmatrix}_{abc \times 1}$$

## Motivating Example # 4

$$\text{Cov}(\epsilon) = \Sigma = \begin{pmatrix} R_1 & 0 & \dots & 0 \\ R_2 & \dots & \vdots & \\ \ddots & & 0 & \\ & & & R_{ab} \end{pmatrix}_{abc \times abc}.$$

The  $\text{Cov}(\epsilon)$  is a block diagonal matrix. We will have  $ab c \times c$  blocks with each

$$R_i = \begin{pmatrix} \sigma_1^2 + \sigma_2^2 & \sigma_1^2 & \dots & \sigma_1^2 \\ \sigma_1^2 + \sigma_2^2 & \dots & \sigma_1^2 & \\ \ddots & & \sigma_1^2 & \\ & & \sigma_1^2 + \sigma_2^2 & \end{pmatrix}_{c \times c}.$$

The  $\epsilon_{ijk}$ 's are not independent for all  $i, j, k$ . Thus,  $\text{Cov}(\epsilon_{ijk}, \epsilon_{i'j'k'}) \neq 0$ .

# Motivating Example # 4

Claim:

$$\text{Cov}(\epsilon_{ijk}, \epsilon_{i'j'k'}) = \begin{cases} 0 & \text{If } i \neq i' \text{ or } j \neq j' \\ \sigma_1^2 & \text{If } i = i' \text{ and } j = j' \text{ and } k \neq k' \\ \sigma_1^2 + \sigma_2^2 & \text{If } i = i', j = j', k = k' \end{cases}$$

Proof:

$$\begin{aligned} \text{Cov}(\epsilon_{ijk}, \epsilon_{i'j'k'}) &= \text{Cov}\left(\epsilon_{ij}^{(1)} + \epsilon_{ijk}^{(2)}, \epsilon_{i'j'}^{(1)} + \epsilon_{i'j'k'}^{(2)}\right) \\ &= \text{Cov}\left(\epsilon_{ij}^{(1)}, \epsilon_{i'j'}^{(1)}\right) + \text{Cov}\left(\epsilon_{ij}^{(1)}, \epsilon_{i'j'k'}^{(2)}\right) \\ &\quad + \text{Cov}\left(\epsilon_{ijk}^{(2)}, \epsilon_{i'j'}^{(1)}\right) + \text{Cov}\left(\epsilon_{ijk}^{(2)}, \epsilon_{i'j'k'}^{(2)}\right) \\ &= \text{Cov}\left(\epsilon_{ij}^{(1)}, \epsilon_{i'j'}^{(1)}\right) + \text{Cov}\left(\epsilon_{ijk}^{(2)}, \epsilon_{i'j'k'}^{(2)}\right) + 0 + 0 \end{aligned}$$

## Motivating Example # 4

If  $i = i'$ ,  $j = j'$ , and  $k \neq k'$ , then

$$\begin{aligned} &= \text{Cov}\left(\epsilon_{ij}^{(1)}, \epsilon_{ij}^{(1)}\right) + \text{Cov}\left(\epsilon_{ijk}^{(2)}, \epsilon_{ijk'}^{(2)}\right) \\ &= \text{Var}\left(\epsilon_{ij}^{(1)}\right) = \sigma_1^2. \end{aligned}$$

If  $i' = i$ ,  $j' = j$ , and  $k = k'$ ,

$$\text{cov}(\epsilon_{ijk}, \epsilon_{ijk}) = \text{var}(\epsilon_{ijk}) = \text{var}(\epsilon_{ij}^{(1)} + \epsilon_{ijk}^{(2)}) = \text{var}(\epsilon_{ij}^{(1)}) + \text{var}(\epsilon_{ijk}^{(2)}) = \sigma_1^2 + \sigma_2^2.$$

## Vector Spaces

Christensen, Appendix A.

### Other references

- Halmos
- Strang
- Schaum's outline series

The type of vector spaces we consider are finite dimensional real vector spaces.

### Defn:

A real vector space  $\mathcal{M}$  is a set of elements (called vectors) with the following properties:

# Vector Spaces

## A. Addition axioms:

For  $x, y \in \mathcal{M}$ , there corresponds a vector  $x + y \in \mathcal{M}$  called the sum of  $x$  and  $y$  such that

A1.  $x + y = y + x$  (commutative)

A2.  $x + (y + z) = (x + y) + z$  (associative)

A3. There exists a unique vector  $0$ , the null vector such that for all  $x \in \mathcal{M}$ ,  $x + 0 = x$

A4. for all  $x \in \mathcal{M}$ , there exists a unique element  $-x$  such that  $x + (-x) = 0$

Note:  $x$  and  $y$  need not be vectors of real numbers.

## B. Scalar multiplication axioms

For any real number  $\alpha$  and for any  $x \in \mathcal{M}$ , there exists a member of  $\mathcal{M}$ ,  $\alpha x$ , called product of  $\alpha$  and  $x$  such that

B1.  $\alpha(x + y) = \alpha x + \alpha y$  (distributive)

B2.  $(\alpha + \beta)x = \alpha x + \beta x$  (distributive)

B3.  $\alpha(\beta x) = (\alpha\beta)x$  (associative)

B4.  $1 \times x = x$ , where  $1$  is a unique scalar.

# Vector Spaces

## Examples of Vector Spaces

Example 1  $\mathcal{M} = (n \text{ dimensional Euclidean space}), R^n.$

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \quad y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

$$x + y = \begin{pmatrix} x_1 + y_1 \\ \vdots \\ x_n + y_n \end{pmatrix} \quad \alpha x = \begin{pmatrix} \alpha x_1 \\ \vdots \\ \alpha x_n \end{pmatrix} .$$

Note: Christensen defines a vector space as a space within  $R^n$ . We have given a more general definition.  $R^n$  is an example of a vector space. Thus, a vector space need not be defined via  $R^n$ .

# Vector Spaces

Example 2 – The space of all  $2 \times 2$  matrices with real elements is a vector space.  $\mathcal{M}$  = space of a  $2 \times 2$  matrices.  
 $x$  is a  $2 \times 2$  matrix.

$$0 = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$$

$$x = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \quad y = \begin{pmatrix} 5 & 6 \\ 7 & 8 \end{pmatrix}$$

$$x + y = \begin{pmatrix} 6 & 8 \\ 10 & 12 \end{pmatrix}$$

and so on.

Example 3 – The polynomials of degree  $n$  with real coefficients constitutes a vector space.  $\mathcal{M}$  = polynomials of degree  $n$ .

$$x = \beta_0 + \beta_1 t + \dots + \beta_n t^n.$$

$$y = \gamma_0 + \gamma_1 t + \dots + \gamma_n t^n.$$

# Vector Spaces

Defn: A set of vectors  $D = \{x_1, \dots, x_r\}$  is called linearly dependent if there is a set of scalars  $\alpha_1, \dots, \alpha_r$ , not all zero, such that

$$\sum_{i=1}^r \alpha_i x_i = 0.$$

If  $\sum_{i=1}^r \alpha_i x_i = 0 \Rightarrow \alpha_i = 0, i = 1, \dots, r$ , then  $D = \{x_1, \dots, x_r\}$  are linearly independent.

Note: If  $0 \in D$ ,  $D$  is linearly dependent. If  $\phi = D$ ,  $D$  is linearly independent.

If  $D$  is linearly independent, then  $D_1 \subset D$  is linearly independent (proof by contradiction).

Example 1:  $M = R^3$ .

$$x_1 = \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix} \quad x_2 = \begin{pmatrix} 16 \\ 12 \\ 3 \end{pmatrix} \quad x_3 = \begin{pmatrix} 0 \\ 28 \\ 3 \end{pmatrix}$$

$D = \{x_1, x_2, x_3\}$  is linearly dependent since  $16x_1 - x_2 + x_3 = 0$ .

# Vector Spaces

## Example 2:

Consider the linear model  $y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$ ,  $\beta_i \neq 0$ ,  $i = 1, 2, 3$ , and the  $x_i$ 's are given by Example 1.

What is the meaning of linear dependence for this linear model? Since, for the example given above  $x_2 - x_3 = 16x_1$ , or  $x_1 = (x_2 - x_3)/16$ , by substituting for  $x_1$ , we write:

$$\begin{aligned}y &= \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon \\&= x_2(\beta_2 + \beta_1/16) + x_3(\beta_3 - \beta_1/16) + \epsilon \\&= \gamma_1 x_2 + \gamma_2 x_3 + \epsilon\end{aligned}$$

Where the  $\gamma_i$ 's are defined by the above equation. This shows that the model with three parameters is equivalent to a model with only two parameters.

## Example 3: $\mathcal{M} = \mathbb{R}^n$

$$e_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \quad e_2 = \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix} \quad \dots \quad e_n = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}$$

$D = \{e_1, \dots, e_n\}$  is linearly independent.

# Vector Spaces

Suppose  $D = \{x_1, \dots, x_r\}$  is a linearly dependent. This means that

$$\sum_{i=1}^r \alpha_i x_i = 0, \text{ and not all } \alpha_i = 0.$$

Let  $k$  be such that  $\alpha_k \neq 0$ . Then

$$x_k = \sum_{i=1}^r -\frac{\alpha_i}{\alpha_k} x_i , \\ i \neq k.$$

We are led to the following theorem.

**Theorem** A set of vectors is linearly dependent if and only if some vector of the set can be written as a linear combination of the others. That is there exists a  $k$  such that

$$x_k = \sum_{i=1}^r -\frac{\alpha_i}{\alpha_k} x_i ,$$

$$\alpha_k \neq 0, i \neq k.$$

**Proof:** Exercise.

# Vector Spaces

## Defn:

A basis in a vector space  $\mathcal{M}$  is a set of linearly independent vectors such that every  $x \in \mathcal{M}$  is a linear combination of vectors in the set.

Example:  $\mathcal{M} = \mathbb{R}^3$

The vectors

$$\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \quad \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \quad \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

are a basis for  $\mathbb{R}^3$ .

Every vector in  $\mathbb{R}^3$  can be written as a linear combination of the vectors above.  
Bases are not unique. The vectors

$$\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \quad \begin{pmatrix} 3 \\ 3 \\ 0 \end{pmatrix} \quad \begin{pmatrix} 2 \\ 2 \\ 2 \end{pmatrix}$$

are also a basis for  $\mathbb{R}^3$ .

# Vector Spaces

**Note:** The representation of a vector as a linear combination of basis elements is unique. The coefficients of the vectors are often called coordinates with respect to that basis.

To illustrate, suppose  $B = \{x_1, \dots, x_r\}$  is a basis and  $z = \sum_{i=1}^r \alpha_i x_i$ .

Then,  $z = \sum_{i=1}^r \beta_i x_i \Rightarrow \sum_{i=1}^r (\alpha_i - \beta_i) x_i = 0 \Rightarrow \alpha_i = \beta_i, i = 1, \dots, n$ .

In this example  $(\alpha_1, \dots, \alpha_r)$  are the coordinates of  $z$  with respect to  $B$ .

**Defn:** A vector space  $\mathcal{M}$  is said to be finite dimensional if it has a basis with a finite number of elements.

## Examples

$\mathbb{R}^n$  is finite dimensional

$P^n$  is finite dimensional (polynomials of degree  $n$ ).

Many useful results in linear algebra do not involve coordinates. Many of these carry over to properties of linear models. So picking a basis will be unnecessary. Not selecting a basis is called coordinate free. In this course, we will not go the coordinate free route. We will select coordinates. No generality is lost.

# Vector Spaces

Let  $D$  be a set of vectors.

- ① The set of all possible linear combinations of elements of  $D$  is a vector space, called the span of  $D$  written  $\mathcal{S}(D)$ .
- ②  $\mathcal{S}(D) = \cap_i \mathcal{M}_i$  of all vector spaces containing  $D$ .
- ③ A basis for  $\mathcal{M}$  is a linearly independent set of elements of  $\mathcal{M}$  whose span is  $\mathcal{M}$ .

## Theorem

Every basis of a vector space  $\mathcal{M}$  contains the same number of elements.

This number is called the dimension of  $\mathcal{M}$ , written  $\dim(\mathcal{M})$ . This number is also called the rank of  $\mathcal{M}$ , written  $r(\mathcal{M})$ .

Example:  $\mathcal{M} = \mathbb{R}^n$ . We have the following facts:

- ①  $\dim(\mathbb{R}^n) = n$  ,  $r(\mathbb{R}^n) = n$
- ②  $n + 1$  vectors in an  $n$ -dimensional space must be linearly dependent.
- ③  $n$  vectors in an  $n$ -dimensional space form a basis for  $\mathbb{R}^n$  if and only if they are linearly independent.
- ④  $n$  vectors in an  $n$ -dimensional space form a basis for  $\mathbb{R}^n$  if and only if they span the space.
- ⑤  $D = \{x_1, \dots, x_r\}$ ,  $r < n$ , cannot be a basis for  $\mathbb{R}^n$

# Vector Spaces

**Theorem** If  $\{x_1, \dots, x_r\}$  is a linearly independent set of vectors in  $\mathcal{M}$ , and  $\dim(\mathcal{M}) = n$ ,  $r < n$ , then there exists elements  $x_{r+1}, \dots, x_n$  such that  $\{x_1, \dots, x_n\}$  is a basis for  $\mathcal{M}$ .

This theorem says that any set of linearly independent set of vectors can be extended to a basis.

# Vector Spaces

**Linear Subspaces** (Manifolds) Defn: Let  $\mathcal{M}$  be a vector space and let  $N$  be a set with  $N \subset \mathcal{M}$ . Then  $N$  is a subspace of  $\mathcal{M}$  if and only if  $N$  is a vector space.

## Theorem

Let  $\mathcal{M}$  be a vector space and let  $N$  be a nonempty subset of  $\mathcal{M}$ . If  $N$  is closed under addition and scalar multiplication, then  $N$  is a subspace of  $\mathcal{M}$ .

**Closed under addition and multiplication implies the other vector space axioms**

# Vector Spaces

## Examples of Subspaces

- a) Let  $\mathcal{M} = \mathbb{R}^3$ . Choose a vector  $x_0 \in \mathbb{R}^3$ , where  $x_0 \neq 0$ . Consider all vectors of the form  $\alpha x_0$ ,  $\alpha \in \mathbb{R}^1$ .

$$\mathcal{S}(x_0) = \{\alpha x_0 : \alpha \in \mathbb{R}^1\} .$$

$\mathcal{S}(x_0)$  is a subspace of  $\mathbb{R}^3$ .

- b) Choose vectors  $x_0$  and  $x_1$  which are linearly independent. The set

$$\{\alpha x_0 + \beta x_1 : \alpha, \beta \in \mathbb{R}^1\}$$

is a subspace of  $\mathbb{R}^3$ . The set above equals  $\mathcal{S}(x_0, x_1)$ .

# Vector Spaces

## Example

Let  $D$  be any set of vectors in a vector space  $\mathcal{M}$ . Then  $\mathcal{S}(D)$  is a linear subspace in  $\mathcal{M}$ .

Sometimes it is useful to consider linear subspaces translated from the origin. Let  $\mathcal{M}$  be a vector space and let  $N$  be a subspace of  $\mathcal{M}$ . Let  $y_0$  be an element of  $\mathcal{M}$ .

A flat or coset consists of

All vector spaces contain the origin. Sometimes we want to translate these spaces out of the origin. They won't be subspaces anymore. A flat is not necessarily a subspace.

$$\{x + y_0 : x \in N\} .$$

We write  $y_0 + N$ , where  $N$  is a subspace to indicate a flat. We note here that  $y_0$  is not unique. It can be any element of the flat. If  $y_0 \in N$ , then  $N + y_0 = N$ .

Intercept term acts as  $y_0$  in linear models

# Vector Spaces

## Example

Let  $\mathcal{M} = \mathbb{R}^2$ , and consider

$$N = \left\{ \alpha \begin{pmatrix} 1 \\ 2 \end{pmatrix}; \alpha \in \mathbb{R}^1 \right\}, y_0 = \begin{pmatrix} 2 \\ 2 \end{pmatrix}.$$

The flat  $y_0 + N$  is given by the set

$$y_0 + N = \left\{ \begin{pmatrix} 2 \\ 2 \end{pmatrix} + \alpha \begin{pmatrix} 1 \\ 2 \end{pmatrix}; \alpha \in \mathbb{R}^1 \right\},$$

which is a straight line which passes through  $(1, 0)$ .

Clearly,  $y_0$  is not unique. It can be any point of the form  $y = y_0 + y_\alpha$ , where

$$y_\alpha = \alpha \begin{pmatrix} 1 \\ 2 \end{pmatrix}.$$

For any  $y_0$  not of this form, we simply get a different flat. As mentioned earlier, the choice of  $y_0$  is not unique. It can be any element of the flat. Flats are not subspaces in general.

# Vector Spaces

## Example

Consider the linear model

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \beta + \epsilon,$$

where  $\beta$  is a scalar, and  $E(\epsilon) = 0$ . This is a special case of the linear model

$Y = X\beta + \epsilon$ . Clearly  $Y \in R^2$ , but  $\mu = E(Y) = \beta \begin{pmatrix} 1 \\ 1 \end{pmatrix}$  is in a one-dimensional subspace of  $R^2$ .

If we want to estimate  $\mu$ , we should take into account that

$$\mu = S \left\{ \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right\} = \left\{ \beta \begin{pmatrix} 1 \\ 1 \end{pmatrix} : \beta \in R^1 \right\}.$$

Now suppose the model was given by

$$Y = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \epsilon.$$

# Vector Spaces

This model now includes an intercept, and is just the simple linear regression model. We have

$$\mu = E(Y) = \begin{pmatrix} \beta_0 + \beta_1 x_1 \\ \beta_0 + \beta_1 x_2 \\ \beta_0 + \beta_1 x_3 \end{pmatrix} = \begin{pmatrix} \beta_0 \\ \beta_0 \\ \beta_0 \end{pmatrix} + \beta_1 \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}.$$

We recognize this as a flat with  $y_0 = (\beta_0, \beta_0, \beta_0)'$  and

$$N = \mathcal{S} \left\{ \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \right\}.$$

Let  $H$  and  $K$  be two linear subspaces. Define the sum  $H + K$  as

$$H + K = \{x + y : x \in H, y \in K\} .$$

Moreover, define  $H \cap K = \{x : x \in H, x \in K\}$ . We are led to the following theorem.

### Theorem

Both  $H + K$  and  $H \cap K$  are linear subspaces.

proof: Homework

# Vector Spaces

## Defn:

Two subspaces are disjoint if  $H \cap K = 0$ , where 0 is the null vector.

## Theorem

If  $H \cap K = 0$  and  $z \in H + K$ , then the decomposition  $z = x + y$  with  $x \in H$  and  $y \in K$  is unique.

## Proof

Suppose  $z = x + y$  and  $z = x' + y'$ . Then  $x - x' \in H$  and  $y - y' \in K$ . Therefore, we must have  $x + y = x' + y'$ , which implies  $x - x' = y' - y$ . This in turn requires that

$x - x' = y' - y = 0$  since 0 is the only vector common to  $H$  and  $K$ . Thus  $x = x'$  and  $y = y'$ , and this completes the proof.

# Vector Spaces

## Theorem

If  $H \cap K = 0$ , then  $r(H + K) = r(H) + r(K)$ . In general, we have

$$r(H + K) = r(H) + r(K) - r(H \cap K).$$

## Defn:

If  $N$  and  $N^c$  are disjoint subspaces of  $\mathcal{M}$  and  $\mathcal{M} = N + N^c$ , then  $N^c$  is called the complement of  $N$ .

## Remark:

The complement is not unique. In  $R^2$ , a subspace  $N$  of dimension 1 consists of a line through the origin. A complement of  $N$  is given by any other line  $N^c \neq \alpha N$  through the origin, because any two such lines span  $R^2$ .

e.g. let  $N=S(1,1)$ ,  $Nc=S(1,2)$ . They span  $R^2$ . Now let  $Nc=S(1,3)$ . In the case above,  $N$  is the span of the line, not the line itself

# Vector Spaces

Defn:

We can define orthogonality in other spaces, such as function spaces, but we're not interested in that

Suppose  $\mathcal{M}$  is a vector space in  $R^n$ . Let  $x$  and  $y$  be two vectors in  $\mathcal{M}$ . Then  $x$  and  $y$  are said to be orthogonal, written  $x \perp y$ , if  $x'y = 0$ , where  $x'$  denotes the transpose of  $x$ .

Two subspaces  $N_1$  and  $N_2$  are said to be orthogonal if for every  $x \in N_1$  and  $y \in N_2$  implies  $x'y = 0$ .

Defn:

Suppose  $N$  is a subspace of  $R^n$ . Then  $\{x_1, \dots, x_r\}$  is an orthogonal basis for  $N$  if for every  $i \neq j$ ,  $x_i'x_j = 0$ .  $\{x_1, \dots, x_r\}$  is an orthonormal basis if in addition,  $x_i'x_i = 1$ , for  $i = 1, \dots, r$ .

Note: Two orthogonal vectors are necessarily linearly independent.

# Vector Spaces

## Theorem (Gram-Schmidt)

Let  $N$  be a subspace of  $R^n$  with basis  $\{x_1, \dots, x_r\}$ . Then there exists an orthonormal basis for  $N$ ,  $\{y_1, \dots, y_r\}$  with  $y_s \in \mathcal{S}(x_1, \dots, x_s)$ ,  $s = 1, \dots, r$ .

Explicitly, the  $y_s$ 's are given by

$$y_1 = (x_1' x_1)^{-1/2} x_1$$

$$w_s = x_s - \sum_{i=1}^{s-1} (x_s' y_i) y_i, \quad s = 2, \dots, r$$

$$y_s = (w_s' w_s)^{-1/2} w_s, \quad s = 2, \dots, r$$

# Vector Spaces

Defn: (Orthogonal Complement)

Let  $N$  be a subspace of a vector space  $\mathcal{M} \subset R^n$ . Define

$$N^\perp = \{y \in \mathcal{M} : y \perp N\}.$$

$N^\perp$  is called the orthogonal complement of  $N$  with respect to  $\mathcal{M}$ .

If  $\mathcal{M} = R^n$ , then  $N^\perp$  is referred to the orthogonal complement of  $N$ .

## Theorem

Let  $\mathcal{M}$  be a vector space and let  $N^\perp$  be the orthogonal complement of  $N$  with respect to  $\mathcal{M}$ . Then  $N^\perp$  is a subspace of  $\mathcal{M}$ , and if  $x \in \mathcal{M}$ ,  $x$  can be written uniquely as  $x = x_0 + x_1$ , with  $x_0 \in N$  and  $x_1 \in N^\perp$ . The ranks of these subspaces satisfy

$$r(\mathcal{M}) = r(N) + r(N^\perp).$$

Also

Note that this is addition, NOT union! Also, that  $N$ -perp is a subspace is also important

$$\mathcal{M} = N + N^\perp = \left\{x : x = x_0 + x_1, x_0 \in N, x_1 \in N^\perp\right\}.$$

# Vector Spaces

## Matrices

We're now in the  $R^n$  world. No more general vector spaces.

Suppose  $A$  is an  $n \times p$  matrix.  $A'$  will denote the transpose of  $A$ .

A matrix can be defined as a linear transformation on vector space.

Defn:

Suppose  $\mathcal{M}$  is an arbitrary vector space. A linear transformation  $A$  on a vector space  $\mathcal{M}$  is a function mapping  $\mathcal{M} \rightarrow \mathcal{M}$  such that

$$A(\alpha x + \beta y) = \alpha A(x) + \beta A(y)$$

For all  $\alpha, \beta \in R^1$  and  $x, y \in \mathcal{M}$ .

Suppose  $A$  is an  $n \times p$  matrix. Then each column of  $A$  is a vector in  $R^n$ . We can write

$$A = (x_1, \dots, x_p)$$

where each  $x_i \in R^n$ ,  $i = 1, \dots, p$ .

The space spanned by the columns of  $A$  is called the column space of  $A$ , written  $C(A)$ . That is  $S(A) = C(A)$ . Also  $r(A)$  will denote the rank of  $A$ .

# Vector Spaces

Example  
Suppose

$$A = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{pmatrix},$$

Then

$$\begin{aligned} C(A) &= \mathcal{S} \left\{ \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 2 \end{pmatrix} \right\} \\ &= \left\{ \alpha \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} + \beta \begin{pmatrix} 0 \\ 1 \\ 2 \end{pmatrix} : \alpha, \beta \in R^1 \right\} \end{aligned}$$

Here  $r(A) = 2$ , since the two vectors are linearly independent.

# Vector Spaces

## Theorem

$$r(A) = r(A').$$

**NOTE: If its not specified, A is not square**

## Defn:

Suppose  $A$  is an  $n \times n$  square matrix with  $ij$ th element  $a_{ij}$ . The trace of  $A$  is defined as

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

## Theorem

$$tr(A + B) = tr(A) + tr(B) .$$

Properties of the trace operator: The trace is invariant under cyclic permutations. Suppose  $A, B, C$  are  $n \times n$  square matrices. Then

$$tr(ABC) = tr(BCA) = tr(CAB) .$$

# Vector Spaces

## Defn:

Suppose  $A$  is an  $n \times n$  square matrix. Then  $A$  is said to be nonsingular if there exists a matrix  $A^{-1}$  such that  $A^{-1}A = AA^{-1} = I$ . If no such matrix exists, then  $A$  is said to be singular.

If  $B$  is nonsingular, then  $\text{tr}(A) = \text{tr}(BAB^{-1}) = \text{tr}(B^{-1}AB)$ .

## Theorem

An  $n \times n$  matrix  $A$  is nonsingular if and only if  $r(A) = n$ , i.e.,  $C(A)$  is a basis for  $R^n$ . Thus  $A$  is nonsingular if and only if all of its columns are linearly independent.

$A$  is singular if and only if there exists a nonzero vector  $x$  such that  $Ax = 0$ ,  $x \in R^n$ .

## Defn:

The set of all  $x$  such that  $Ax = 0$  is a vector space and is called the null space of A, written  $\mathcal{N}(A)$ .

Null space can have different dimensions than  
**A**

## Theorem

Suppose  $A$  is  $n \times n$ . If  $r(A) = r$ , then  $r(\mathcal{N}(A)) = n - r$ .

# Vector Spaces

## Eigenvalues and Eigenvectors

Suppose  $A$  is an  $n \times n$  square matrix. An eigenvector of  $A$  is any nonzero vector  $x$  satisfying

$$Ax = \lambda x, \quad \lambda \in R^1.$$

$\lambda$  is called an eigenvalue of  $A$ .

Eigenvectors are not unique. To see this, note that

$A(cx) = cAx = c\lambda x = \lambda(cx)$ , so that  $cx$  is an eigenvector of  $\lambda$ .

### Theorem

If  $x_1$  and  $x_2$  are eigenvectors with the same eigenvalue, then any nonzero linear combination of  $x_1$  and  $x_2$  is also an eigenvector with the same eigenvalue.

# Vector Spaces

## Defn:

A square matrix  $A$  is said to symmetric if  $A = A'$ .

## Theorem

If  $A$  is a symmetric matrix, then the eigenvalues of  $A$  are real.

## Theorem

$\lambda$  is an eigenvalue of  $A$  if and only if  $A - \lambda I$  is singular.

## Some facts about eigenvalues and eigenvectors

- 1) If  $\lambda$  is an eigenvalue of  $A$  and  $\lambda \neq 0$ , then the eigenvectors corresponding to  $\lambda$  form a subspace of  $C(A)$ .
- 2) If  $A$  is symmetric, and  $\lambda$  and  $\gamma$  are distinct eigenvalues, then the eigenvectors corresponding to  $\lambda$  and  $\gamma$  are orthogonal. These eigenvectors forms a basis for a subspace of  $C(A)$ .
- 3)  $A$  is nonsingular if and only if all of its eigenvalues are nonzero.
- 4) Theorem

If  $A$  is a symmetric matrix, then there exists a basis for  $C(A)$  consisting of eigenvectors of nonzero eigenvalues. The eigenvectors corresponding to the nonzero eigenvalues of  $A$  are a basis for  $C(A)$ .



# Vector Spaces

- 5) If  $A$  is  $n \times n$  and symmetric, and all of its eigenvalues are nonzero, then  $C(A) = R^n$ , and the eigenvectors of  $A$  are a basis for  $R^n$ .
- 6) If  $A$  is  $n \times n$  and symmetric, and some of its eigenvalues are 0, then the eigenvectors corresponding to the nonzero eigenvalues are a basis for  $C(A) \subset R^n$ . Thus,  $r(A) = \text{number of nonzero eigenvalues of } A$ .
- 7) If  $A$  is  $n \times n$  and symmetric, then the eigenvectors corresponding to the 0 eigenvalues (if any) are a basis for  $\mathcal{N}(A)$ .
- 8) If  $A$  is symmetric, then  $\mathcal{N}(A) = C(A)^\perp$ . That is, the null space of  $A$  corresponds to the orthogonal complement of  $A$ .
- 9) Theorem

Suppose  $A$  is  $n \times n$  and symmetric. Then there exists eigenvectors of  $A$  that are an orthogonal basis for  $C(A)$ . If  $A$  is nonsingular, then they are an orthogonal basis for  $R^n$ . If we normalize these eigenvectors, they are an orthonormal basis.

# Vector Spaces

## Theorem

Suppose  $A$  is  $n \times n$  symmetric of rank  $r \leq n$ . Then

- i)  $\mathcal{N}(A) = C(A)^\perp$ .
- ii)  $C(A) \cap \mathcal{N}(A) = 0$ .
- iii)  $C(A) + \mathcal{N}(A) = R^n$ .
- iv)  $r(A) = r$ ,  $r(\mathcal{N}(A)) = n - r$ .

# Vector Spaces

The eigenvalues of a matrix  $A$  are found by finding the zeroes of the equation

$$\det(A - \lambda I) = 0$$

where  $\det(A)$  denotes the determinant of  $A$ .

Suppose  $A$  is  $n \times n$  with eigenvalues  $\lambda_1, \dots, \lambda_n$ . Then

- 1)  $\det(A) = \prod_{i=1}^n \lambda_i$ .
- 2)  $A$  is singular if and only if  $\det(A) = 0$ .
- 3) If  $A$  is nonsingular, (i.e., all eigenvalues nonzero), then  $A^{-1}$  exists and the eigenvalues are given by  $\lambda_1^{-1}, \dots, \lambda_n^{-1}$ .
- 4) The eigenvalues of  $A'$  are the same as those of  $A$ .
- 5)  $\text{tr}(A) = \sum_{i=1}^n \lambda_i$  and  $\text{tr}(A^{-1}) = \sum_{i=1}^n \lambda_i^{-1}$ .
- 6) If  $A$  is symmetric, then  $\text{tr}(A^r) = \sum_{i=1}^n \lambda_i^r$  for any integer  $r$ .

# Vector Spaces

## Defn:

A square matrix  $P$  is said to be orthogonal if  $P' = P^{-1}$ . If  $P$  is an orthogonal matrix, so is  $P'$ . Thus a square matrix is orthogonal if  $PP' = P'P = I$ .

Columns are orthonormal

## Theorem

The product of two orthogonal matrices is orthogonal.

## Proof:

If  $P_1$  and  $P_2$  are orthogonal matrices, then

$$(P_1 P_2)(P_1 P_2)' = P_1 P_2 P_2' P_1' = P_1 I P_1' = P_1 P_1' = I .$$

## Theorem

An  $n \times n$  matrix  $P$  is orthogonal if and only if the columns of  $P$  form an orthonormal basis for  $R^n$ .

## Defn:

Suppose  $A$  is an  $n \times p$  matrix. Then

$$C(A) = \{z : Ax = z, x \in R^p\}$$

# Vector Spaces

## Theorem

Suppose  $A$  is an  $n \times p$  matrix. Then

$$\mathcal{N}(A) = C(A')^\perp.$$

## Proof

1)  $\Rightarrow$

We first show that  $\mathcal{N}(A) \subset C(A')^\perp$ . Let  $x \in \mathcal{N}(A)$ . Then  $Ax = 0$ .  $Ax = 0 \Rightarrow x'A' = 0'$ . Now write  $A' = (a_1, \dots, a_n)$ , where  $a_j$  is  $p \times 1$  and is the  $j$ th column of  $A'$ . Also write  $0' = (0, \dots, 0)$ , so that  $x'A' = (x'a_1, \dots, x'a_n) = (0, \dots, 0)$ . This implies that  $x'a_j = 0$  for each  $j = 1, \dots, n$ . Now if  $z \in C(A')$ , then we can write  $z = \alpha_1 a_1 + \dots + \alpha_n a_n$ . Therefore,

$x'z = x'(\alpha_1 a_1 + \dots + \alpha_n a_n) = \alpha_1 x'a_1 + \dots + \alpha_n x'a_n = 0 + \dots + 0 = 0$ . Thus  $x \in C(A')^\perp$ .

2)  $\Leftarrow$

Now we must show  $C(A')^\perp \subset \mathcal{N}(A)$ .

Let  $x \in C(A')^\perp$ . Then for any vector  $z \in C(A')$ ,  $x'z = 0$ . Now  $C(A') = \mathcal{S}(a_1, \dots, a_n) = \{z : z = \alpha_1 a_1 + \dots + \alpha_n a_n\}$ . Therefore,  $x'z = 0 \Rightarrow \alpha_1 x'a_1 + \dots + \alpha_n x'a_n = 0$  for all  $(\alpha_1, \dots, \alpha_n)$ , and  $z \in C(A')$ . This implies that  $x'a_j = 0$  for all  $j = 1, \dots, n$ ,  $\Rightarrow x'(a_1, \dots, a_n) = (0, \dots, 0)$ ,  $\Rightarrow Ax = 0$ , and thus  $x \in \mathcal{N}(A)$ . Thus  $C(A')^\perp \subset \mathcal{N}(A)$ . This completes the proof.

# Vector Spaces

## Miscellaneous Results

- 1) Suppose  $A$  is an  $n \times p$  matrix of rank  $r$ . Then  $r \leq \min(n, p)$ .
- 2) Suppose  $A$  is an  $n \times p$  matrix of rank  $r$  and  $B$  is a  $p \times s$  matrix where  $r(AB) = r$ . Then  $C(A) = C(AB)$ . If  $r(AB) < r$ , then  $C(AB) \subset C(A)$ .
- 3) If  $B$  is a square nonsingular matrix, then  $C(A) = C(AB)$ .
- 4) In general,  $C(AB) \subset C(A)$  and  $\mathcal{N}(B) \subset \mathcal{N}(AB)$ .

In general, roughly speaking, #2 is equivalent to saying that if  $r(A)=r(B)=r$  then  $C(A)=C(AB)$  (which means that  $r(AB)=r$  as well)  
rank-preserving means post-multiplication preserves the column space

# Matrix Decompositions

## Matrix Decompositions

Dealing with matrices is generally made easier by decomposing the matrix into a product of matrices, each of which is relatively easy to work with, and has some special structure of interest.

### Spectral Decomposition

The spectral decomposition allows the representation of any symmetric matrix in terms of an orthogonal matrix and a diagonal matrix of eigenvalues.

### Theorem (Spectral Theorem)

Suppose  $A$  is an  $n \times n$  symmetric matrix. Then there exists an orthogonal matrix  $P$  such that

$$A = P\Lambda P'$$
,

where  $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_n)$  is an  $n \times n$  diagonal matrix of the eigenvalues of  $A$  with  $\lambda_1 \leq \lambda_2 \dots \leq \lambda_n$ . Here,  $P$  is the orthogonal matrix of eigenvectors corresponding to the eigenvalues of  $A$ .

### Theorem

The rank of a symmetric matrix is the number of nonzero eigenvalues.

# Matrix Decompositions

Defn:

A symmetric  $n \times n$  matrix  $A$  is positive definite if  $x'Ax > 0$  for all  $x \neq 0, x \in R^n$ .

Defn:

A symmetric  $n \times n$  matrix  $A$  is positive semidefinite if  $x'Ax \geq 0$  for all  $x \neq 0, x \in R^n$ .

Theorem

The eigenvalues of a positive definite matrix are all positive, and the eigenvalues of a positive semidefinite matrix are all nonnegative.

You can't go the other directions: positive eigenvalues does NOT imply positive definite alone

# Matrix Decomposition

This is the cholesky decomposition

## Theorem

$A$  is positive semidefinite if and only if there exists a matrix  $Q$  such that  $r(A) = r(Q)$  and  $A = QQ'$ .

## Corollary

$A$  is positive definite if and only if there exists a nonsingular matrix  $Q$  such that  $A = QQ'$ .

We construct  $Q$  as follows:

We know that  $A = P\Lambda P'$ , and define  $\Lambda^{1/2} = \text{diag}(\lambda_1^{1/2}, \dots, \lambda_n^{1/2})$ . Therefore,  $A = P\Lambda^{1/2}\Lambda^{1/2}P' = (P\Lambda^{1/2})(P\Lambda^{1/2})' = QQ'$ . Clearly  $Q$  is nonsingular since  $P$  is orthogonal,  $|P| = \pm 1$  and  $\Lambda$  is nonsingular.

Note: Covariance matrices are positive semidefinite.

# Matrix Decompositions

Note,  $C(X) \cup C(X)^{\perp} \neq \mathbb{R}^n$ , but  $C(X) + C(X)^{\perp} = \mathbb{R}^n$

## Theorem

If  $A$  is positive definite and  $A = P\Lambda P'$ , then

$$A^{-1} = P\Lambda^{-1}P'$$

where  $\Lambda^{-1} = \text{diag}(\lambda_1^{-1}, \dots, \lambda_n^{-1})$ .

## Theorem

Suppose  $A$  is  $m \times n$ . Then  $A'A$  and  $AA'$  are positive semidefinite.

# Matrix Decompositions

## Theorem

Suppose  $A$  is an  $n \times n$  symmetric matrix. Let  $a_{ii}$  denote the  $i$ th diagonal element of  $A$ . Then  $A$  is positive definite if and only if

- 1)  $a_{ii} > 0$ , for all  $i = 1, \dots, n$ .
- 2) The determinant of every submatrix is positive. That is

$$\det \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} > 0, \det \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} > 0, \dots,$$

$$\det(A) > 0.$$

- 3)  $A$  is positive semidefinite if we replace  $> 0$  in 1) and 2) above by  $\geq 0$ .

# Matrix Decompositions

## Example

$$A = \begin{bmatrix} 2 & -1 & 1 & 1 \\ -1 & 4 & 0 & 2 \\ 1 & 0 & 4 & 3 \\ 1 & 2 & 3 & 5 \end{bmatrix}$$

$A$  is positive definite since

1)  $a_{ii} > 0$ ,  $i = 1, \dots, 4$ .

2)

$$\det \begin{pmatrix} 2 & -1 \\ -1 & 4 \end{pmatrix} = 8 - 1 = 7 > 0 ,$$

$$\det \begin{pmatrix} 2 & -1 & 1 \\ -1 & 4 & 0 \\ 1 & 0 & 4 \end{pmatrix} = 24, \quad \det(A) = 33.$$

# Matrix Decompositions

## Theorem

If  $A$  is an  $n \times n$  positive semidefinite matrix with non-zero eigenvalues  $\lambda_1, \dots, \lambda_r$ , then there exists an  $n \times r$  matrix  $Q = Q_1 Q_2^{-1}$  such that  $Q_1$  is an  $n \times r$  matrix with orthonormal columns and  $C(A) = C(Q_1)$  and  $Q_2$  is a diagonal matrix with positive diagonal elements, and  $Q' A Q = I$ .

## Proof

Not important. Can forget this. Next slide IS important.

Let  $Q_1 = (v_1, \dots, v_r)$  where  $v_i$  is an eigenvector of  $\lambda_i$ . Then the columns of  $Q$  are an orthonormal basis for  $C(A)$ . Thus  $A = Q_1 \Lambda Q_1'$ , where  $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_r)$ . Therefore  $\Lambda = Q_1' A Q_1$ .

Now take  $Q_2 = \Lambda^{1/2}$ ,  $\Lambda^{1/2} = \text{diag}(\lambda_1^{1/2}, \dots, \lambda_r^{1/2})$  and  $Q = Q_1 Q_2^{-1}$ . Clearly,  $C(A) = C(Q_1)$ , and  $Q_2$  is diagonal with positive elements. Moreover,

$$\begin{aligned} Q' A Q &= (Q_1 Q_2^{-1})' A (Q_1 Q_2^{-1}) \\ &= (Q_1 \Lambda^{-1/2})' (Q_1 \Lambda Q_1') (Q_1 \Lambda^{-1/2}) \\ &= \Lambda^{-1/2} Q_1' Q_1 \Lambda Q_1' Q_1 \Lambda^{-1/2} \\ &= \Lambda^{-1/2} I \Lambda \Lambda^{-1/2} \\ &= \Lambda^{1/2} \Lambda^{-1/2} \\ &= I \end{aligned}$$

# Matrix Decompositions

## Singular Value Decomposition (SVD)

### Theorem

Suppose  $A$  is an  $n \times p$  matrix of rank  $r$ , ( $r \leq \min(n, p)$ ). There exists orthogonal matrices  $U_{p \times p}$  and  $V_{n \times n}$  such that

$$V'AU = \begin{pmatrix} \Delta & 0 \\ 0 & 0 \end{pmatrix}$$

where  $\Delta = \text{diag}(\delta_1, \dots, \delta_r)$  is an  $r \times r$  diagonal matrix with  $\delta_1 \geq \delta_2 \dots \geq \delta_r > 0$ . The  $\delta_i$ 's are called the singular values of  $A$ .

# Matrix Decompositions

## Implications of SVD

1)  $A = VDU'$ , where  $D = \begin{pmatrix} \Delta & 0 \\ 0 & 0 \end{pmatrix}$ .

2) Split  $V = (V_1, V_2)$  and  $U = (U_1, U_2)$ , where  $V_1$  is  $n \times r$ ,  $V_2$  is  $n \times (n - r)$ ,  $U_1$  is  $p \times r$  and  $U_2$  is  $p \times (p - r)$ . Then

$$\begin{aligned} A = VDU' &= (V_1, V_2) \begin{pmatrix} \Delta & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} U'_1 \\ U'_2 \end{pmatrix} \\ &= V_1 \Delta U'_1 + V_2 0 U'_2 \\ &= V_1 \Delta U'_1 \end{aligned}$$

This implies that

$$\begin{aligned} C(A) &= \{z : z = Ax = V_1 \Delta U'_1 x, x \in R^p\} \\ &= \{z : z = V_1 x^*, x^* \in R^r\} \\ &= C(V_1) \end{aligned}$$

Note that since  $\Delta$  and  $U'_1$  have rank  $r$ , we can set  $x^* = \Delta U'_1 x$  and get the equality above. Thus the columns of  $V_1$  span the same space as the columns of  $A$ . Since  $V$  is an orthonormal basis for  $R^n$ , the columns of  $V_1$  are an orthonormal basis for  $C(A)$ .

# Matrix Decompositions

- 3) Similar arguments show that

$$C(A') = C(U_1)$$

- 4) Claim:  $\mathcal{N}(A) = C(U_2) = C(A')^\perp$ .

Proof:

Since  $A = V_1 \Delta U'_1 + V_2 0 U'_2$ , we have  $AU_2 = V_1 \Delta U'_1 U_2 + V_2 0 U'_2 U_2$ , and thus  $AU_2 = 0 + V_2 0 = 0$ . Therefore  $AU_2 = 0$  which implies that  $\mathcal{N}(A) = C(U_2)$ .

- 5) Similar arguments show that  $\mathcal{N}(A') = C(V_2) = C(V_1)^\perp$ .

We have the following summary:

Matrix	Column space	Null space
$A$	$C(V_1)$	$C(U_2)$
$A'$	$C(U_1)$	$C(V_2)$

- 6) The columns of  $V_1$  are the eigenvectors corresponding to the nonzero eigenvalues of  $AA'$ , and the columns of  $U_1$  are the eigenvectors corresponding to the nonzero eigenvalues of  $A'A$ .

- 7) If  $r(A) = r$ , then  $\delta_1^2, \dots, \delta_r^2$  are the eigenvalues of  $A'A$ .

Can alternatively define SVD by taking  $U_1$  and  $V_1$  as the Gramschitted columns of  $A/A^T$ , but Delta is now a non-singular matrix. Gramschitting can be easier than getting eigenvalues/eigenvectors of  $AA^T$

# Matrix Decompositions

## Q-R factorization

Suppose  $A$  is an  $n \times p$  matrix with linearly independent columns. Then  $A$  can be written uniquely in the form:

$$A = QR$$

where  $Q_{n \times p}$  has orthonormal columns and  $R_{p \times p}$  is an upper triangular matrix with positive diagonal elements.

## Proof:

We construct  $Q$  and  $R$  by using the Gram-Schmidt orthogonalization process. Write  $A = (a_1, \dots, a_p)$ , where  $a_j$  is the  $j$ th column of  $A$ , given by

$$a_j = \begin{pmatrix} a_{1j} \\ \vdots \\ a_{nj} \end{pmatrix}_{n \times 1}$$

Now apply Gram-Schmidt to yield an orthogonal basis  $(u_1, \dots, u_p)$  for  $C(A)$ , such that

$$\mathcal{S}(a_1, \dots, a_k) = \mathcal{S}(u_1, \dots, u_k), \quad k \leq p .$$

# Matrix Decompositions

Note that  $u_k$  is a linear combination of  $(a_1, \dots, a_k)$ ,  $k = 1, \dots, p$ . Thus we can write the Q-R decomposition as  $U_{n \times p} = (u_1, \dots, u_p)$ , where  $u_k$  is an  $n \times 1$  vector corresponding to the  $k$ th column of  $u$ , given by

$$u_k = a_k - \sum_{i=1}^{k-1} \left( \frac{a'_k u_i}{u'_i u_i} \right) u_i . \quad (2)$$

Since  $u_k = S(a_1, \dots, a_k)$ , this means we can write

$$u_k = s_{1k} a_1 + \dots + s_{kk} a_k$$

for some set of scalars  $s_{1k}, \dots, s_{kk}$ . Now define the upper triangular matrix  $S$  as

$$S = \begin{pmatrix} s_{11} & s_{12} & \dots & s_{1p} \\ 0 & s_{22} & \dots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & s_{pp} \end{pmatrix} .$$

Now in matrix notation we can write the Gram-Schmidt representation of  $U$  as

$$\begin{aligned} U &= (u_1, \dots, u_p) \\ &= (s_{11} a_1, \dots, s_{1k} a_1 + \dots + s_{kk} a_k, \dots, s_{1p} a_1 + \dots + s_{pp} a_p) \\ &= AS \end{aligned}$$

# Matrix Decompositions

By definition of Gram-Schmidt, the diagonal elements of  $S$  are all 1. Now we normalize the columns of  $U$  by postmultiplying  $U$  by  $D$ , where

$$D = \text{diag}((u_1' u_1)^{-1/2}, \dots, (u_p' u_p)^{-1/2}).$$

Thus  $UD$  now has orthonormal columns and therefore

$$ASD = UD = Q.$$

Since  $D$  and  $S$  are nonsingular,

$$ASD = Q \Rightarrow A = Q(SD)^{-1}.$$

Note that  $(SD)^{-1}$  is an upper triangular matrix with positive diagonal elements. Therefore, we define  $R = (SD)^{-1}$ , and thus  $A = QR$ , where  $Q$  has orthonormal columns and  $R$  is upper triangular with positive diagonal elements.

# Projections

## Projections

Projections are special matrices (linear transformations) that are extremely useful in linear models. Most of the theory of linear models has to do with projections.

### General Definition of a Projection

Suppose  $\mathcal{M}$  is a vector space and  $N_1$ , and  $N_2$  are two subspaces in  $\mathcal{M}$ , where  $N_1 + N_2 = \mathcal{M}$  and  $N_1 \cap N_2 = 0$ . Consider the unique decomposition  $z = x + y$ , where  $x \in N_1$ , and  $y \in N_2$ . The linear transformation

Easiest example is a column space and its orthogonal compliment, but it doesn't have to be

$$P_{N_1|N_2} z = x$$

is called the projection of  $z$  onto the subspace  $N_1$  along the subspace  $N_2$ .

Note: The vector  $x$  is the result of the projection of  $z$ .

### Theorem

The projection operator onto  $N_2$  along  $N_1$  is given by

$$P_{N_2|N_1} = I - P_{N_1|N_2} .$$

In this course, we will be concerned with projections in  $\mathcal{M} = \mathbb{R}^n$ .

# Projections

## Projection in 3-dimensional space

Remarks:

- 1) If the projection is at a “right angle”, then it is called an orthogonal projection and such projections are unique.
- 2) If the projection is not at a right angle, then it is called a projection and it is not unique.

Defn:

Let  $A$  be an  $n \times n$  matrix.  $A$  is said to be a projection operator onto  $C(A)$  along  $\mathcal{N}(A)$  if for any  $v \in C(A)$ ,

$$Av = v .$$

Defn: If  $A^2 = A$ , then  $A$  is said to be an idempotent matrix.

Theorem  $A^2 = A$  if and only if  $A$  is a projection matrix.

Defn:  $M$  is a perpendicular (orthogonal) projection operator (matrix) onto  $C(X)$  if and only if

- i)  $v \in C(X) \Rightarrow Mv = v$  (projection)
- ii)  $w \in C(X)^\perp \Rightarrow Mw = 0$  (orthogonal)

# Projections

## Theorem

If  $M$  is a orthogonal projection operator onto  $C(X)$ , then  $C(M) = C(X)$ .

## Proof

We must show that  $C(M) \subset C(X)$  and  $C(X) \subset C(M)$ .

1)  $\Rightarrow C(X) \subset C(M)$ .

Let  $v \in C(X)$ , then  $Mv \in C(M)$ , since

$C(M) = \{z : Mt = z, z \in R^n\}$ . But  $Mv = v$  since  $M$  is a projection, thus  $v \in C(M)$ .

2)  $\Leftarrow C(M) \subset C(X)$ .

Let  $v \in C(M)$ . Then there exists a  $t$  such that  $Mt = v$ . Write

$t = t_1 + t_2$ , where  $t_1 \in C(X)$ ,  $t_2 \in C(X)^\perp$ . Now

$Mt = M(t_1 + t_2) = Mt_1 + Mt_2 = v$ . Since  $M$  is an orthogonal projection operator onto  $C(X)$ ,  $Mt_2 = 0$ , and thus  $Mt_1 = v$ . Since  $t_1 \in C(X)$ ,  $Mt_1 = t_1$ , and this implies  $t_1 = v$ . Thus for  $v \in C(M)$ , we have  $Mv = v$  which implies  $v \in C(X)$ .

# Projections

## Theorem

$M$  is an orthogonal projection operator onto  $C(M)$  if and only if  $M = M^2$  and  $M = M'$ .

Thus a matrix is an orthogonal projection operator if it is idempotent and symmetric.

## Proof:

1)  $\Rightarrow$

Suppose  $M$  is an orthogonal projection operator. We want to show that  $M^2 = M$  and  $M = M'$ .

Let  $v \in R^n$ , and write  $v = v_1 + v_2$ , where  $v_1 \in C(M)$ ,  $v_2 \in C(M)^\perp$ . Then

$$\begin{aligned}M^2v &= M^2(v_1 + v_2) \\&= M^2v_1 + M^2v_2 = M(Mv_1) + M(Mv_2) \\&= Mv_1 + Mv_2 = Mv_1 = M(v_1 + v_2) = Mv.\end{aligned}$$

Thus  $M^2v = Mv$  for any  $v \in R^n$ . This implies that  $(M^2 - M)v = 0$  for any  $v \in R^n$ , and thus  $M^2 - M = 0 \Rightarrow M^2 = M$ .

# Projections

To see that  $M = M'$ , let  $w = w_1 + w_2$ , where  $w_1 \in C(M)$ ,  $w_2 \in C(M)^\perp$ . Also write  $v = v_1 + v_2$ , where  $v_1 \in C(M)$  and  $v_2 \in C(M)^\perp$ . Since

$$\begin{aligned}(I - M)v &= (I - M)(v_1 + v_2) \\&= v_1 + v_2 - (Mv_1 + Mv_2) \\&= v_1 + v_2 - v_1 - Mv_2 \\&= (I - M)v_2 = v_2.\end{aligned}$$

Thus we get

$w' M'(I - M)v = w'_1 M'(I - M)v_2 = w'_1 v_2 = 0$ . This is true for any  $v$  and  $w$ , so that  $M'(I - M) = 0$  implies  $M' = M'M$ . Since  $M'M$  is symmetric,  $M'$  must be symmetric, hence  $M = M'$ .

# Projections

## Theorem

Orthogonal projection operators are unique.

## Proof:

Let  $M$  and  $P$  be two orthogonal projection operators onto the same space  $C(M)$ . Let  $v \in R^n$  and write  $v = v_1 + v_2$ ,  $v_1 \in C(M)$  and  $v_2 \in C(M)^\perp$ .

$$Mv = M(v_1 + v_2) = Mv_1 + Mv_2 = Mv_1 = v_1, \quad v_1 \in C(M).$$

Now it must be the case that  $Pv = P(v_1 + v_2) = Pv_1 = v_1$ , and this implies  $Mv = Pv$ , which implies  $(M - P)v = 0$ . Thus  $M = P$ .

Note: Projection operators are not unique in general.

# Projections

## Illustration of orthogonal projection in linear models

Consider the linear model

$$Y = X\beta + \epsilon,$$

where  $E(\epsilon) = 0$ , and  $\text{Cov}(\epsilon) = \sigma^2 I$ . We have  $E(Y) = \mu = X\beta$ . If  $X$  has full rank (i.e.,  $r(X) = p$ ), then  $X'X$  is invertible and the least squares estimator of  $\beta$  is  $\hat{\beta} = (X'X)^{-1}X'Y$ . The least squares estimator of  $\mu = X\beta$  is  $\hat{\mu} = X\hat{\beta} = X(X'X)^{-1}X'Y = MY$ , where  $M = X(X'X)^{-1}X'$ .  $\hat{\mu}$  is the orthogonal projection of  $Y$  onto  $C(X)$ .

# Projections

## Theorem

Suppose  $M$  is an  $n \times n$  orthogonal projection operator of rank  $r \leq n$ . Then

- 1) The eigenvalues of  $M$  are 0 or 1.
- 2)  $r(M) = \text{tr}(M) = r$ .
- 3)  $M$  is a positive semidefinite matrix.

## Proof of 1)

Since  $M$  is an orthogonal projection operator,  $M = M^2$  and  $M = M'$ . Let  $\lambda$  be an eigenvalue of  $M$  with eigenvector  $x$ . Thus  $Mx = \lambda x$  and

$M^2x = M(Mx) = M(\lambda x) = \lambda Mx = \lambda^2 x$ . But  $M = M^2$  implies  $Mx = M^2x$ , which implies  $\lambda x = \lambda^2 x \Rightarrow (\lambda - \lambda^2)x = 0 \Rightarrow \lambda(1 - \lambda) = 0 \Rightarrow \lambda = 0$  or  $\lambda = 1$ .

# Projections

Proof of 2)

$r(M) = r$ . By definition,

$$\begin{aligned} \text{tr}(M) &= \sum_{i=1}^n \lambda_i \\ &= \sum_{i=1}^r \lambda_i + \sum_{i=r+1}^n \lambda_i \\ &= \sum_{i=1}^r 1 + \sum_{i=r+1}^n 0 = r \end{aligned}$$

since all eigenvalues are either 0 or 1, and exactly  $r$  of them are 1.

Proof of 3)

This proof is immediate since  $M = M'$  is symmetric, and therefore all eigenvalues are nonnegative. Therefore,  $M$  is positive semidefinite.

# Projections

## Theorem

Suppose  $X$  is an  $n \times p$  of rank  $r \leq \min(n, p)$ . Suppose  $\{a_1, \dots, a_r\}$  is an orthonormal basis for  $C(X)$ . Let  $A = (a_1, \dots, a_r)$ , where  $a_i$  is  $n \times 1$ . Then,

$$AA' = \sum_{i=1}^r a_i a_i'$$

is the unique orthogonal projection operator onto  $C(X)$ .

## Proof

Clearly  $AA'$  is symmetric. We need to show that  $(AA')^2 = AA'$  and  $C(AA') = C(X)$ . Now  $(AA')^2 = (AA')(AA') = A(A'A)A' = AI_{r \times r}A' = AA'$ .

Note here that

$$A'A = \begin{pmatrix} a_1' \\ \vdots \\ a_r' \end{pmatrix}_{r \times n} (a_1, \dots, a_r)_{n \times r}$$

$$= \begin{pmatrix} a_1'a_1 & a_1'a_2 & \dots & a_1'a_r \\ a_1'a_2 & a_2'a_2 & \dots & a_2'a_r \\ \vdots & \vdots & \ddots & \vdots \\ a_1'a_r & \dots & \dots & a_r'a_r \end{pmatrix}$$

# Projections

$$= \begin{pmatrix} 1 & \cdots & 0 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 1 \end{pmatrix}_{r \times r} = I_{r \times r}.$$

Now we need to show  $C(AA') = C(X)$ .

1)  $\Rightarrow$

Let  $x \in C(AA')$ . Then  $x = AA't$  for some  $t$ . Let  $t^* = A't$ . Then  $AA't = At^* = x$ , and thus  $x \in C(X)$ , since  $C(X) = \{z : At^* = z\}$ . Thus  $C(AA') \subset C(X)$ .

2)  $\Leftarrow$

let  $x \in C(X)$ , then  $x \in C(A)$  since  $C(A) = \mathcal{S}(a_1, \dots, a_r) = C(X)$ . Now  $x \in C(A)$  implies  $x = At$  for some  $t$ . Since  $A$  has full rank, (i.e.,  $r(A) = r$ ), any vector  $t \in R^r$  can be written as  $t = A'z$ , where  $z \in R^n$ . Thus  $x = At = AA'z$ , which implies  $x \in C(AA')$ . Thus  $C(X) \subset C(AA')$ , and this completes the proof.

# Projections

## Theorem (special case)

Suppose  $X$  is an  $n \times p$  matrix of rank  $p$ . Define  $M = X(X'X)^{-1}X'$ . Then  $M$  is the orthogonal projection operator onto  $C(X)$  (along  $C(X)^\perp$ ). If  $X$  is  $n \times 1$  then  $M = \frac{XX'}{X'X}$ .

$X$  has rank  $P$ , meaning  $X^T$  has rank  $P$ , and  $X^TX$  is rank preserving (pg. 58). If you post-multiply by a matrix of the same rank, you get the same rank (prop 2, pg. 58). Furthermore,  $(X^TX)$  is clearly symmetric, which means its inverse is symmetric

## Proof:

Since  $X$  has full rank  $p$ ,  $(X'X)^{-1}$  exists. Clearly  $M$  is symmetric since  $(X(X'X)^{-1}X')' = (X')'(X'X)^{-1}X' = X(X'X)^{-1}X'$ . Now  $M^2 = (X(X'X)^{-1}X')(X(X'X)^{-1}X') = X(X'X)^{-1}(X'X)(X'X)^{-1}X' = X(X'X)^{-1}X' = M$ .

Finally, we need to show that  $C(M) = C(X)$ .  $C(X) = \{z : Xt = z\}$  and  $C(M) = \{z^* : X(X'X)^{-1}X't = z^*\} = \{z^* : Xt^* = z^*\} = C(X)$ . Thus  $M$  is the orthogonal projection onto  $C(X)$ .

## Theorem

Suppose  $M = X(X'X)^{-1}X'$  as above. Then  $M$  can be written  $M = QQ' = V_1V_1'$  where  $Q$  and  $V_1$  are defined as in the  $QR$  and  $SVD$  decompositions, respectively.

# Projections

## Theorem

$I - M$  is the unique orthogonal projection operator onto  $C(X)^\perp$ .

## Proof

- 1)  $(I - M)' = I - M' = I - M$ , thus  $I - M$  is symmetric.
- 2)  $(I - M)(I - M) = I - M - M + M^2 = I - M - M + M = I - M$ , thus  $I - M$  is idempotent.

Thus 1) and 2) prove that  $I - M$  is an orthogonal projection operator.

- 3) We need to show  $C(I - M) = C(X)^\perp$ . Let  $x \in C(M)$ . We want to show that for any  $y \in C(I - M)$ ,  $x'y = 0$ .

It is enough to show that  $(I - M)x = 0$ , where  $x \in C(M)$ .

$(I - M)x = x - Mx = x - x = 0$ . Thus  $C(X)^\perp = C(I - M)$ .

# Projections

## Theorem

Suppose  $M$  is an orthogonal projection operator. Let  $m_{ii}$  be the  $i$ th diagonal element of  $M$ . Then  $0 \leq m_{ii} \leq 1$ . Suppose  $M = X(X'X)^{-1}X'$  and  $1 \in C(X)$ , then  $\frac{1}{n} \leq m_{ii} \leq 1$ . If the number of rows of  $X$  exactly equal to  $x_i$  is  $c$ , then  $\frac{1}{n} \leq m_{ii} \leq \frac{1}{c}$ .

Proof is an exercise.

## Theorem

Suppose  $X$  is a  $n \times p$  matrix of rank  $p$ . Write  $X = (X_1, X_2)$  Where  $X_1$  is  $n \times k$  and  $X_2$  is  $n \times (p - k)$  where  $r(X_1) = k$  and  $r(X_2) = p - k$ . Let  $M_j = X_j(X_j'X_j)^{-1}X_j'$ ,  $j = 1, 2$ , and let  $M = X(X'X)^{-1}X'$ . Further, let

$$X_j^* = (I - M_{3-j})X_j, \quad j = 1, 2,$$

and

$$M_j^* = X_j^{*'}(X_j^{*'}X_j^*)^{-1}X_j^{*'}.$$

Thus  $X_1^*$  consists of columns which are orthogonal to  $X_2$  and  $X_2^*$  has columns which are orthogonal to  $X_1$ . Therefore,  $M = M_1 + M_2^*$  and  $M = M_2 + M_1^*$ .

# Generalized Inverses

## Generalized inverses

Singular matrices arise much in the theory of linear models. Suppose  $X$  is  $n \times p$  of rank  $r < \min(n, p)$ . Then  $(X'X)^{-1}$  does not exist. In these cases, we will need the notion of a generalized inverse of  $X'X$  so that estimates can be computed.

## Definition of Generalized Inverse

Consider the linear transformation  $A : R^p \longrightarrow R^n$ . A generalized inverse of  $A$  is the linear transformation  $A^-$  such that

$$AA^-y = y \quad \text{for all } y \in C(A).$$

Note: Since  $A : R^p \longrightarrow R^n$ ,  $A^- : R^n \longrightarrow R^p$ .

This definition is equivalent to the following.

Defn:

Suppose  $A$  is an  $n \times p$  matrix, then  $A_{p \times n}^-$  is a generalized inverse of  $A$  if

$$AA^-A = A.$$

Recall that  $y \in C(A)$  means  $y = Ax$  for some  $x \in R^p$ . Thus, substituting this into the definition, we get

$$AA^-y = y, \quad y \in R^n,$$

and  $y = Ax$ .

# Generalized Inverses

Note that by the definition of a generalized inverse, we have

$$(A^\perp A)(A^\perp A) = A^\perp(AA^\perp A) = A^\perp A .$$

Thus  $A^\perp A$  is idempotent, and hence a projection.

The generalized inverse is not unique. We want to focus on a specific type of generalized inverse that satisfies additional properties.

## Moore-Penrose Generalized Inverse

Suppose  $A$  is an  $n \times p$  matrix. The Moore-Penrose generalized inverse (M-P g-inverse) of  $A$  is a  $p \times n$  matrix  $A^+$  such that

- 1)  $(AA^+)' = AA^+$  ( $AA^+$  is symmetric)
- 2)  $(A^+A)' = A^+A$  ( $A^+A$  is symmetric)
- 3)  $AA^+A = A$  ( $A^+$  is a g-inverse of  $A$ )
- 4)  $A^+AA^+ = A^+$  ( $A$  is a g-inverse of  $A^+$ )

From the definition, it is immediate that  $AA^+$  and  $A^+A$  are orthogonal projection operators.

# Generalized Inverses

## Theorem

Every matrix of  $A$  has a Moore-Penrose generalized inverse.

## Proof

Suppose  $A$  has rank  $r$ . From the SVD of  $A$ , we can write  $A = V_1 \Delta U_1'$  where  $\Delta$  is an  $r \times r$  diagonal matrix with positive elements, and  $V_1$  and  $U_1$  have orthonormal columns. Define  $A^+ = U_1 \Delta^{-1} V_1'$ . We must show that  $A^+$  satisfies the four conditions of a MP g-inverse.

- 1)  $AA^+ = V_1 \Delta U_1' U_1 \Delta^{-1} V_1' = V_1 V_1'$  which is clearly symmetric.
- 2)  $A^+A = U_1 \Delta^{-1} V_1' V_1 \Delta U_1' = U_1 U_1'$  which is symmetric.
- 3)  $AA^+A = V_1 \Delta U_1' U_1 \Delta^{-1} V_1' V_1 \Delta U_1' = V_1 \Delta U_1' = A$ .
- 4)  $A^+AA^+ = U_1 \Delta^{-1} V_1' V_1 \Delta U_1' U_1 \Delta^{-1} V_1' = A^+$ .

## Theorem

The Moore-Penrose generalized inverse is unique.

Proof is an exercise.

# Generalized Inverses

Examples of Moore-Penrose Generalized Inverses.

- 1) If  $M$  is an orthogonal projection operator then  $M^+ = M$ .
- 2) If  $A$  is an  $n \times n$  non-singular matrix then  $A^+ = A^{-1}$
- 3) If  $A = \text{diag}(a_{11}, \dots, a_{nn})$  then  $A^+$  is a diagonal matrix with elements  $1/a_{ii}$  if  $a_{ii} \neq 0$  and 0 if  $a_{ii} = 0$ .
- 4) Suppose  $A$  is an  $n \times p$  matrix and  $r(A) = n$ . Then  $A^+ = A'(AA')^{-1}$ .
- 5) Suppose  $A$  is an  $n \times p$  matrix of rank  $p$ , then  $A^+ = (A'A)^{-1}A'$ .
- 6)  $r(A) = r(A^+)$ .
- 7) For any matrix  $A$ ,  $(A^+)' = (A')^+$ .
- 8) If  $A$  is symmetric, then  $A^+$  is symmetric.
- 9)  $(A^+)^+ = A$ .
- 10) Suppose  $A, B$  are square and either  $A$  or  $B$  is singular. Then  $(AB)^+ \neq B^+A^+$  in general.
- 11)  $(AB)^+ = B^+A^+$  if  $A$  and  $B$  are both non-singular.
- 12) If  $A$  and  $B$  are not square, then  $(AB)^+ \neq B^+A^+$  in general.

# Generalized Inverses

## Counterexample

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad B = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad AB = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

$(AB)^+ = (1, 0)$  and  $B^+ = (\frac{1}{2}, \frac{1}{2})$ .

$$A^+ = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$$

$$B^+ A^+ = (\frac{1}{2}, 0) \neq (AB)^+.$$

There are two important special cases for which the equality will hold.

1) For any  $A$ ,  $(A'A)^+ = A^+(A')^+$ .

2) If  $A$  is  $n \times p$  of rank  $p$  and  $B$  is  $p \times r$  of rank  $p$ , then  $(AB)^+ = B^+ A^+$ .

# Generalized Inverses

## Theorem

$M = XX^+$  is the unique orthogonal projection onto  $C(X)$ .

## Proof

Suppose  $X$  is  $n \times p$  of rank  $r$ . We see that  $M' = (XX^+)' = XX^+$  and  $M^2 = (XX^+)(XX^+) = X(X^+XX^+) = XX^+$ . Thus  $M$  is an orthogonal projection. Finally we need to show that  $C(XX^+) = C(X)$ .

1)  $\Rightarrow$

We first show that  $C(XX^+) \subset C(X)$ .

Let  $v \in C(XX^+)$ , then  $v = XX^+t$  for some  $t \in R^n$ . Now

$v = XX^+t = Xt^*$ . Thus  $v = Xt^*$ , and therefore  $v \in C(X)$ . Thus  $C(XX^+) \subset C(X)$ .

2) Instead of showing that  $C(X) \subset C(XX^+)$ , we can show that  $r(XX^+) = r(X)$ . This along with 1) will complete the proof. Now  $r(XX^+) = \text{tr}(XX^+)$  since  $XX^+$  is an orthogonal projection operator.

Thus  $r(XX^+) = \text{tr}(XX^+) = \text{tr}(V_1 \Delta U_1' U_1 \Delta^{-1} V_1') = \text{tr}(V_1 V_1') = \text{tr}(V_1' V_1) = \text{tr}(I_{r \times r}) = r$ .

# Generalized Inverses

## Theorem

Let  $X^-$  be any generalized inverse of an  $n \times p$  matrix  $X$ . Then

$$X^* = X^- XX^- + (I - X^- X)A + B(I - XX^-)$$

is also a generalized inverse of  $X$  for any  $p \times n$  matrices  $A$  and  $B$ . If  $\dot{X}$  is another generalized inverse of  $X$ , then there is a choice of  $A$  and  $B$  for which  $X^* = \dot{X}$ .

## Proof

To show the first part, we need to show  $XX^*X = X$  for any  $A$  and  $B$ . To show the second part, take  $X^* = \dot{X}X\dot{X} + (I - \dot{X}X)A + B(I - X\dot{X})$  and set  $A = \dot{X}$  and  $B = \dot{X}X\dot{X}$ .

# Square Root of a Matrix

Defn:

Suppose  $A$  is an  $n \times n$  positive semidefinite matrix with spectral decomposition  $A = P\Lambda P'$ . The square root of  $A$  is defined as

$$A^{1/2} = P\Lambda^{1/2}P'$$

where  $\Lambda^{1/2} = \text{diag}(\lambda_1^{1/2}, \dots, \lambda_n^{1/2})$ .

Note that

$$\begin{aligned} A^{1/2}A^{1/2} &= (P\Lambda^{1/2}P')(P\Lambda^{1/2}P') \\ &= (P\Lambda^{1/2})(P'P)(\Lambda^{1/2}P') \\ &= P\Lambda^{1/2}\Lambda^{1/2}P' \\ &= P\Lambda P' \\ &= A \end{aligned}$$

If  $A^{-1}$  exists (i.e.,  $A$  is positive definite), then

$$A^{-1/2} = P\Lambda^{-1/2}P' .$$

where  $\Lambda^{-1/2} = \text{diag}(\lambda_1^{-1/2}, \dots, \lambda_n^{-1/2})$ . Note that

$$A^{-1/2} = (A^{1/2})^{-1}$$

# Generalized Inverses

## Theorem

For any matrix  $A$ , there exists a generalized inverse of  $A$ .

## Theorem

If  $G_1$  and  $G_2$  are generalized inverses of  $A$ , then so is  $G_1AG_2$ .

## Proof:

$$\begin{aligned} A(G_1AG_2)A &= (AG_1A)G_2A \\ &= AG_2A \\ &= A \end{aligned}$$

Theorem If  $A$  is symmetric, then there exists a g-inverse of  $A$  that is symmetric, i.e.,  $(A^-)' = A^-$ .

Note: The MP g-inverse is symmetric.

## Defn:

A generalized inverse  $A^-$  for a matrix  $A$  that has the property

$$A^-AA^- = A^-$$

is said to be reflexive.

Note: MP g-inverse is reflexive.

# Projections

Theorem Suppose  $X$  is an  $n \times p$  matrix of rank  $r$ . Consider the matrix  $X'X$ . (note that  $X'X$  is also of rank  $r$ ). If  $G$  and  $H$  are generalized inverses of  $X'X$ , then

- i)  $XGX'X = XHX'X = X$ .
- ii)  $XGX' = XHX'$ .

Proof:

i) Let  $v \in R^n$ , and write  $v = v_1 + v_2$ , where  $v_1 \in C(X)$ ,  $v_2 \in C(X)^\perp$ . Since  $v_1 \in C(X)$ ,  $v_1 = Xb$  for some  $b \in R^p$ . Then

$$\begin{aligned} v'XGX'X &= (v'_1 + v'_2)XGX'X \\ &= v'_1XGX'X \quad \text{since } v'_2X = 0 \\ &= b'X'(XGX'X) \\ &= b'(X'X)G(X'X) \\ &= b'(X'X) \\ &= v'_1X \\ &= v'X. \end{aligned}$$

Thus,  $v'XGX'X = v'X$ . Since  $v$  and  $G$  are arbitrary, this implies  $XGX'X = X$  for any  $G$ .

# Projections

- ii) Let  $v \in R^n$ ,  $v = v_1 + v_2$ , with  $v_1 \in C(X)$  and  $v_2 \in C(X)^\perp$ . Then  $v_1 = Xb$  for some  $b \in R^p$ . Thus

$$\begin{aligned} XGX'v &= XGX'(v_1 + v_2) \\ &= XG(X'v_1 + X'v_2) \\ &= XGX'v_1 \\ &= XGX'Xb \\ &= XHX'Xb \quad (\text{by part i) of this theorem}) \\ &= XHX'v. \end{aligned}$$

Since  $v$  is arbitrary, we have  $XGX' = XHX'$  for any choice of  $G$  and  $H$ .

## Remarks

- 1) Part ii) of this theorem says that  $X(X'X)^{-}X'$  is invariant with respect to the choice of generalized inverse.
- 2) It also follows that  $X(X'X)^{-}X'$  is symmetric for any choice of generalized inverse, because by a previous theorem, we know that there exists a generalized inverse of  $X'X$  that is symmetric. This, together with part ii) of the theorem, implies that  $X(X'X)^{-}X'$  is symmetric for any choice of generalized inverse.

This theorem now leads to the following result.

# Projections

## Theorem

Suppose  $X$  is an  $n \times p$  matrix of rank  $r$ . Let  $M = X(X'X)^{-}X'$ , where  $-$  denotes any generalized inverse. Then  $M$  is the unique orthogonal projection operator onto  $C(X)$ .

## Proof:

We prove this theorem by using the definition of an orthogonal projection.

1) Let  $v \in C(X)$ . Then  $v = Xb$  for some  $b \in R^p$ . Thus

$$\begin{aligned} Mv &= X(X'X)^{-}X'v \\ &= X(X'X)^{-}X'Xb \\ &= Xb \quad \text{by part i) of the previous theorem} \\ &= v . \end{aligned}$$

Thus for any  $v \in C(X)$ ,  $Mv = v$ , which means that  $M$  is a projection operator onto  $C(X)$  by its definition.

# Projections

2) To show orthogonality, we consider  $w \in C(X)^\perp$ . Then  $w'x_j = 0$ ,  $j = 1, \dots, p$ , where  $x_j$  is the  $j$ th column of  $X$ . By taking transposes, we have  $x_j'w = 0$ ,  $j = 1, \dots, p$ . Thus

$$\begin{aligned} Mw &= X(X'X)^{-1}X'w \\ &= X(X'X)^{-1}0 \\ &= 0 \end{aligned}$$

Thus  $M$  is the unique orthogonal projection operator onto  $C(X)$ .

# Projections

## Different forms for $M$

Suppose  $X$  is an  $n \times p$  matrix of rank  $r$ . Let  $M$  denote the orthogonal projection operator onto  $C(X)$ . Then

- i)  $M = QQ'$  where  $Q$  is the matrix from the QR decomposition of  $X = QR$ .
- ii)  $M = V_1 V_1'$  where  $V_1$  comes from the SVD of  $X = V_1 \Delta U_1'$ .
- iii)  $M = XX^+$  where  $X^+$  is the MP g-inverse of  $X$ .
- iv)  $M = X(X'X)^-X'$  where  $X'X^-$  is any g-inverse of  $X'X$ .

# Projections

## Theorem

Suppose  $M$  and  $M_0$  are orthogonal projection operators with  $C(M_0) \subset C(M)$ .

Then

- i)  $MM_0 = M_0M = M_0$ .
- ii)  $M - M_0$  is an orthogonal projection operator.
- iii)  $C(M_0) \perp C(M - M_0)$ .
- iv)  $C(M - M_0) = C(M) \cap C(M_0)^\perp$ .
- v)  $M - M_0$  is the orthogonal projection operator onto  $C(M - M_0)$ .

# Projections

Proof:

i)  $\Rightarrow$

Let  $x \in C(M_0M)$ , then  $M_0My = x$  for some  $y$ . Let  $z = My$ . Thus  $M_0z = x$  which implies  $x \in C(M_0)$ . Thus  $C(M_0M) \subset C(M_0)$ .

$\Leftarrow$

Let  $x \in C(M_0)$ . Now  $M_0Mx = M_0x = x$  since if  $x \in C(M_0)$ , then  $x \in C(M)$  and therefore  $Mx = x$ . Thus  $M_0Mx = x$  and thus  $x \in C(M_0M)$ . Therefore,  $C(M_0M) = C(M_0)$ .

Since  $C(M_0M) = C(M_0)$  and  $M_0$  is an orthogonal projection onto  $C(M_0M) = C(M_0)$ , this immediately implies that  $M_0M$  is an orthogonal projection onto  $C(M_0)$  and is equal to  $M_0$  since orthogonal projections onto the same space are unique, as shown earlier in the notes. Thus  $M_0M = M_0$ .

Now to show  $M_0 = MM_0$ , we make use of the fact that we just showed  $M_0M = M_0$ . We have  $M_0 = M'_0 = (M_0M)' = M'M'_0 = MM_0$ . Thus  $M_0 = MM_0$ .

Note: Recall that for any two orthogonal projection operators  $M_1$  and  $M_2$ ,  $C(M_1) = C(M_2)$  if and only if  $M_1 = M_2$ . This follows from the fact that orthogonal projection operators onto the same space are unique.



# Projections

ii)

$$\begin{aligned}(M - M_0)^2 &= (M - M_0)(M - M_0) \\&= M^2 - M_0M - MM_0 + M_0^2 \\&= M - M_0 - M_0 + M_0 \\&= M - M_0\end{aligned}$$

Thus  $M - M_0$  is a projection operator.

$$(M - M_0)' = M' - M'_0 = M - M_0$$

and thus  $M - M_0$  is an orthogonal projection operator.

iii)  $(M - M_0)M_0 = MM_0 - M_0^2 = M_0 - M_0 = 0.$

iv) Let  $x \in C(M - M_0)$ . Then  $(M - M_0)x = x$ , and thus  $Mx - M_0x = x$ .  
 $M_0x = 0$  since  $C(M - M_0) \perp C(M_0)$ . This implies that  $Mx = x$ , so that  
 $x \in C(M)$ . Hence  $x \in C(M) \cap C(M_0)^\perp$ .

Now let  $x \in C(M) \cap C(M_0)^\perp$ . Then  $x \in C(M)$  and  $x \in C(M_0)^\perp$ . Now  
 $(M - M_0)x = Mx - M_0x = x - 0 = x$ . Thus,  $x \in C(M - M_0)$ . Therefore  
 $C(M - M_0) = C(M) \cap C(M_0)^\perp$ .

v)  $(M - M_0)x = x$  for any  $x \in C(M - M_0)$  by part iv). Thus  $M - M_0$  is the  
orthogonal projection operator onto  $C(M - M_0)$ .

# Projections

## Theorem

Suppose  $M$  and  $M_0$  are orthogonal projection operators with  $C(M_0) \subset C(M)$ . Then  $C(M - M_0)$  is the orthogonal complement of  $C(M_0)$  with respect to  $C(M)$ .

Proof: We see that  $C(M - M_0) \perp C(M_0)$  from iii) above. This implies that  $C(M - M_0)$  is contained in the orthogonal complement of  $C(M_0)$  with respect to  $C(M)$ . If  $x \in C(M)$  and  $x \in C(M_0)^\perp$ , then

$$Mx = x = (M - M_0)x + M_0x = (M - M_0)x, \text{ and thus } x \in C(M - M_0).$$

Therefore, the orthogonal complement of  $C(M_0)$  with respect to  $C(M)$  is contained in  $C(M - M_0)$ .

This theorem now implies that

$$C(M) = C(M_0) + C(M - M_0)$$

and thus  $r(M) = r(M_0) + r(M - M_0)$ .

# Projections

## Theorem

Suppose  $M_1$  and  $M_2$  are two orthogonal projection operators in  $R^n$ . Then  $M_1 + M_2$  is the orthogonal projection operator onto  $C(M_1, M_2)$  if and only if  $C(M_1) \perp C(M_2)$ .

Note:  $C(M_1) \perp C(M_2)$  if and only if  $M_1M_2 = M_2M_1 = 0$ .

## Theorem

If  $M_1$  and  $M_2$  are symmetric matrices, with  $C(M_1) \perp C(M_2)$  and  $M_1 + M_2$  is an orthogonal projection operator, then  $M_1$  and  $M_2$  are orthogonal projection operators.

# Solving Systems of Linear Equations

## Solutions to Systems of Linear Equations

Consider the matrix equation

$$Y = X\beta \quad (1)$$

where  $Y_{n \times 1}$ ,  $X_{n \times p}$ , and  $\beta_{p \times 1}$ .

We ask the following question: For a given  $X$  and  $Y$ , does there exist a solution  $\beta$  to the equation above?

### Characterization of Solution

- 1) If  $p = n$  and  $X$  is nonsingular, the answer is YES and the unique solution is

$$\beta = X^{-1}Y.$$

In general, the solution depends on  $Y$ .

# Solving Systems of Linear Equations

- 2) Suppose  $p \leq n$ . If  $Y \in C(X)$ , the answer to the question above is YES again, since  $Y$  can be expressed as a linear combination of the columns of  $X$ . In fact,

$$\beta = X^{-} Y \quad (2)$$

is a solution, since by the definition of a generalized inverse

$$X\beta = XX^{-} Y = Y \quad \text{for all } Y \in C(X).$$

Is the solution in (2) unique? This depends on  $r(X)$ . If  $X$  has full rank, (i.e.,  $r(X) = p$ ), then the columns of  $X$  form a basis for  $C(X)$  and the coordinates of  $Y$  relative to that basis are unique and therefore the solution is unique.

However, the solution is not unique if  $r(X) < p$ . Thus, if  $\beta^*$  is a solution to  $Y = X\beta$ , then so is  $\beta^* + w$ , where  $w \in N(X)$ . Thus the set of all solutions is of the form

$$X^{-} Y + (I - X'(XX')^{-} X)z, \quad z \in R^p.$$

Note that  $X'(XX')^{-} X$  is the orthogonal projection operator onto  $C(X')$  and  $I - X'(XX')^{-} X$  is the orthogonal projection operator onto  $C(X')^\perp = N(X)$ .

# Solving Systems of Linear Equations

- 3) If  $Y \notin C(X)$ , and  $p < n$ , then no solution exists. This is the usual situation in linear models. In this case, we look for a vector in  $C(X)$  that is "closest" to  $Y$  and solve the system above with that vector instead of  $Y$ .

Let  $M = X(X'X)^{-}X'$ . We know that

$$Y = MY + (I - M)Y$$

Why does an orthogonal projection imply closeness?  
Define closeness independent of least squares.

**MY** is the orthogonal projection of  $Y$  onto  $C(X)$  and thus  $MY$  is the closest vector to  $Y$  in  $C(X)$ . We now solve the system

$$MY = X\beta .$$

We know the general solution to this system is

$$X^{-}MY + (I - X'(XX')^{-}X)z . \quad (3)$$

We can simplify this general solution in (3) a bit more. Consider the SVD of  $X$  and write  $X = V_1 \Delta U_1'$ . The MP g-inverse of  $X$  is  $X^+ = U_1 \Delta^{-1} V_1'$ . Choose the MP g-inverse in (3).

# Solving Systems of Linear Equations

3) Thus

$$\begin{aligned} X^+ M Y &= X^+ X (X' X)^+ X' Y \\ &= (U_1 \Delta^{-1} V'_1) (V_1 \Delta U'_1) (U_1 \Delta^{-2} U_1)' (U_1 \Delta V'_1) Y \\ &= U_1 \Delta^{-1} V'_1 Y \\ &= X^+ Y \end{aligned}$$

Thus, the general solution can be written as

$$X^+ Y + (I - X'(X X')^+ X)z, \quad z \in R^p.$$

If  $r(X) = p$ , then  $(I - X'(X X')^+ X)z = 0$  for any  $z \in R^p$ , and  $X^+ = (X' X)^{-1} X'$  so that

Go back to why this is. I think  
the Null space intersection with  
the column space is zero  
under certain scenarios.

$$X^+ Y = (X' X)^{-1} X' Y .$$

# Random Vectors and Matrices

## Random Vectors and Matrices

Suppose

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}$$

is a random vector with  $E(Y_i) = \mu_i$ ,  $\text{Var}(Y_i) = \sigma_{ii}$ , and  $\text{Cov}(Y_i, Y_j) = \sigma_{ij}$ . The expectation of  $Y$  is defined as

$$E(Y) = \begin{pmatrix} E(Y_1) \\ E(Y_2) \\ \vdots \\ E(Y_n) \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix} = \mu .$$

# Random Vectors and Matrices

Defn:

Suppose  $Z$  is an  $n \times p$  matrix of random variables. Then

$$E(Z) = \begin{pmatrix} E(Z_{11}) & \dots & E(Z_{1p}) \\ \vdots & \dots & \vdots \\ E(Z_{n1}) & \dots & E(Z_{np}) \end{pmatrix}$$

Thus the expectation of a random matrix is the matrix of the expectations.

Suppose  $Y$  is an  $n \times 1$  vector of random variables. The covariance matrix of  $Y$  is defined as

$$\begin{aligned} \text{Cov}(Y) &= E[(Y - \mu)(Y - \mu)'] \\ &= \Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1n} \\ \sigma_{21} & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{n1} & \dots & \dots & \sigma_{nn} \end{pmatrix} \end{aligned}$$

where  $\sigma_{ij} = E[(Y_i - \mu_i)(Y_j - \mu_j)]$ ,  $i, j = 1, \dots, n$ .

# Random Vectors and Matrices

## Theorem

Suppose  $Y$  is a random  $n \times 1$  vector with mean  $\mu = E(Y)$  and covariance matrix  $\Sigma = \text{Cov}(Y)$ . Moreover, suppose  $A$  is an  $r \times n$  matrix of constants, and  $b$  is an  $r \times 1$  vector of constants. Then

$$E(AY + b) = AE(Y) + b = A\mu + b$$

and

$$\text{Cov}(AY + b) = AC\text{Cov}(Y)A' = A\Sigma A' .$$

We leave the proof as an exercise.

## Defn

Let  $Y$  be an  $s \times 1$  random vector and  $W$  be an  $r \times 1$  random vector, with  $E(Y) = \mu$  and  $E(W) = \gamma$ . We define  $\text{Cov}(W, Y)$  as

$$\text{Cov}(W, Y) = E [(W - \gamma)(Y - \mu)'] .$$

Note that  $\text{Cov}(W, Y)$  is an  $r \times s$  matrix of covariances with  $ij$ th element  $\text{Cov}(W_i, Y_j)$ .

## Theorem

Let  $Y$  be an  $s \times 1$  random vector and  $W$  is an  $r \times 1$  random vector with  $\text{Cov}(W) = \Sigma_w$ ,  $\text{Cov}(Y) = \Sigma_y$ ,  $\text{Cov}(W, Y) = \Sigma_{wy}$ , and  $\text{Cov}(Y, W) = \Sigma_{yw}$ . Moreover, let  $A$  be an  $n \times r$  matrix of constants, and  $B$  is an  $n \times s$  matrix of constants. Then

$$\text{Cov}(AW + BY) = A\Sigma_w A' + B\Sigma_y B' + A\Sigma_{wy}B' + B\Sigma_{yw}A' .$$

# Random Vectors and Matrices

## Theorem

Covariance matrices are always positive semidefinite.

## Proof:

For  $Y_{n \times 1}$ ,  $\text{Cov}(Y) = E[(Y - \mu)(Y - \mu)']$ , where  $\mu = E(Y)$ . To show that  $\text{Cov}(Y)$  is positive semidefinite, we need to show that for any vector  $x \in R^n$ ,  $x' \text{Cov}(Y)x \geq 0$ . Let  $Z = Y - \mu$  for convenience. Then

$$\begin{aligned} x' \text{Cov}(Y)x &= x'E(ZZ')x \\ &= E(x'ZZ'x) \\ &= E(w'w) \text{ where } w = Z'x \\ &= E\left(\sum_{i=1}^n w_i^2\right) \\ &= \sum_{i=1}^n E(w_i^2) \geq 0 , \end{aligned}$$

since the expectation of a positive random variable is always nonnegative.

# Estimability

## Estimation

We now consider the problem of least squares and weighted least squares estimation in the linear model.

Consider the linear model

$$Y = X\beta + \epsilon \quad (4)$$

where

$$E(\epsilon) = 0, \quad \text{Cov}(\epsilon) = \sigma^2 I, \quad (5)$$

and  $Y$  is an  $n \times 1$  random vector,  $X$  is an  $n \times p$  fixed matrix of rank  $r \leq p$ ,  $\beta = (\beta_1, \dots, \beta_p)'$  is a  $p \times 1$  vector of regression coefficients, and  $\epsilon$  is an  $n \times 1$  vector of random errors.

Our goal is to estimate  $\beta$  or more generally, a linear function of  $\beta$ , say  $\lambda'\beta$ , where  $\lambda$  is a  $p \times 1$  vector of constants.

For example, if  $\lambda = (1, -1, 0, \dots, 0)'$ , then  $\lambda'\beta = \beta_1 - \beta_2$ .

We need to develop the notion of an estimable function.

# Estimability

## Definition of Estimability

$\lambda'\beta$  is estimable if there exists an  $n \times 1$  vector of constants  $\rho$ , such that

$$E(\rho' Y) = \lambda'\beta$$

for any  $\beta$ .

### Defn:

An estimate  $f(Y)$  of  $\lambda'\beta$  is said to be unbiased for  $\lambda'\beta$  if

$$E(f(Y)) = \lambda'\beta .$$

### Defn:

$f(Y)$  is a linear estimate of  $\lambda'\beta$  if

$$f(Y) = a_0 + a'Y$$

for some vectors of constants  $a_0$  and  $a$ .

Thus,  $\lambda'\beta$  is estimable if there exists a linear unbiased estimate of it. This leads to the following theorem.

# Estimability

## Theorem

$a_0 + a'Y$  is unbiased for  $\lambda'\beta$  if and only if  $a_0 = 0$  and  $a'X = \lambda'$ .

## Proof:

1) **Necessity:** If  $a_0 = 0$  and  $a'X = \lambda'$ , then  $E(a_0 + a'Y) = 0 + a'X\beta = \lambda'\beta$ .

Shouldn't this be reversed?

2) **Sufficiency:** If  $a_0 + a'Y$  is unbiased for  $\lambda'\beta$ , then

$\lambda'\beta = E(a_0 + a'Y) = a_0 + a'X\beta$  for any  $\beta$ . Subtracting  $a'X\beta$  from both sides gives

$$(\lambda' - a'X)\beta = a_0 \text{ for any } \beta.$$

This can only be true if  $a_0 = 0$  and  $\lambda' = a'X$ .

Why? Can  $a_0$  not depend on beta?

## Corollary

$\lambda'\beta$  is estimable if and only if there exists an  $n \times 1$  vector  $\rho$  such that

$$\rho'X = \lambda'.$$

why is this a corollary?

The statement above implies  $\lambda = X'\rho$ . Thus  $\lambda'\beta$  is estimable if  $\lambda \in C(X')$ .

We note that the concept of estimability is based entirely on the assumption that  $E(Y) = X\beta$ . Estimability does not depend on  $\text{Cov}(Y)$ .

# Estimability

## Defn:

Suppose  $\Lambda$  is a  $p \times s$  matrix of constants. Then the  $s \times 1$  vector of linear functions  $\Lambda'\beta$  is estimable if and only if there exists an  $n \times s$  matrix of constants  $P$  such that

$$P'X = \Lambda' .$$

## Remarks

- 1)  $\Lambda'\beta$  is estimable if each of its components is estimable.
- 2)  $P$  above is not unique. However  $MP$  is unique, where  $M = X(X'X)^{-1}X'$ . To see that  $MP$  is unique, let  $P_1, P_2$  be such that  $P_1'X = \Lambda'$  and  $P_2'X = \Lambda'$ . Then

$$\begin{aligned} MP_1 &= X(X'X)^{-1}X'P_1 \\ &= X(X'X)^{-1}\Lambda \\ &= X(X'X)^{-1}X'P_2 \\ &= MP_2 . \end{aligned}$$

- 3) The components of  $\beta$  need not be estimable, but linear combinations of the components of  $\beta$  may be estimable.
- 4) If  $X$  is of full rank, then  $\beta$  is estimable and every linear combination of the components is estimable.
- 5) If  $X$  is of full rank, then we can pick  $P' = (X'X)^{-1}X'$  and thus  $P'X\beta = \beta$ . Note here that  $\Lambda' = P'X = (X'X)^{-1}(X'X) = I_{p \times p}$ .

# Least Squares Estimation (LSE)

## Least Squares Estimation

Notation: The squared length of a vector will be denoted by  $\|.\|^2$ . Thus

$$\|Y\|^2 = Y'Y .$$

If  $A$  is an  $s \times n$  matrix, then

$$\|AY\|^2 = Y'A'AY .$$

If  $A$  is a orthogonal projection operator, then

$$\|AY\|^2 = Y'A'AY .$$

Defn:

The least squares estimate of  $\beta$ , denoted  $\hat{\beta}$ , satisfies

$$(Y - X\hat{\beta})'(Y - X\hat{\beta}) = \min_{\beta} (Y - X\beta)'(Y - X\beta) .$$

Thus, the least squares estimate of  $\beta$  minimizes the squared Euclidean distance between  $Y$  and its mean  $\mu = X\beta$ . If  $Y \notin C(X)$ , then we know that a solution of  $Y = X\beta$  does not exist in general. If  $Y \in C(X)$ , a solution exists. Thus the least squares solution will be the closest vector to  $Y$  in  $C(X)$ . We know that this vector is  $MY$ . We are led to the following theorem

# Least Squares Estimation (LSE)

## Theorem

$\hat{\beta}$  is a least squares solution to  $\beta$  if and only if

$$X\hat{\beta} = MY$$

where  $M = X(X'X)^{-1}X'$ . We note here that  $\hat{\beta}$  is not necessarily unique.  
Uniqueness will depend on estimability.

## Proof

Let  $\tilde{\beta}$  be an arbitrary estimate of  $\beta$ . We can write

$$\begin{aligned}(Y - X\tilde{\beta})'(Y - X\tilde{\beta}) &= \\(Y - MY + MY - X\tilde{\beta})'(Y - MY + MY - X\tilde{\beta}) &= \\&= (Y - MY)'(Y - MY) + (Y - MY)'(MY - X\tilde{\beta}) \\&\quad + (MY - X\tilde{\beta})'(Y - MY) + (MY - X\tilde{\beta})'(MY - X\tilde{\beta})\end{aligned}$$

We note here that:

$$(Y - MY)'(MY - X\tilde{\beta}) = Y'(I - M)MY - Y'(I - M)X\tilde{\beta} = 0 - 0 = 0.$$

Substitution now gives

$$\begin{aligned}(Y - X\tilde{\beta})'(Y - X\tilde{\beta}) &= \\&= (Y - MY)'(Y - MY) + (MY - X\tilde{\beta})'(MY - X\tilde{\beta})\end{aligned}$$

# Least Squares Estimation (LSE)

Both terms on the right hand side are nonnegative and the first term does not depend on  $\tilde{\beta}$ . Thus  $(Y - X\tilde{\beta})'(Y - X\tilde{\beta})$  will be minimized by minimizing  $(MY - X\tilde{\beta})'(MY - X\tilde{\beta})$ , which is the squared distance between  $MY$  and  $X\tilde{\beta}$ . This distance is 0 if and only if  $MY = X\tilde{\beta}$ .

## Corollary

$(X'X)^{-}X'Y$  is a least squares estimate of  $\beta$ . The set of all least squares estimates of  $\beta$  are of the form

$$\hat{\beta} = X^+Y + (I - X'(XX')^{-}X)z, \quad z \in R^p.$$

It turns out that least squares estimates of  $\lambda'\beta$  are unique if and only if  $\lambda'\beta$  is estimable. We are led to the following theorem.

# Least Squares Estimation (LSE)

## Theorem

$\lambda' = \rho' X$  if and only if  $\lambda' \hat{\beta}_1 = \lambda' \hat{\beta}_2$  for any  $\hat{\beta}_1, \hat{\beta}_2$  satisfying

$$X \hat{\beta}_1 = MY, \quad X \hat{\beta}_2 = MY.$$

## Proof:

1)  $\Rightarrow$

If  $\lambda' = \rho' X$ , then  $\lambda' \hat{\beta}_1 = \rho' X \hat{\beta}_1 = \rho' M Y = \rho' M Y = \rho' X \hat{\beta}_2 = \lambda' \hat{\beta}_2$ .

2)  $\Leftarrow$

Decompose  $\lambda$  into vectors in  $C(X')$  and  $C(X')^\perp$ . Moreover, let  $N = X'(XX')^- X$ .  $N$  is the orthogonal projection operator onto  $C(X')$ , and  $I - N$  is the orthogonal projection operator onto  $C(X')^\perp$ . Thus  $\lambda = X' \rho_1 + (I - N) \rho_2$ , where  $\rho_1 \in R^n$  and  $\rho_2 \in R^p$ . We have  $\lambda' = \rho_1' X + \rho_2' (I - N)$ ,  $X' \rho_1 \in C(X')$ , and  $(I - N) \rho_2 \in C(X')^\perp$ .

Thus

$$\begin{aligned} & \lambda' (\hat{\beta}_1 - \hat{\beta}_2) = 0 \\ \Leftrightarrow & (\rho_1' X + \rho_2' (I - N)) (\hat{\beta}_1 - \hat{\beta}_2) = 0 \\ \Leftrightarrow & \rho_1' (X \hat{\beta}_1 - X \hat{\beta}_2) + \rho_2' (I - N) (\hat{\beta}_1 - \hat{\beta}_2) = 0 \\ \Leftrightarrow & \rho_1' (M Y - M Y) + \rho_2' (I - N) (\hat{\beta}_1 - \hat{\beta}_2) = 0 \\ \Leftrightarrow & \rho_2' (I - N) (\hat{\beta}_1 - \hat{\beta}_2) = 0. \end{aligned}$$

# Least Squares Estimation (LSE)

2) Thus

$$\rho_2'(I - N)(\hat{\beta}_1 - \hat{\beta}_2) = 0 \quad (6)$$

for any  $\hat{\beta}_1 - \hat{\beta}_2$ . Let  $t = \hat{\beta}_1 - \hat{\beta}_2$ , and decompose  $t = t_1 + t_2$ , where  $t_1 \in C(X')$  and  $t_2 \in C(X')^\perp$ . Thus (6) implies

$$\begin{aligned} \rho_2'(I - N)(t_1 + t_2) &= 0 \Rightarrow \\ \rho_2'(I - N)t_1 + \rho_2'(I - N)t_2 &= 0 \end{aligned}$$

Now  $\rho_2'(I - N)t_1 = 0$  since  $t_1 \in C(X')$  by definition. This implies  $\rho_2'(I - N)t_2 = 0$  for any  $t_2 \in C(X')^\perp$ , and thus for any  $t \in R^p$ ,  $\rho_2'(I - N)t = 0$ . This implies that  $\rho_2'(I - N) = 0$ , which implies  $(I - N)\rho_2 = 0$ . Thus  $\lambda = X'\rho_1 + (I - N)\rho_2 = X'\rho_1 + 0 = X'\rho_1$ . Therefore  $\lambda \in C(X')$ . This completes the proof.

# Least Squares Estimation (LSE)

## Corollary

The unique least squares estimate of  $\rho'X\beta = \rho'MY$ .

## Corollary

The unique least squares estimate of  $P'X\beta$  is  $P'MY$ .

## Corollary

The unique least squares estimator of  $\mu = X\beta$  is  $MY$ .

If  $\lambda'\beta$  is estimable, then its unique least squares estimate is unbiased. This leads to the following theorem.

## Theorem

If  $\lambda' = \rho'X$ , then  $E(\rho'MY) = \lambda'\beta$ .

## Proof:

$$E(\rho'MY) = \rho'ME(Y) = \rho'MX\beta = \rho'X\beta = \lambda'\beta.$$

# Least Squares Estimation (LSE)

The squared length of  $MY$  is the regression sums of squares. That is

$$\|MY\|^2 = Y' MY .$$

## Theorem

Suppose  $Y$  is a random  $n \times 1$  vector, with  $E(Y) = \mu$  and  $\text{Cov}(Y) = \Sigma$ . Moreover, suppose  $A$  is any  $n \times n$  matrix. Then

$$E(Y'AY) = \mu'A\mu + \text{tr}(A\Sigma) . \quad (7)$$

The proof is a homework problem.

# Least Squares Estimation (LSE)

## Estimation of $\sigma^2$

We have the decomposition

$$Y = MY + (I - M)Y$$

Now  $MY$  is the least squares estimate of  $X\beta$ . We note that

$$MY = MX\beta + M\epsilon = X\beta + M\epsilon$$

so that

$$MY = X\beta + M\epsilon$$

where  $E(M\epsilon) = ME(\epsilon) = M0 = 0$ . Similarly, we have

$$(I - M)Y = (I - M)X\beta + (I - M)\epsilon = (I - M)\epsilon$$

so that  $(I - M)Y$  depends only on  $\epsilon$ . Since  $(I - M)Y$  depends only on  $\epsilon$ , it is reasonable to use some function of  $(I - M)Y$  to estimate  $\sigma^2$ . The function we use is the squared length of  $(I - M)Y$ .

# Least Squares Estimation (LSE)

## Theorem

Suppose  $r(X) = r$ . Then

$$\frac{\|(I - M)Y\|^2}{n - r} = \frac{Y'(I - M)Y}{n - r}$$

is an unbiased estimate of  $\sigma^2$ .

## Proof

Using the result in (7), we have

$$\begin{aligned} E(Y'(I - M)Y) &= (X\beta)'(I - M)X\beta + tr(\sigma^2 I(I - M)) \\ &= \beta' X'(I - M)X\beta + \sigma^2 tr(I - M) \\ &= 0 + \sigma^2(n - r) \end{aligned} \quad \text{See pg. 54}$$

Thus  $E\left(\frac{Y'(I - M)Y}{n - r}\right) = \sigma^2$ .

$(I - M)Y$  is the residual vector and its squared length is the error sum of squares. Thus  $\|(I - M)Y\|^2 = Y'(I - M)Y$  is the error sum of squares (SSE).  $\|(I - M)Y\|^2/(n - r)$  is the mean square error (MSE).

# Best Linear Unbiased Estimates (BLUE)

## Properties of Estimators

It is desirable to have estimators which satisfy certain properties such as

- 1) Unbiasedness
- 2) Minimum variance
- 3) Efficiency
- 4) Asymptotic normality

Suppose our goal is to estimate  $\lambda'\beta$  and we want to find the “best” linear unbiased estimate of it. The word “best” is in the sense of minimum variance. Thus, we seek linear estimators in  $Y$ , say  $a'Y$ , such that

$$E(a'Y) = \lambda'\beta$$

and

$$\text{Var}(a'Y) \leq \text{Var}(b'Y) \text{ for any } b \in R^n.$$

Can we find such an estimator? The answer is YES, and this leads us to the Gauss-Markov theorem.

# Best Linear Unbiased Estimates (BLUE)

## Theorem (Gauss-Markov)

Consider the linear model

$$Y = X\beta + \epsilon$$

where  $E(\epsilon) = 0$  and  $\text{Cov}(\epsilon) = \sigma^2 I$ ,  $\sigma^2 > 0$ . If  $\lambda' \beta$  is estimable, then the (unique) least squares estimate of  $\lambda' \beta$ ,  $\rho' MY$ , is the unique best linear unbiased estimator (BLUE) of  $\lambda' \beta$ .

### Proof

Let  $M = X(X'X)^{-1}X'$ . Since  $\lambda' \beta$  is estimable, let  $\lambda' = \rho' X$  for some  $\rho$ . We need to show that if  $a' Y$  is an unbiased estimate of  $\lambda' \beta$ , then

$$\text{Var}(a' Y) \geq \text{Var}(\rho' MY) \quad \text{for any } a \in R^n.$$

Since  $a' Y$  is unbiased for  $\lambda' \beta$ ,  $\lambda' \beta = E(a' Y) = a' E(Y) = a' X \beta$  for any  $\beta$ .

Therefore,  $\rho' X = \lambda' = a' X$ . Now write

$$\begin{aligned}\text{Var}(a' Y) &= \text{Var}(a' Y - \rho' MY + \rho' MY) \\ &= \text{Var}(a' Y - \rho' MY) + \text{Var}(\rho' MY) \\ &\quad + 2\text{Cov}(a' Y - \rho' MY, \rho' MY)\end{aligned}$$

Since  $\text{Var}(a' Y - \rho' MY) \geq 0$ , if we show that  $\text{Cov}(a' Y - \rho' MY, \rho' MY) = 0$ , then this will establish minimum variance.

# Best Linear Unbiased Estimates (BLUE)

$$\begin{aligned}\text{Cov}(a'Y - \rho'MY, \rho'MY) &= \text{Cov}((a' - \rho'M)Y, \rho'MY) \\&= (a' - \rho'M)\text{Cov}(Y)(\rho'M)' \\&= (a' - \rho'M)(\sigma^2 I)M\rho \\&= \sigma^2(a' - \rho'M)M\rho \\&= \sigma^2(a'M - \rho'M)\rho\end{aligned}$$

As shown above,  $a'X = \rho'X$ . This implies  $a'X(X'X)^{-}X' = \rho'X(X'X)^{-}X'$ , and therefore  $a'M = \rho'M$ . Thus it follows that  $\sigma^2(a'M - \rho'M)\rho = 0$ . This establishes minimum variance. Now we want to show that  $\rho'MY$  is unique. We have just shown that for any linear unbiased estimate  $a'Y$  of  $\lambda'\beta$ ,

$$\text{Var}(a'Y) = \text{Var}(\rho'MY) + \text{Var}(a'Y - \rho'MY).$$

Thus if  $a'Y$  is BLUE of  $\lambda'\beta$ , then it must be true that  $\text{Var}(a'Y - \rho'MY) = 0$ .

# Best Linear Unbiased Estimates (BLUE)

Since  $a'Y$  and  $\rho'MY$  are both unbiased, it is clear that

$$\begin{aligned} 0 &= \text{Var}(a'Y - \rho'MY) \\ &= \text{Var}((a' - \rho'M)Y) \\ &= (a' - \rho'M)(\sigma^2 I)(a' - \rho'M)' \\ &= \sigma^2(a - M\rho)'(a - M\rho) \\ &= \sigma^2 \|a - M\rho\|^2 \end{aligned}$$

Thus  $\sigma^2 \|a - M\rho\|^2 = 0$  if and only if  $a - M\rho = 0$ , which implies  $a = M\rho$ . Thus  $\rho'MY$  is the unique BLUE of  $\lambda'\beta$ .

## Remarks

- 1)  $C(X)$  is sometimes called the estimation space and  $C(X)^\perp$  is sometimes called the error space.

# Weighted Least Squares

## Weighted Least Squares

Consider the linear model

$$Y = X\beta + \epsilon \quad (8)$$

where

$$E(\epsilon) = 0 \text{ and } \text{Cov}(\epsilon) = \sigma^2 V$$

Examples of unknown  $V$  include generalized linear mixed effects models (GLMMs).  $\sigma^2$ ,  $\epsilon$  and  $\beta$  are still unknown

and  $V$  is a **known** positive definite matrix.

this is the key here

Applications with covariance matrices equal to  $\sigma^2 V$

- 1) Repeated measures. Several structures for  $V$  have been proposed in the literature.  $V$  is typically unknown.
- 2) Split-plot design models. Here the covariance matrix is the intraclass covariance matrix. This means that all of the variances are equal and all of the covariances are equal.  $V$  is typically unknown.
- 3) When the observations consist of averages, say  $\bar{y}_i = \frac{1}{m_i} \sum_{j=1}^{m_i} y_{ij}$ ,  $i = 1, \dots, n$ , then the covariance matrix  $\sigma^2 V$  of the response vector is a diagonal matrix matrix and the  $i$ th diagonal element of  $V$  is  $1/m_i$ . Here the elements  $V$  are all known, and this problem fits into the weighted least squares framework.

# Weighted Least Squares

KNOWN

We want characterize the least squares estimates of  $(\beta, \sigma^2)$ . Since  $V$  is positive definite, we can write  $V$  as  $V = QQ'$  for some nonsingular matrix  $Q$ . It follows that  $Q^{-1}VQ'^{-1} = I$ .

Now instead of working with (8) we can equivalently work with

$$Q^{-1}Y = Q^{-1}X\beta + Q^{-1}\epsilon. \quad (9)$$

From (9), we notice that  $E(Q^{-1}\epsilon) = Q^{-1}E(\epsilon) = Q^{-1}0 = 0$ , and

$$\text{Cov}(Q^{-1}\epsilon) = Q^{-1}(\sigma^2 V)Q'^{-1} = \sigma^2 Q^{-1}VQ'^{-1} = \sigma^2 I.$$

This transformation leads us back to the usual linear model. Thus a least squares estimate of  $\beta$  is a minimizer of

$$\begin{aligned} & (Q^{-1}Y - Q^{-1}X\beta)'(Q^{-1}Y - Q^{-1}X\beta) \\ = & (Y - X\beta)'(Q'^{-1}Q^{-1})(Y - X\beta) \\ = & (Y - X\beta)'V^{-1}(Y - X\beta). \end{aligned}$$

We now present a theorem characterizing estimability of  $\lambda'\beta$ .

# Weighted Least Squares

## Theorem

- a)  $\lambda'\beta$  is estimable in (8) if and only if  $\lambda'\beta$  is estimable in (9).
- b)  $\hat{\beta}$  is a weighted least squares estimate of  $\beta$  if and only if

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

- c) For any estimable function  $\lambda'\beta$ ,  $\lambda' = \rho'X$ , the unique weighted least squares estimate of  $\lambda'\beta$  is  $\rho'AY$ , where  $A = X(X'V^{-1}X)^{-}X'V^{-1}$ . The unique weighted least least squares estimate of  $\mu = X\beta$  is  $AY$ .
- d) For any estimable function  $\lambda'\beta$ ,  $\lambda' = \rho'X$ ,  $\rho'AY$  is the BLUE of  $\lambda'\beta$ , where  $A$  is the matrix in part c).

# Weighted Least Squares

## Properties of Weighted Least Squares Estimates

### Theorem

Let  $A = X(X'V^{-1}X)^{-1}X'V^{-1}$ . Then

- a)  $A$  is invariant with respect to the choice of generalized inverse.
- b)  $A$  is a projection operator onto  $C(X)$  along  $\mathcal{N}(A)$ .

### Proof

- a) Let  $B = V^{-1/2}X$ . Then we can write

$$\begin{aligned} A &= X(X'V^{-1}X)^{-1}X'V^{-1} \\ &= V^{1/2}(V^{-1/2}X)(X'V^{-1/2}V^{-1/2}X)^{-1}(X'V^{-1/2})V^{-1/2} \\ &= V^{1/2}B(B'B)^{-1}B'V^{-1/2}. \end{aligned}$$

We know from previous results that  $B(B'B)^{-1}B'$  is the orthogonal projection operator onto  $C(B)$  and is invariant with respect to the choice of generalized inverse. Since  $V$  is nonsingular, it follows that  $A$  is invariant with respect to the choice of generalized inverse.

What does non-singularity have to do with anything? Existence of  $V^{-1/2}$ , or rank-preserving? What does rank preservation have to do with invariant preservation?

# Weighted Least Squares

b) We have  $V = QQ'$ , where  $Q$  is nonsingular. Consider the orthogonal projection operator onto  $C(Q^{-1}X)$ , denoted by  $P$ , which is given by

$$\begin{aligned} P &= (Q^{-1}X) \left( (Q^{-1}X)'(Q^{-1}X) \right)^{-1} (Q^{-1}X)' \\ &= (Q^{-1}X)(X'(QQ')^{-1}X)^{-1} X' Q'^{-1} \\ &= Q^{-1}X(X'V^{-1}X)^{-1} X' Q'^{-1}. \end{aligned}$$

By the definition of a projection operator, we have

$$\begin{aligned} PQ^{-1}X &= Q^{-1}X \\ \Leftrightarrow Q^{-1}X(X'V^{-1}X)^{-1}X'(QQ')^{-1}X &= Q^{-1}X \\ \Leftrightarrow Q^{-1}X(X'V^{-1}X)^{-1}X'V^{-1}X &= Q^{-1}X \\ \Leftrightarrow Q^{-1}AX &= Q^{-1}X \\ \Leftrightarrow AX &= X. \quad \text{this last line sufficient to show } AZ=Z \text{ for } Z \in C(X) \end{aligned}$$

It is now obvious by the definition of a projection operator that  $A$  is a projection operator onto  $C(X)$ .

# Weighted Least Squares

To formally finish the proof, let  $v \in C(X)$ , and write  $X = (x_1, \dots, x_n)$ , where  $x_j$  is the  $j$ th column of  $X$ . Then  $v = \alpha_1 x_1 + \dots + \alpha_n x_n$ , where the  $\alpha_j$ 's are scalars. Then

$$\begin{aligned}Av &= A(\alpha_1 x_1 + \dots + \alpha_n x_n) \\&= \alpha_1 A x_1 + \dots + \alpha_n A x_n \\&= \alpha_1 x_1 + \dots + \alpha_n x_n \quad \text{from b) above} \\&= v.\end{aligned}$$

Thus for any  $v \in C(X)$ ,  $Av = v$ , and so  $A$  is a projection operator onto  $C(X)$ .

# Weighted Least Squares

## Remarks:

- 1) We can see that  $A$  is not an orthogonal projection operator with respect to the inner product  $x'y$ . This is easily seen by noting that  $A$  is NOT symmetric.
- 2) If one defines the inner product between two vectors  $x$  and  $y$  as  $x'V^{-1}y$ , then  $A$  is an orthogonal projection operator with respect to this inner product. In this inner product,  $x$  and  $y$  are orthogonal if  $x'V^{-1}y = 0$ , where  $x \in C(X)$  and  $y \in C(X)^\perp$ .

To see that  $A$  is an orthogonal projection operator with respect to the  $x'V^{-1}y$  inner product, we note that for any  $x \in C(X)$ , and  $y \in C(X)^\perp$ ,

$$Ay = X(X'V^{-1}X)^{-1}X'V^{-1}y = X(X'V^{-1}X)^{-1}0 = 0$$

- 3) The weighted least squares estimator equals the ordinary least squares estimator under certain conditions. This is given in the next two theorems.

# Weighted Least Squares

## Theorem

Suppose  $X$  is  $n \times p$  of rank  $r$ , and  $V$  is a positive definite matrix. Then  $C(V^{-1}X) = C(X)$  if and only if  $C(VX) = C(X)$ .

## Proof

Exercise.

## Theorem

Consider the linear model

$$Y = X\beta + \epsilon,$$

where  $E(\epsilon) = 0$  and  $\text{Cov}(\epsilon) = \sigma^2 V$ , where  $V$  is a known positive definite matrix and  $X$  has rank  $r$ . Consider estimating  $\rho' X\beta$ . Then

$$\rho' AY = \rho' MY$$

if and only if  $C(VX) = C(X)$ , where  $M = X(X'X)^{-1}X'$  and  $A = X(X'V^{-1}X)^{-1}X'V^{-1}$ .

The proof is left as an exercise.

This theorem says that the weighted least squares estimate equals the ordinary least squares estimate if and only if  $C(VX) = C(X)$ .

# Weighted Least Squares

## Corollary

Suppose  $X$  has full rank  $p$ . Then the weighted least squares estimate of  $\beta$  is given by  $\tilde{\beta} = (X'V^{-1}X)^{-1}X'V^{-1}Y$ , and the ordinary least squares estimate of  $\beta$  is given by  $\hat{\beta} = (X'X)^{-1}X'Y$ . Then  $\tilde{\beta} = \hat{\beta}$  if and only if  $C(VX) = C(X)$ .

## Proof

1)  $\Rightarrow$

Suppose  $\tilde{\beta} = \hat{\beta}$ . Then

$$\begin{aligned}(X'V^{-1}X)^{-1}X'V^{-1}Y &= (X'X)^{-1}X'Y \\ \Leftrightarrow (X'V^{-1}X)^{-1}X'V^{-1} &= (X'X)^{-1}X' \\ \Leftrightarrow V^{-1}X(X'V^{-1}X)^{-1} &= X(X'X)^{-1}\end{aligned}$$

Now let  $T_1 = (X'V^{-1}X)^{-1}$ , and  $T_2 = (X'X)^{-1}$ . The columns of  $T_1$  and  $T_2$  are bases for  $R^p$ . Now we have  $V^{-1}XT_1 = XT_2$  which implies  $V^{-1}X = XT_2T_1^{-1}$ . Let  $T_3 = T_2T_1^{-1}$ . Then the columns of  $T_3$  are a basis for  $R^p$  and  $V^{-1}X = XT_3$ , which implies  $X = VXT_3$ . Since  $T_3$  is nonsingular,  $C(VXT_3) = C(VX)$ , and thus  $C(X) = C(VX)$ .

2) The proof of  $\Leftarrow$  is left as an exercise.

# Weighted Least Squares

## Estimation of $\sigma^2$

For the model in (8), the residual vector is given by  $(I - A)Y$ . Thus the estimate of  $\sigma^2$  is some function of  $(I - A)Y$ . We search for an unbiased estimate. If we use the transformed model in (9), it is easy to find such an estimate.

Recall that the transformed model is given by  $Q^{-1}Y = Q^{-1}X\beta + Q^{-1}\epsilon$ . Let  $M^*$  denote the orthogonal projection operator onto  $C(Q^{-1}X)$ . Thus  $M^* = (Q^{-1}X)(X'V^{-1}X)^{-1}X'Q'^{-1}$ .

An unbiased estimator of  $\sigma^2$  is given by

$$\hat{\sigma}^2 = \frac{\|(I - M^*)Q^{-1}Y\|^2}{n - r} .$$

We note that

$$\begin{aligned}(I - M^*)Q^{-1} &= Q^{-1} - Q^{-1}X(X'V^{-1}X)^{-1}X'Q'^{-1}Q^{-1} \\ &= Q^{-1} - Q^{-1}X(X'V^{-1}X)^{-1}X'V^{-1} \\ &= Q^{-1} - Q^{-1}A \\ &= Q^{-1}(I - A) .\end{aligned}$$

# Weighted Least Squares

Thus  $\|(I - M^*)Q^{-1}Y\|^2 = \|Q^{-1}(I - A)Y\|^2$ , and therefore

$$\begin{aligned}\hat{\sigma}^2 &= \frac{\|Q^{-1}(I - A)Y\|^2}{n - r} \\ &= \frac{Y'(I - A)'(QQ')^{-1}(I - A)Y}{n - r} \\ &= \frac{Y'(I - A)'V^{-1}(I - A)Y}{n - r}.\end{aligned}$$

# Weighted Least Squares

## Theorem

$$V^{-1}(I - A) = (I - A)' V^{-1}(I - A).$$

## Proof

We have

$$(I - A)' V^{-1}(I - A) = V^{-1}(I - A) - A' V^{-1}(I - A).$$

The proof will be completed if we can show  $A' V^{-1}(I - A) = 0$ , since this will imply  $A' V^{-1} = A' V^{-1}A$ . Now

$$\begin{aligned} A' V^{-1} A &= \left[ V^{-1} X (X' V^{-1} X)^{-1} X' \right] V^{-1} \\ &\times \left[ X (X' V^{-1} X)^{-1} X' V^{-1} \right] \\ &= V^{-1} X (X' V^{-1} X)^{-1} (X' V^{-1} X) \\ &\times (X' V^{-1} X)^{-1} X' V^{-1} \\ &= V^{-1} X (X' V^{-1} X)^{-1} X' V^{-1} \\ &= A' V^{-1}. \end{aligned}$$

## Theorem

$$AVA = AV = VA' \quad (\text{The proof is similar to the one given above.})$$

# Weighted Least Squares

## Covariance Matrices of Estimates

- a) The BLUE of  $\rho'X\beta$  in the usual linear model (i.e., when  $\text{Cov}(\epsilon) = \sigma^2 I$ ) is  $\rho'MY$ .

$$\begin{aligned}\text{Cov}(\rho'MY) &= (\rho'M)\text{Cov}(Y)(\rho'M)' \\ &= (\rho'M)(\sigma^2 I)(\rho'M)' = \sigma^2 \rho'M\rho.\end{aligned}$$

- b) The covariance matrix of the residual vector for the usual linear model is

$$\begin{aligned}\text{Cov}((I - M)Y) &= (I - M)(\sigma^2 I)(I - M)' \\ &= \sigma^2(I - M).\end{aligned}$$

Note that the covariance matrix of the residuals is singular. It is an  $n \times n$  positive semidefinite matrix of rank  $n - r$ .

- c) For weighted least squares,

$$\begin{aligned}\text{Cov}(\rho'AY) &= (\rho'A)(\sigma^2 V)(\rho'A)' \\ &= \sigma^2 \rho' A V A' \rho.\end{aligned}$$

Also,

$$\text{Cov}((I - A)Y) = \sigma^2(I - A)V(I - A)'.$$

# Weighted Least Squares

- d) For the usual linear model with an intercept, the residuals sum to 0. To see this, let  $J = (1, \dots, 1)'$  denote the  $n \times 1$  vector of ones. Then

$$\begin{aligned} J'(I - M)Y &= J'Y - J'MY \\ &= J'Y - J'Y = 0 . \end{aligned}$$

We note that since  $J \in C(X)$ ,  $J \in C(M)$  since  $C(X) = C(M)$ , so  $MJ = J$ , which implies  $(MJ)' = J'M = J'$ .

We have a similar result for weighted least squares. That is

$$\begin{aligned} J'(I - A)Y &= J'Y - J'AY \\ &= J'Y - J'Y = 0 . \end{aligned}$$

## Major Note

We have NOT made ANY distributional assumptions on  $\epsilon$  to obtain least squares estimates, weighted least squares estimates, or BLUE's.

We need distributional assumptions to construct tests of hypotheses, confidence regions, and prediction regions.

# Distribution Theory

## Review of Distribution Theory

### Chi-square distribution

A random variable  $X$  is said to have a central chi-square distribution with  $n$  degrees of freedom, written  $X \sim \chi^2(n)$ , if  $X$  has density

$$f(x) = \frac{1}{\Gamma(n/2)} \left(\frac{1}{2}\right)^{n/2} x^{n/2-1} e^{-x/2}.$$

The moment generating function (MGF) of a random variable  $X$ , denoted  $\psi_X(t)$  is defined as

$$\psi_X(t) = E(e^{tX}) = \int_{-\infty}^{\infty} e^{tx} f(x) dx.$$

The integral is replaced by a sum if  $X$  is discrete.

### Theorem

If  $X \sim \chi^2(n)$ , then  $\psi_X(t) = (1 - 2t)^{-n/2}$ .

The proof is left as an exercise.

# Distribution Theory

## Normal distribution

A random variable  $X$  is said to have a normal distribution with mean  $\mu$  and variance  $\sigma^2$ , written  $X \sim N(\mu, \sigma^2)$  if  $X$  has density

$$f(x) = (2\pi)^{-1/2}\sigma^{-1} \exp\left\{\frac{-1}{2\sigma^2}(x - \mu)^2\right\}.$$

If  $\mu = 0$  and  $\sigma = 1$ , then  $X \sim N(0, 1)$  and we say that  $X$  has a standard normal distribution.

## Theorem

If  $X \sim N(\mu, \sigma^2)$ , then  $\psi_X(t) = \exp\{t\mu + \frac{1}{2}t^2\sigma^2\}$ .

## Proof

To prove this result we need to know how to complete the square. Recall that

$$\begin{aligned} ax^2 + bx &= a\left(x^2 + \frac{bx}{a}\right) \\ &= a\left(x + \frac{b}{2a}\right)^2 - \frac{b^2}{4a}. \end{aligned}$$

# Distribution Theory

Now

$$\begin{aligned}\psi_X(t) &= \int_{-\infty}^{\infty} \exp\{tx\} (2\pi)^{-1/2} \sigma^{-1} \exp\left\{\frac{-1}{2\sigma^2}(x - \mu)^2\right\} dx \\ &= \int_{-\infty}^{\infty} (2\pi)^{-1/2} \sigma^{-1} \exp\{tx\} \\ &\quad \times \exp\left\{\frac{-1}{2\sigma^2}(x^2 - 2\mu x + \mu^2)\right\} dx \\ &= \exp\left\{\frac{-\mu^2}{2\sigma^2}\right\} \\ &\quad \times \int_{-\infty}^{\infty} (2\pi)^{-1/2} \sigma^{-1} \exp\left\{\frac{-1}{2\sigma^2}(x^2 - 2x(\mu + \sigma^2 t))\right\} dx \\ &= \exp\left\{\frac{-\mu^2}{2\sigma^2}\right\} \exp\left\{\frac{1}{2\sigma^2}(\mu + \sigma^2 t)^2\right\} \\ &\quad \times \int_{-\infty}^{\infty} (2\pi)^{-1/2} \sigma^{-1} \exp\left\{\frac{-1}{2\sigma^2}(x - (\mu + \sigma^2 t))^2\right\} dx \\ &= \exp\left\{\frac{-\mu^2}{2\sigma^2}\right\} \exp\left\{\frac{1}{2\sigma^2}(\mu + \sigma^2 t)^2\right\} \\ &= \exp\left\{\mu t + \frac{1}{2}\sigma^2 t^2\right\}.\end{aligned}$$

# Distribution Theory

## Theorem

Suppose  $Z_1, \dots, Z_n$  are independently identically distributed (*i.i.d.*)  $N(0, 1)$  random variables. Define

$$X = \sum_{i=1}^n Z_i^2 .$$

Then  $X \sim \chi^2(n)$ .

# Distribution Theory

## Noncentral chi-square distribution

A random variable  $X$  is said to have a noncentral chi-square distribution with  $n$  degrees of freedom and noncentrality parameter  $\gamma$ , written  $X \sim \chi^2(n, \gamma)$ , if  $X$  has density

$$f(x) = \sum_{i=0}^{\infty} \left( \frac{\gamma^i e^{-\gamma}}{i!} \right) \frac{\left(\frac{1}{2}\right)^{\frac{2i+n}{2}} x^{\frac{2i+n}{2}-1}}{\Gamma((2i+n)/2)} \exp\{-x/2\}.$$

We see that a noncentral chi-square density is a infinite Poisson mixture of central chi-square densities. Noncentral chi-square distributions arise in hypothesis testing situations in linear models when one is interested in finding the distribution of the test statistic under the alternative hypothesis.

## Theorem

Suppose  $Y_1, \dots, Y_n$  are independent and  $Y_i \sim N(\mu_i, \sigma^2)$ ,  $i = 1, \dots, n$ . Define

$$X = \frac{1}{\sigma^2} \sum_{i=1}^n Y_i^2 .$$

Then  $X \sim \chi^2(n, \gamma)$ , where  $\gamma = \frac{1}{2\sigma^2} \sum_{i=1}^n \mu_i^2$ .

# Distribution Theory

## Properties of noncentral chi-square

- 1) If  $X \sim \chi^2(n, \gamma)$ , then

$$\psi_X(t) = (1 - 2t)^{-n/2} \exp\left\{\frac{2\gamma t}{1 - 2t}\right\}.$$

This can be proved using the definition of the noncentral chi-square density above and interchanging the order of integration and summation.

- 2) If  $X \sim \chi^2(n, \gamma)$ , then

$$E(X) = n + 2\gamma$$

and

$$\text{Var}(X) = 2n + 8\gamma.$$

This can be proved using the MGF in 1).

- 3) If  $X \sim \chi^2(n, \gamma)$  and  $\gamma = 0$ , then this corresponds to a central chi-square random variable with  $n$  degrees of freedom. That is,  
 $X \sim \chi^2(n, 0) = \chi^2(n)$ .

# Distribution Theory

## The $t$ distribution

Suppose  $X \sim N(0, 1)$ ,  $Y \sim \chi^2(n)$ , and  $X$  and  $Y$  are independent. Define the random variable

$$T = \frac{X}{\sqrt{Y/n}} .$$

Then  $T$  is said to have a  $t$  distribution with  $n$  degrees of freedom. We write  $T \sim t(n)$ .

## Noncentral $t$ distribution

Suppose  $X \sim N(\mu, 1)$  and  $Y \sim \chi^2(n)$ , and  $X$  and  $Y$  are independent. Define the random variable

$$W = \frac{X}{\sqrt{Y/n}} .$$

Then  $W$  is said to have a noncentral  $t$  distribution with  $n$  degrees of freedom and noncentrality parameter  $\mu$ . We write  $W \sim t(n, \mu)$ . If  $\mu = 0$ , then  $W$  reduces to a central  $t$  distribution with  $n$  degrees of freedom.

# Distribution Theory

## The $F$ distribution

Suppose  $X_1 \sim \chi^2(n_1, \gamma_1)$  and  $X_2 \sim \chi^2(n_2, \gamma_2)$ , and  $X_1$  and  $X_2$  are independent. Define the random variable

$$F = \frac{X_1/n_1}{X_2/n_2}.$$

Then  $F$  is said to have a doubly noncentral  $F$  distribution with  $(n_1, n_2)$  degrees of freedom and noncentrality parameters  $(\gamma_1, \gamma_2)$ . We write  $F \sim F(n_1, n_2, \gamma_1, \gamma_2)$ .

- a) If  $\gamma_2 = 0$ , then  $F$  is said to have a noncentral  $F$  distribution. We represent this as  $F \sim F(n_1, n_2, \gamma_1)$ .
- b) If  $\gamma_1 = 0$  and  $\gamma_2 = 0$ , then  $F$  is said to have a central  $F$  distribution. We represent this as  $F \sim (n_1, n_2)$ .

The central  $F$  distribution arises in hypothesis tests of nested linear models. In this setting, the distribution of the test statistic under the null hypothesis often has a central  $F$  distribution. Noncentral  $F$  distributions arise from distributions of the test statistic under the alternative hypothesis. The distribution of the test statistic under the alternative hypothesis is important for power calculations.

# Distribution Theory

## Properties of $F$ distribution

If  $F \sim F(n_1, n_2, \gamma)$ , then

a)  $E(F) = \frac{n_2(n_1+2\gamma)}{n_1(n_2-2)}$  ,  $n_2 > 2$

b)

$$\text{Var}(F) = 2 \left( \frac{n_2}{n_1} \right)^2 \frac{(n_1 + 2\gamma)^2 + (n_1 + 4\gamma)(n_2 - 2)}{(n_2 - 2)^2(n_2 - 4)}, \quad n_2 > 4 .$$

The formulas for the mean and variance for a central  $F$  are obtained by setting  $\gamma = 0$  in a) and b) above.

# Distribution Theory

## Multivariate Generating Functions

Suppose  $X = (X_1, \dots, X_n)'$  is an  $n \times 1$  random vector with  $n$  dimensional density  $f(x_1, \dots, x_n)$ . The multivariate moment generating function of  $X$  is defined as

$$\begin{aligned}\psi_{X_1, \dots, X_n}(t_1, \dots, t_n) &= E(e^{t_1 X_1 + \dots + t_n X_n}) \\ &= \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} e^{t_1 X_1 + \dots + t_n X_n} f(x_1, \dots, x_n) dx_1 \dots dx_n\end{aligned}$$

We can represent the multivariate MGF in vector notation by letting  $t = (t_1, \dots, t_n)'$  and  $X = (X_1, \dots, X_n)'$ . Then

$$\psi_X(t) = E(e^{t' X}).$$

## Properties of the multivariate MGF

The properties of the multivariate MGF are similar to the univariate MGF.

Again let  $X = (X_1, \dots, X_n)'$  and  $t = (t_1, \dots, t_n)'$ .

- 1)  $\psi_X(0) = 1$ .
- 2) If  $X_1, \dots, X_n$  are independent, then

$$\psi_X(t) = \prod_{i=1}^n \psi_{X_i}(t_i)$$

where  $\psi_{X_i}(t_i)$  is the univariate MGF of  $X_i$ .

# Distribution Theory

- 3) Moments can be obtained by differentiating the multivariate MGF.

$$\frac{\partial^{k_1+\dots+k_n}}{\partial t_1^{k_1} \dots \partial t_n^{k_n}} \psi_X(t_1, \dots, t_n) |_{t_1=\dots=t_n=0} = E(X_1^{k_1} \dots X_n^{k_n}).$$

For example suppose  $n = 2$  so that  $X = (X_1, X_2)'$ , and  $\psi_X(t_1, t_2) = E(e^{t_1 X_1 + t_2 X_2})$ . We have

$$\frac{\partial^5}{\partial t_1^2 \partial t_2^3} \psi_X(t_1, t_2) |_{t_1=t_2=0} = E(X_1^2 X_2^3).$$

- 4) The MGF for any marginal distribution of  $X$  is obtained by setting equal to 0 those  $t_j$ 's that correspond to the  $X_j$ 's not in the marginal distribution. For example, suppose  $n = 4$ , so that  $X = (X_1, X_2, X_3, X_4)'$  and  $\psi_X(t_1, t_2, t_3, t_4)$  is the multivariate MGF of  $X$ . Then  $\psi_{X_1, X_3}(t_1, t_3) = \psi_X(t_1, 0, t_3, 0)$ ,  $\psi_{X_1}(t_1) = \psi_X(t_1, 0, 0, 0)$  and so on.
- 5) The multivariate characteristic function is defined as

$$\phi_X(t) = E(e^{it'X})$$

where  $i = \sqrt{-1}$ . The characteristic function always exists for any random variable or vector, but the MGF may NOT exist for some random variables. Thus the characteristic function is a bit more useful than the MGF in proving certain results.

Assume  
Independence?

# Distribution Theory

For example, consider the Cauchy distribution with median 0. The density for this Cauchy distribution is

$$f(x) = \frac{1}{\pi(1+x^2)}, \quad -\infty < x < \infty.$$

For the Cauchy distribution, the MGF does NOT exist, but the characteristic function is given by

$$\phi_X(t) = \exp\{-|t|\}.$$

The relationship between the characteristic function and the MGF is given by

$$\psi_X(it_1, \dots, it_n) = \phi_X(t_1, \dots, t_n).$$

# Distribution Theory

## Multivariate Normal Distribution

Perhaps the most widely used distribution in statistics is the multivariate normal distribution. A lot of estimation and hypothesis testing results in linear models are derived assuming a multivariate normal distribution for  $\epsilon$ .

### Defn

Suppose  $Z_1, \dots, Z_n$  are *i.i.d.*  $N(0, 1)$  random variables. Let  $Z = (Z_1, \dots, Z_n)'$ . We have  $E(Z) = 0$  and  $\text{Cov}(Z) = I$ . We say that  $Y$  has an  $r$  dimensional multivariate normal distribution if  $Y$  has the same distribution as  $AZ + b$  for some  $r \times n$  matrix of constants  $A$  and an  $r \times 1$  vector of constants  $b$ . We denote the distribution of  $Y$  by

$$Y \sim N_r(b, AA')$$

Note that  $E(Y) = E(AZ + b) = AE(Z) + b = A0 + b = b$ , and  $\text{Cov}(Y) = \text{Cov}(AZ + b) = AC\text{Cov}(Z)A' = AIA' = AA'$ . Thus the notation above indicates that  $b$  is the mean vector of  $Y$  and  $AA'$  is the covariance matrix of  $Y$ .

Thus, when we write  $Y \sim N_n(\mu, \Sigma)$ , this means that  $Y$  has an  $n$  dimensional multivariate normal distribution with mean vector  $\mu$  and covariance matrix  $\Sigma$ . We will abbreviate multivariate normal by MVN.

# Distribution Theory

Recall that  $\Sigma$  must be positive semidefinite since any covariance matrix must be positive semidefinite. If  $\Sigma$  is singular, then the MVN distribution is said to be singular normal. In these cases, the density does not exist. The density of a MVN distribution exists only when  $\Sigma$  is positive definite.

## Defn

Suppose  $X = (X_1, \dots, X_n)'$ . Then  $X$  is said to have an  $n$  dimensional multivariate normal distribution with mean  $\mu$  and covariance matrix  $\Sigma$  if  $X$  has density

$$f(x) = (2\pi)^{-n/2} |\Sigma|^{-1/2} \exp \left\{ \frac{-1}{2} (x - \mu)' \Sigma^{-1} (x - \mu) \right\}.$$

We note that this definition requires  $\Sigma$  to be positive definite. Thus, this definition is a bit less general than the definition of MVN given above.

# Distribution Theory

Now we give two useful results for quadratic forms.

## Square completion in $n$ dimensions

Suppose  $x = (x_1, \dots, x_n)'$  is an  $n \times 1$  vector,  $A$  is an  $n \times n$  nonsingular matrix, and  $b$  is an  $n \times 1$  vector. Then

$$x'Ax + b'x = \left(x + \frac{A^{-1}b}{2}\right)'A\left(x + \frac{A^{-1}b}{2}\right) - \frac{b'A^{-1}b}{4}$$

This is the multivariate analog of the one dimensional square completion given earlier. This result is very useful in computing multivariate normal integrals as we will see shortly.

## Combining quadratic forms

Suppose  $x$  is an  $n \times 1$  vector,  $\mu_1$ , and  $\mu_2$  are  $n \times 1$  vectors,  $A_1$ ,  $A_2$  are  $n \times n$  matrices such that  $A_1 + A_2$  is nonsingular. Then

$$\begin{aligned} & (x - \mu_1)'A_1(x - \mu_1) + (x - \mu_2)'A_2(x - \mu_2) \\ = & (x - \mu^*)'(A_1 + A_2)(x - \mu^*) + \mu_1'A_1\mu_1 + \mu_2'A_2\mu_2 \\ - & \mu^*(A_1 + A_2)\mu^*, \end{aligned}$$

where

$$\mu^* = (A_1 + A_2)^{-1}(A_1\mu_1 + A_2\mu_2).$$

# Distribution Theory

We can generalize the above result to combining  $m$  quadratic forms. We have

$$\begin{aligned} & (x - \mu_1)' A_1 (x - \mu_1) + \dots + (x - \mu_m)' A_m (x - \mu_m) \\ = & (x - \mu^*)' B (x - \mu^*) + \sum_{i=1}^m \mu_i' A_i \mu_i - \mu'^* B \mu^* \end{aligned}$$

where  $B = \sum_{i=1}^m A_i$  and  $\mu^* = B^{-1} (\sum_{i=1}^m A_i \mu_i)$ .

## Properties of MVN distributions

1) Suppose  $X \sim N_n(\mu, \Sigma)$ , then

$$\psi_X(t) = \exp\left\{t'\mu + \frac{1}{2}t'\Sigma t\right\}.$$

We note here that the MGF does not require the inversion of  $\Sigma$ . Thus the MGF of MVN always exists and is given above even if  $\Sigma$  is singular. Thus the MGF of singular MVN distributions exists but the density does not.

# Distribution Theory

## Proof of 1)

Assume  $\Sigma$  is positive definite. Then

$$\begin{aligned}\psi_X(t) &= E(e^{t'X}) \\ &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} (2\pi)^{-n/2} |\Sigma|^{-1/2} \exp\left\{t'x\right\} \\ &\quad \times \exp\left\{\frac{-1}{2}(x - \mu)' \Sigma^{-1} (x - \mu)\right\} dx_1 \dots dx_n \\ &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} (2\pi)^{-n/2} |\Sigma|^{-1/2} \exp\{t'x\} \\ &\quad \times \exp\left\{\frac{-1}{2}(x' \Sigma^{-1} x - 2x' \Sigma^{-1} \mu + \mu' \Sigma^{-1} \mu)\right\} dx_1 \dots dx_n \\ &= \exp\left\{\frac{-1}{2}\mu' \Sigma^{-1} \mu\right\} \\ &\quad \times \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} (2\pi)^{-n/2} |\Sigma|^{-1/2} \\ &\quad \times \exp\left\{\frac{-1}{2}(x' \Sigma^{-1} x - 2x' (\Sigma^{-1} \mu + t))\right\} dx_1 \dots dx_n\end{aligned}$$

# Distribution Theory

$$\begin{aligned}&= \exp\left\{\frac{-1}{2}\mu'\Sigma^{-1}\mu\right\} \exp\left\{\frac{1}{2}(\mu + \Sigma t)'\Sigma^{-1}(\mu + \Sigma t)\right\} \\&\times \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} (2\pi)^{-n/2} |\Sigma|^{-1/2} \\&\times \exp\left\{\frac{-1}{2}(x - (\mu + \Sigma t))'\Sigma^{-1}(x - (\mu + \Sigma t))\right\} dx_1 \dots dx_n \\&= \exp\left\{\frac{-1}{2}\mu'\Sigma^{-1}\mu\right\} \exp\left\{\frac{1}{2}(\mu + \Sigma t)'\Sigma^{-1}(\mu + \Sigma t)\right\} \\&= \exp\left\{t'\mu + \frac{1}{2}t'\Sigma t\right\}.\end{aligned}$$

This completes the proof. The characteristic function is given by

$$\phi_X(t) = \psi_X(it) = \exp\left\{it'\mu - \frac{1}{2}t'\Sigma t\right\}.$$

# Distribution Theory

## Properties of MVN distributions (cont'd)

2) A linear transformation of MVN's is MVN. Suppose  $X \sim N_n(\mu, \Sigma)$ , and define  $Y = AX + b$ , where  $A$  is an  $r \times n$  matrix of constants and  $b$  is an  $r \times 1$  vector of constants. Then

$$Y \sim N_r(A\mu + b, A\Sigma A')$$

This result can easily be proved using the MGF. We have

$$\begin{aligned}\psi_Y(t) &= E(e^{t' Y}) \\ &= E(e^{t' (AX+b)}) \\ &= e^{t' b} E(e^{At' X}) \\ &= e^{t' b} E(e^{(A't)' X}) \\ &= e^{t' b} \psi_X(A't) = e^{t' b} e^{(A't)' \mu + \frac{1}{2} ((A't)' \Sigma (A't))} \\ &= e^{t' (A\mu + b) + \frac{1}{2} t' A \Sigma A' t}.\end{aligned}$$

We can now recognize the MGF above as the MGF of a MVN distribution with mean  $A\mu + b$  and covariance matrix  $A\Sigma A'$ .

# Distribution Theory

3) A linear combination of independent MVN's is MVN. Suppose  $X_1, \dots, X_k$  are independent and each  $X_i \sim N_n(\mu_i, \Sigma_i)$ ,  $i = 1, \dots, k$ . Suppose  $a_1, \dots, a_k$  are scalars and define

$$Y = a_1 X_1 + \dots + a_k X_k .$$

Then

$$Y \sim N_n(\mu^*, \Sigma^*)$$

where  $\mu^* = \sum_{i=1}^k a_i \mu_i$  and  $\Sigma^* = \sum_{i=1}^k a_i^2 \Sigma_i$ . This again can be proved using MGF's.

4) Marginal distributions of MVN are MVN. Suppose  $X \sim N_n(\mu, \Sigma)$ . Partition

$X$  into  $X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$  where  $X_1$  is  $r \times 1$  and  $X_2$  is  $(n - r) \times 1$ . Partition  $\mu$  as

$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$  where  $\mu_1$  is  $r \times 1$  and  $\mu_2$  is  $(n - r) \times 1$ . Similarly partition  $\Sigma$  as

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} ,$$

where  $\Sigma_{11}$  is  $r \times r$ ,  $\Sigma_{12}$  is  $r \times (n - r)$ ,  $\Sigma_{21} = \Sigma'_{12}$  is  $(n - r) \times r$ , and  $\Sigma_{22}$  is  $(n - r) \times (n - r)$ .

# Distribution Theory

The marginal distribution of  $X_1$  is given by

$$X_1 \sim N_r(\mu_1, \Sigma_{11}) .$$

This can be proved using the MGF by putting in zeroes to the  $t_j$ 's corresponding to  $X_2$ .

5) Conditional distributions of MVN are MVN. Suppose  $X \sim N_n(\mu, \Sigma)$ . Using the partition in 4), we have

$$X_1 | X_2 = x_2 \sim N_r(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2), \Sigma_{11.2})$$

where  $\Sigma_{11.2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$ .

Using MGF's to prove things about conditional distributions is usually hard since we do not have nice results for MGF's for conditional distributions. One usually has to rely on the density to prove results concerning conditional distributions.

## Exercise

Suppose  $X = (X_1, X_2, X_3)'$  is a  $3 \times 1$  random vector and  $X \sim N_3(\mu, \Sigma)$ . Derive the conditional distribution of  $(X_1, X_3) | X_2 = x_2$ .

# Distribution Theory

## Connection of conditional distributions with linear regression

Consider the usual linear model

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p + \epsilon .$$

Let  $X = (X_1, \dots, X_p)'$ , and suppose

$$\begin{pmatrix} Y \\ X \end{pmatrix} \sim N_{p+1}(\mu, \Sigma) ,$$

where

$$\mu = \begin{pmatrix} E(Y) \\ E(X_1) \\ \vdots \\ E(X_p) \end{pmatrix} = \begin{pmatrix} \mu_y \\ \mu_1 \\ \vdots \\ \mu_p \end{pmatrix}$$

and

$$\Sigma = \begin{pmatrix} \sigma_y^2 & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

where  $\sigma_y^2 = \text{Var}(Y)$ ,  $\Sigma_{22} = \text{Cov}(X)$ , and  $\Sigma_{12}$  is a  $1 \times p$  vector consisting of  $\text{Cov}(Y, X)$ .

# Distribution Theory

Let  $\mu_x = E(X) = (\mu_1, \dots, \mu_p)'$ . By property 5), we know that

$$Y | X = x \sim N_1(\mu_y + \Sigma_{12}\Sigma_{22}^{-1}(x - \mu_x), \sigma_y^2 - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}) .$$

Thus

$$\begin{aligned} E(Y | X = x) &= \mu_y + \Sigma_{12}\Sigma_{22}^{-1}(x - \mu_x) \\ &= (\mu_y - \Sigma_{12}\Sigma_{22}^{-1}\mu_x) + \Sigma_{12}\Sigma_{22}^{-1}x \end{aligned}$$

Now make the transformation

$$\begin{aligned} \beta_0 &= \mu_y - \Sigma_{12}\Sigma_{22}^{-1}\mu_x , \\ \beta' &= \Sigma_{12}\Sigma_{22}^{-1} , \end{aligned}$$

and

$$\sigma^2 = \sigma_y^2 - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} ,$$

where  $\beta' = (\beta_1, \dots, \beta_p)$ . It can be shown that this transformation is one-to-one. The transformation implies

$$\begin{aligned} E(Y | X = x) &= \beta_0 + \beta'x \\ &= \beta_0 + \beta_1x_1 + \dots + \beta_px_p . \end{aligned}$$

This result tells us that if we assume a multivariate normal distribution on the response  $Y$  and the (random) regressors  $X$ , then the regression function is the conditional expectation of  $Y | X = x$ .

# Distribution Theory

## Theorem

If  $X \sim N_n(\mu, \Sigma)$ , then all marginals, conditionals, and linear combinations of the components of  $X$  are MVN.

Note: The converse of the above theorem is NOT true. For example, if all of the marginals are MVN, this does NOT imply that the joint distribution is MVN.

## Example

Suppose  $(X_1, X_2)$  have joint density

$$\begin{aligned} f(x_1, x_2) &= \frac{1}{\pi\sqrt{2}} \exp\left\{-(x_1^2 + x_2^2)\right\} \\ &\times \left( \exp\left\{x_1^2/2\right\} + \exp\left\{x_2^2/2\right\} - \sqrt{2} \right) \end{aligned}$$

for  $-\infty < x_1 < \infty$  and  $-\infty < x_2 < \infty$ .

A simple calculation shows that  $X_1 \sim N(0, 1)$  and  $X_2 \sim N(0, 1)$ , but the joint distribution of  $(X_1, X_2)$  is not bivariate normal.

The multivariate normal distribution is completely characterized by its mean vector and covariance matrix. This means that once the mean vector and covariance matrix are specified, the density and MGF of the MVN are completely determined.

## Independence of MVN

### General Definition of Independence

Two random vectors are independent if their joint density  $f(x, y)$  factors into

$$f(x, y) = f_1(x)f_2(y)$$

where  $f_1(x)$  is the marginal density of  $X$  and  $f_2(y)$  is the marginal density of  $Y$ .

### Theorem

If  $X$  and  $Y$  are independent random vectors then  $G(X)$  and  $H(Y)$  are independent where  $G(\cdot)$  and  $H(\cdot)$  are arbitrary functions. For example if  $X$  and  $Y$  are independent then  $X^2$  and  $Y^{10} + \exp\{Y\}$  are independent.

### Theorem

Suppose  $X \sim N_n(\mu, \Sigma)$ . Define  $Y_1 = AX$  and  $Y_2 = BX$  where  $A$  is an  $r \times n$  matrix of constants and  $B$  is an  $s \times n$  matrix of constants. Then  $Y_1$  and  $Y_2$  are independent if and only if  $A\Sigma B' = 0$ . If  $\Sigma = \sigma^2 I$ , then  $Y_1$  and  $Y_2$  are independent if and only if  $AB' = 0$ .

# Distribution Theory

## Theorem

Suppose  $X \sim N_n(\mu, \Sigma)$ . Partition  $X$  into  $X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$  where  $X_1$  is  $r \times 1$  and  $X_2$  is  $(n - r) \times 1$ . Partition  $\mu$  as  $\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$  where  $\mu_1$  is  $r \times 1$  and  $\mu_2$  is  $(n - r) \times 1$ . Similarly partition  $\Sigma$  as

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix},$$

where  $\Sigma_{11}$  is  $r \times r$ ,  $\Sigma_{12}$  is  $r \times (n - r)$ ,  $\Sigma_{21} = \Sigma'_{12}$  is  $(n - r) \times r$ , and  $\Sigma_{22}$  is  $(n - r) \times (n - r)$ . Then  $X_1$  and  $X_2$  are independent if and only if  $\Sigma_{12} = 0$ .

## Theorem

If  $X \sim N_n(\mu_x, \Sigma_x)$  and  $Y \sim N_m(\mu_y, \Sigma_y)$ , and  $X$  and  $Y$  are independent, then

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N_{n+m}(\mu, \Sigma)$$

where

$$\mu = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}$$

and

$$\Sigma = \begin{pmatrix} \Sigma_x & 0 \\ 0' & \Sigma_y \end{pmatrix}.$$

# Distribution Theory

## Theorem

If  $X \sim N_n(\mu, \Sigma)$ , then

$$E(X) = \text{mode}(X) = \text{median}(X) = \mu .$$

## Proof

We have already proved that the mean is  $\mu$  using the MGF. To show that the mode is  $\mu$ , one needs to minimize  $(x - \mu)' \Sigma^{-1} (x - \mu)$ . We have already done this minimization in weighted least squares estimation. To show that  $\mu$  is the median, we can easily see that the density of  $X$  is symmetric about  $\mu$ , that is,  $f(\mu + x) = f(\mu - x)$ .

## Theorem

Suppose  $X = (X_1, \dots, X_n)'$  has density of the form

$$f(x) = c \exp \{-Q/2\} \tag{10}$$

where  $\exp \{-Q/2\} \propto \exp \left\{ \frac{-1}{2} (x - \mu)' \Sigma^{-1} (x - \mu) \right\}$  and  $c$  is the normalizing constant. Then  $X \sim N_n(\mu, \Sigma)$ .

# Distribution Theory

## Remarks

- 1) We note that if  $f(x)$  has the form given in (10), we can find  $\mu = E(X)$  by finding the mode of  $X$ . This reduces to minimizing  $Q$ . Thus,  $\mu$  is the solution to the equations

$$\frac{\partial Q}{\partial x_j} = 0, \quad j = 1, \dots, n.$$

These  $n$  equations will be linear in the  $x_j$ 's.

- 2) The elements of  $\Sigma^{-1}$  are contained in the quadratic term  $x'\Sigma^{-1}x$ . We note here that

$$\begin{aligned} & (x - \mu)' \Sigma^{-1} (x - \mu) \\ = & x' \Sigma^{-1} x - 2x' \Sigma^{-1} \mu + \mu' \Sigma^{-1} \mu. \end{aligned}$$

# Distribution Theory

## Example

Suppose  $X = (X_1, X_2)'$  has density of the form  $f(x) = c \exp\{-Q/2\}$ , where

$$Q = x_1^2 + 2x_1x_2 + 4x_2^2 + 2x_1 .$$

What is the distribution of  $X$ ?

We know that  $X$  must be multivariate normal by the theorem. To find  $E(X)$ , we have

$$\frac{\partial Q}{\partial x_1} = 2x_1 + 2x_2 + 2 = 0$$

and

$$\frac{\partial Q}{\partial x_2} = 2x_1 + 8x_2 = 0$$

Solving these equations for  $x_1$  and  $x_2$  leads to  $x_1 = -4/3$  and  $x_2 = 1/3$ . Thus  $\mu = (-4/3, 1/3)'$ . To find the elements of  $\Sigma^{-1}$ , we look at the quadratic terms in  $Q$ . Let

$$\Sigma^{-1} = \begin{pmatrix} \sigma^{(11)} & \sigma^{(12)} \\ \sigma^{(12)} & \sigma^{(22)} \end{pmatrix}$$

# Distribution Theory

Thus

$$\begin{aligned} x' \Sigma^{-1} x &= (x_1, x_2) \begin{pmatrix} \sigma^{(11)} & \sigma^{(12)} \\ \sigma^{(12)} & \sigma^{(22)} \end{pmatrix} (x_1, x_2)' \\ &= \sigma^{(11)} x_1^2 + 2\sigma^{(12)} x_1 x_2 + \sigma^{(22)} x_2^2. \end{aligned}$$

Thus for our problem,  $\sigma^{(11)} = 1$ ,  $2\sigma^{(12)} = 2$  and  $\sigma^{(22)} = 4$ . Therefore,

$$\Sigma^{-1} = \begin{pmatrix} 1 & 1 \\ 1 & 4 \end{pmatrix}.$$

Inverting this gives

$$\Sigma = \begin{pmatrix} 4/3 & -1/3 \\ -1/3 & 1/3 \end{pmatrix}.$$

Thus

$$X \sim N_2 \left( \begin{pmatrix} -4/3 \\ 1/3 \end{pmatrix}, \begin{pmatrix} 4/3 & -1/3 \\ -1/3 & 1/3 \end{pmatrix} \right).$$

# Distribution Theory

## Distribution of Quadratic Forms

### Defn

Suppose  $Y$  is an  $n$  dimensional random vector and let  $A$  be an  $n \times n$  matrix of constants. A quadratic form is a random variable defined by  $Y'AY$  for some  $Y$  and  $A$ .

Since  $Y'AY$  is real-valued, we have

$$Y'AY = Y'A'Y = Y' \left( \frac{A + A'}{2} \right) Y .$$

Since  $(A + A')/2$  is always symmetric for any  $A$ , we can without loss of generality restrict ourselves to quadratic forms where  $A$  is symmetric.

### Theorem

Suppose  $Y \sim N_n(0, \sigma^2 I)$ . Then

$$\frac{1}{\sigma^2} (Y'MY) \sim \chi^2(r)$$

if and only if  $M$  is an orthogonal projection operator of rank  $r$ .

# Distribution Theory

## Theorem

Suppose  $Y \sim N_n(\mu, \sigma^2 I)$ . Then

$$\frac{1}{\sigma^2}(Y' M Y) \sim \chi^2(r, \gamma)$$

if and only if  $M$  is an orthogonal projection operator of rank  $r$  and  $\gamma = \frac{\mu' M \mu}{2\sigma^2}$ .

## Theorem

Suppose  $Y \sim N_n(\mu, \sigma^2 M)$  where  $M$  is an orthogonal projection operator of rank  $r$  and  $\mu \in C(M)$ . Then

$$\frac{1}{\sigma^2} Y' Y \sim \chi^2(r, \gamma)$$

where  $\gamma = \frac{\mu' \mu}{2\sigma^2}$ .

# Distribution Theory

## Theorem

Suppose  $Y \sim N_n(\mu, \Sigma)$  where  $\Sigma$  is positive definite. Then

$$Y'AY \sim \chi^2(r, \gamma)$$

where  $\gamma = \frac{\mu'A\mu}{2}$  if and only if any of the following conditions are satisfied:

- i)  $A\Sigma$  is a projection operator of rank  $r$ .
- ii)  $\Sigma A$  is a projection operator of rank  $r$ .
- iii)  $\Sigma$  is a generalized inverse of  $A$  and  $A$  has rank  $r$ .

Note: The noncentrality parameter is always found by replacing  $Y$  with  $E(Y)$  in the quadratic form. For example, for the theorem above, the noncentrality parameter is derived as

$$\gamma = \frac{E(Y)'AE(Y)}{2} = \frac{\mu'A\mu}{2} .$$

# Distribution Theory

## Theorem

Suppose  $Y \sim N_n(\mu, \Sigma)$  and  $\Sigma$  is positive semi-definite. Then

$$Y'AY \sim \chi^2(tr(A\Sigma), \gamma), \gamma = \frac{\mu'A\mu}{2}, \text{ if}$$

- i)  $\Sigma A \Sigma A \Sigma = \Sigma A \Sigma$  and
- ii)  $\mu' A \Sigma A \mu = \mu' A \mu$  and
- iii)  $\Sigma A \Sigma A \mu = \Sigma A \mu$ .

## Theorem

Suppose  $Y \sim N_n(\mu, \Sigma)$ , where  $\Sigma$  is positive definite. Then  $Y'AY$  has the same distribution as the random variable

$$U = \sum_{i=1}^n d_{ii} U_i$$

where  $d_{ii}$  are the eigenvalues of  $A\Sigma$  and  $U_1, \dots, U_n$  are independent non-central chi-square random variables with one degree of freedom.

# Distribution Theory

## Independence of Quadratic Forms

Theorem If  $Y \sim N_n(\mu, \sigma^2 I)$ , then

- i)  $Y'AY$  and  $BY$  are independent if and only if  $AB' = 0$  where  $A$  is a symmetric matrix.
- ii)  $Y'AY$  and  $Y'BY$  are independent if and only if  $AB = 0$ , where  $A$  and  $B$  are symmetric.

Proof of i)

Recall that  $AY$  and  $BY$  are independent if  $\text{Cov}(AY, BY) = 0$ . But  $\text{Cov}(AY, BY) = AC\text{Cov}(Y, BY)B' = A(\sigma^2 I)B' = \sigma^2 AB'$ . This quantity equals 0 if and only if  $AB' = 0$ . Since  $Y'AY$  is a function of  $AY$  it follows that  $Y'AY$  and  $BY$  are independent if and only if  $AB' = 0$ . Note that

$$\begin{aligned} AB' &= 0 \\ \Leftrightarrow (AB')' &= 0 \\ \Leftrightarrow BA' &= 0 \\ \Leftrightarrow BA &= 0 \quad \text{if } A \text{ is symmetric} \\ \Leftrightarrow AB &= 0 \quad \text{if } A \text{ and } B \text{ are symmetric.} \end{aligned}$$

A similar argument can be given for ii).

# Distribution Theory

## Theorem

Suppose  $Y \sim N_n(\mu, \Sigma)$ , and suppose that  $A$ ,  $B$ , and  $\Sigma$  are all positive semidefinite. Then  $Y'AY$  and  $Y'BY$  are independent if  $\Sigma A \Sigma B \Sigma = 0$ . If  $\Sigma$  is positive definite, then  $Y'AY$  and  $Y'BY$  are independent if  $A\Sigma B = 0$ .

## Proof

Since  $A$ ,  $B$ , and  $\Sigma$  are positive semidefinite, we can write  $A = RR'$ ,  $B = SS'$  and  $\Sigma = QQ'$ . Then

$$Y'AY = Y'RR'Y = (R'Y)'(R'Y)$$

and

$$Y'BY = Y'SS'Y = (S'Y)'(S'Y)$$

Thus  $Y'AY$  and  $Y'BY$  are independent if  $R'Y$  and  $S'Y$  are independent.  $R'Y$  and  $S'Y$  are independent if and only if

$$\begin{aligned} \text{Cov}(R'Y, S'Y) &= 0 \\ \Leftrightarrow R'\Sigma S &= 0 \\ \Leftrightarrow R'QQ'S &= 0 \\ \Leftrightarrow C(Q'S) &\perp C(Q'R). \end{aligned}$$

# Distribution Theory

Since  $C(AA') = C(A)$  for any matrix  $A$ , we have

$$\begin{aligned} C(Q'S) \perp C(Q'R) &\Leftrightarrow C(Q'SS'Q) \perp C(Q'RR'Q) \\ &\Leftrightarrow (Q'SS'Q)(Q'RR'Q) = 0 \\ &\Leftrightarrow Q'B\Sigma A Q = 0 \\ &\Leftrightarrow C(Q) \perp C(B\Sigma A Q) \\ &\Leftrightarrow C(QQ') \perp C(B\Sigma A Q) \\ &\Leftrightarrow QQ'B\Sigma A Q = 0 \\ &\Leftrightarrow \Sigma B\Sigma A Q = 0 \end{aligned}$$

Since  $C(Q) = C(QQ') = C(\Sigma)$ , we have

$$\begin{aligned} \Sigma B\Sigma A Q &= 0 \\ &\Leftrightarrow \Sigma B\Sigma A \Sigma = 0 \\ &\Leftrightarrow \Sigma A \Sigma B \Sigma = 0 \quad \text{by taking transposes} \end{aligned}$$

This completes the proof.

# Distribution Theory

If  $\Sigma^{-1}$  exists, we have

$$\begin{aligned}\Sigma A \Sigma B \Sigma &= 0 \\ \Leftrightarrow \Sigma^{-1} \Sigma A \Sigma B \Sigma \Sigma^{-1} &= 0 \\ \Leftrightarrow A \Sigma B &= 0 \\ \Leftrightarrow B \Sigma A &= 0\end{aligned}$$

# Distribution Theory

## Theorem

Suppose  $Y \sim N_n(\mu, \Sigma)$ ,  $\Sigma$  is positive semi-definite, and suppose  $A$  and  $B$  are  $n \times n$  symmetric matrices. If

- i)  $\Sigma A \Sigma B \Sigma = 0$  and
- ii)  $\Sigma A \Sigma B \mu = 0$  and
- iii)  $\Sigma B \Sigma A \mu = 0$  and
- iv)  $\mu' A \Sigma B \mu = 0$  ,

Then  $Y'AY$  and  $Y'BY$  are independent.

Here  $A$  and  $B$  are symmetric and not necessarily positive semidefinite. To prove this theorem, we write  $\Sigma = QQ'$  and  $Y = \mu + QZ$ , where  $Z \sim N_n(0, I)$ . Using this decomposition of  $Y$ , we multiply  $Y'AY$  and  $Y'BY$  out and check independence of the terms using Theorem 1.3.7 of Christensen and the argument contained in the proof of Theorem 1.3.8.

## Remarks

- 1) If  $Y \sim N_n(\mu, \Sigma)$  and  $\Sigma$  is positive semi-definite, then  $AY$  and  $BY$  are independent if and only if  $A\Sigma B' = 0$ . If  $AY$  is independent of  $BY$  then  $Y'AY$  and  $BY$  are independent, and  $Y'AY$  and  $Y'BY$  are independent.
- 2) If  $\Sigma$  is full rank (i.e., positive definite) and  $Y'AY$  and  $BY$  are independent, then  $AY$  and  $BY$  are independent. Also, if  $Y'AY$  and  $Y'BY$  are independent, then  $AY$  and  $BY$  are independent.
- 3) However if  $\Sigma$  is less than full rank then, if  $Y'AY$  and  $BY$  are independent, then this does NOT imply that  $AY$  and  $BY$  are independent. Also, if  $Y'BY$  and  $AY$  are independent, then this does NOT imply that  $AY$  and  $BY$  are independent.

# Distribution Theory

## Maximum Likelihood Estimation

We want to examine maximum likelihood estimation of estimable functions of  $\beta$  and  $\sigma^2$  in the linear model.

Consider the usual linear model

$$Y = X\beta + \epsilon ,$$

where  $\epsilon \sim N_n(0, \sigma^2 I)$  and  $X$  has rank  $r$ . Using properties of MVN, this implies that

$$Y \sim N_n(X\beta, \sigma^2 I) .$$

The joint density of  $Y = (Y_1, \dots, Y_n)'$  is given by

$$f(Y) = (2\pi)^{-n/2} \sigma^{-n} \exp \left\{ \frac{-1}{2\sigma^2} (Y - X\beta)'(Y - X\beta) \right\} .$$

We note here that  $| \sigma^2 I |^{-1/2} = \sigma^{-n}$ . The likelihood function of the parameters, denoted  $L(\beta, \sigma)$ , is any function proportional to the density function  $f(Y)$ .

That is

$$L(\beta, \sigma) \propto f(Y) .$$

dropping the  $(2\pi)^{-n/2}$  term, we have

$$L(\beta, \sigma) = \sigma^{-n} \exp \left\{ \frac{-1}{2\sigma^2} (Y - X\beta)'(Y - X\beta) \right\} .$$

# Distribution Theory

The maximizers of  $L$  are called the maximum likelihood estimates (MLE's). Maximizing  $L(\beta, \sigma)$  is equivalent to maximizing  $\ell(\beta, \sigma) = \log(L(\beta, \sigma))$ . We have

$$\ell(\beta, \sigma) = -n \log(\sigma) - \frac{1}{2\sigma^2} (Y - X\beta)'(Y - X\beta).$$

Now maximizing  $\ell(\beta, \sigma)$  with respect to  $\beta$  is equivalent to minimizing  $g(\beta) = (Y - X\beta)'(Y - X\beta)$  with respect to  $\beta$ . This is just the least squares criterion. Thus the MLE of  $\beta$  reduces the least squares criterion, and thus the MLE of  $\beta$  for the usual linear model, denoted  $\hat{\beta}_{ml}$  satisfies

$$X\hat{\beta}_{ml} = MY$$

where  $M = X(X'X)^{-1}X'$  is the orthogonal projection operator onto  $C(X)$ .

To find the MLE of  $\sigma$ , we substitute the MLE of  $\beta$  into  $\ell(\beta, \sigma)$ . Thus

$$\begin{aligned}\ell(\hat{\beta}_{ml}, \sigma) &= -n \log(\sigma) - \frac{1}{2\sigma^2} (Y - X\hat{\beta}_{ml})'(Y - X\hat{\beta}_{ml}) \\ &= -n \log(\sigma) - \frac{1}{2\sigma^2} (Y - MY)'(Y - MY) \\ &= -n \log(\sigma) - \frac{1}{2\sigma^2} Y'(I - M)Y.\end{aligned}$$

# Distribution Theory

Now we take derivatives with respect to  $\sigma$  and set equal to 0. Thus

$$\begin{aligned}\frac{\partial \ell(\hat{\beta}_{ml}, \sigma)}{\partial \sigma} &= \frac{-n}{\sigma} + \frac{1}{\sigma^3}(Y'(I - M)Y) = 0 \\ \Leftrightarrow -n\sigma^2 + Y'(I - M)Y &= 0. \\ \Leftrightarrow \sigma^2 &= \frac{Y'(I - M)Y}{n}\end{aligned}$$

Thus the maximum likelihood estimate of  $\sigma^2$  is

$$\hat{\sigma}_{ml}^2 = \frac{Y'(I - M)Y}{n}.$$

We see that  $\hat{\sigma}_{ml}^2$  is biased for estimating  $\sigma^2$ . In particular

$$E(\hat{\sigma}_{ml}^2) = \frac{n-r}{n}\sigma^2$$

Which converges to  $\sigma^2$  as  $n \rightarrow \infty$ . Thus  $\hat{\sigma}_{ml}^2$  is asymptotically unbiased.

# Distribution Theory

## Minimum Variance Unbiased Estimation

We saw by the Gauss-Markov theorem that the least squares estimator of  $\rho'X\beta$  is the unique minimum variance unbiased linear estimator. Now we will show that if  $\epsilon \sim N_n(0, \sigma^2 I)$ , then the least squares estimator is the uniform minimum variance unbiased estimator (UMVUE). This is a stronger result since the UMVUE is not restricted to linear estimators. It covers the class of ALL unbiased estimators.

To show this, we need to establish the notion of completeness.

### Defn

Suppose  $T(Y)$  is a vector valued statistic in  $Y$ . Then  $T(Y)$  is said to be a complete sufficient statistic for the family of distributions indexed by  $\theta \in \Theta$ , if  $T(Y)$  is sufficient and

$$E[f(T(Y))] = 0 \Rightarrow f(T(Y)) = 0$$

with probability 1 for all  $\theta \in \Theta$ , where  $\Theta$  denotes the parameter space.

### Theorem

If  $T(Y)$  is a complete sufficient statistic, then  $f(T(Y))$  is the unique UMVUE of  $E(f(T(Y)))$ .

# Distribution Theory

## Complete Sufficient Statistics for Exponential Families

### Theorem

Let  $\theta = (\theta_1, \dots, \theta_p)'$  and let  $Y$  be a random vector with density

$$f(Y) = h(Y)c(\theta) \exp \left\{ \sum_{i=1}^p \theta_i T_i(Y) \right\}.$$

Then  $T(Y) = (T_1(Y), \dots, T_p(Y))'$  is a complete sufficient statistic for  $\theta$ , if  $\theta \in A$ , where  $A$  is an open subset in  $R^p$ .

The requirement for an open subset always exists if there are no restrictions on the parameter vector  $\theta$ . If there are constraints (restrictions) on the components of  $\theta$ , then such an open subset does not exist. This is the case in linear models when the  $X$  matrix is less than full rank. In this case, the components of  $\beta$  have constraints. That is the model is overparameterized and some components of  $\beta$  are redundant.

To overcome this, we consider a reparametrization to remove the redundant components of  $\beta$ . Suppose  $r(X) = r$  and let  $Z$  be a matrix whose columns form a basis for  $C(X)$ . Thus for some matrix  $A$ ,

$$X = ZA$$

where  $Z$  is  $n \times r$  of rank  $r$  and  $A$  is  $r \times p$ .

# Distribution Theory

Let  $\lambda'\beta$  be an estimable function. Then  $\lambda' = \rho'X$  for some  $\rho \in R^n$ . Thus

$$\lambda'\beta = \rho'X\beta = \rho'ZA\beta .$$

Now let  $\gamma = A\beta$ . Thus  $\gamma$  is  $r \times 1$ , Now consider the reparametrized linear model

$$Y = Z\gamma + \epsilon$$

where  $\epsilon \sim N_n(0, \sigma^2 I)$ .  $Z$  is  $n \times r$  of full rank  $r$ . The least squares estimate of  $\lambda'\beta = \rho'Z\gamma$  is  $\rho'MY$  regardless of the rank of the model. That is, regardless of the reparameterized model or the originally parameterized model.

We want to now show that  $\rho'MY$  is the unique minimum variance unbiased estimate of  $\rho'X\beta$ . The density of  $Y$  is given by

$$\begin{aligned} f(Y) &= (2\pi)^{-n/2}\sigma^{-n} \exp \left\{ \frac{-1}{2\sigma^2}(Y - Z\gamma)'(Y - Z\gamma) \right\} \\ &= (2\pi)^{-n/2}\sigma^{-n} \exp \left\{ \frac{-1}{2\sigma^2}(\gamma'Z'Z\gamma) \right\} \\ &\quad \times \exp \left\{ \frac{-1}{2\sigma^2}(Y'Y) + \frac{\gamma'}{\sigma^2}(Z'Y) \right\} \\ &= h(Y)c(\gamma, \sigma^2) \exp \left\{ \frac{-1}{2\sigma^2}(Y'Y) + \frac{\gamma'}{\sigma^2}(Z'Y) \right\} , \end{aligned}$$

# Distribution Theory

where

$$h(Y) = (2\pi)^{-n/2}$$

and

$$c(\gamma, \sigma^2) = \sigma^{-n} \exp \left\{ \frac{-1}{2\sigma^2} (\gamma' Z' Z \gamma) \right\} .$$

This density is of the form of Theorem 2.5.3 of Christensen, and since there are NO restrictions on  $\gamma$ , the parameters  $\theta = (\frac{-1}{2\sigma^2}, \frac{\gamma_1}{\sigma^2}, \dots, \frac{\gamma_r}{\sigma^2})'$  contain an open subset in  $R^{r+1}$ .

It follows that  $(Y'Y, Z'Y)$  are complete sufficient statistics for  $\theta$ .

An unbiased estimate of  $\lambda'\beta = \rho'X\beta$  is  $\rho'MY = \rho'Z(Z'Z)^{-1}Z'Y$ . Thus  $\rho'MY$  is a function of the complete sufficient statistic so it is the unique minimum variance unbiased estimator of  $\rho'X\beta$ .

Moreover,  $\frac{Y'(I-M)Y}{n-r}$  is an unbiased estimate of  $\sigma^2$ , and

$Y'(I - M)Y = Y'Y - (Y'Z)(Z'Z)^{-1}Z'Y$  is a function of the complete sufficient statistic  $(Y'Y, Y'Z)$ . Therefore it is the unique minimum variance unbiased estimator of  $\sigma^2$ . We now have the following result.

# Distribution Theory

## Theorem

Suppose

$$Y = X\beta + \epsilon$$

where  $\epsilon \sim N_n(0, \sigma^2 I)$  and  $r(X) = r$ . Then

$$\frac{Y'(I - M)Y}{n - r}$$

is the unique UMVUE of  $\sigma^2$  and

$$\rho' MY$$

is the unique UMVUE of  $\rho' X\beta$ .

# Distribution Theory

## Sampling Distributions of Estimates

Consider the usual linear model

$$Y = X\beta + \epsilon$$

where  $\epsilon \sim N_n(0, \sigma^2 I)$ , which implies

$$Y \sim N_n(X\beta, \sigma^2 I) .$$

We want to obtain the sampling distributions of the least squares and maximum likelihood estimates.

Suppose  $\Lambda' \beta$  is an estimable vector of linear functions of  $\beta$ , where  $\Lambda$  is a  $p \times s$  matrix. We have  $\Lambda' = P'X$  for some  $n \times s$  matrix  $P$ . The BLUE and UMVUE of  $P'X\beta$  is  $P'MY$ . We note that

i)

$$\begin{aligned} E(P'MY) &= P'ME(Y) \\ &= P'MX\beta \\ &= P'X\beta . \end{aligned}$$

# Distribution Theory

ii)

$$\begin{aligned}\text{Cov}(P' M Y) &= (P' M) \text{Cov}(Y) (P' M)' \\ &= P' M (\sigma^2 I) M P \\ &= \sigma^2 P' M P.\end{aligned}$$

Thus

$$P' M Y \sim N_s(P' X \beta, \sigma^2 P' M P).$$

Note that we can write  $P' M P = P' X (X' X)^{-} X' P = \Lambda' (X' X)^{-} \Lambda$ , and therefore an alternative representation of the sampling distribution of  $P' M Y$  is

iii)  $P' M Y \sim N_s(\Lambda' \beta, \sigma^2 \Lambda' (X' X)^{-} \Lambda).$

# Distribution Theory

## Some special cases of interest

- a) Let  $P = I_{n \times n}$ . Then the least squares estimate of  $E(Y) = \mu = X\beta$  is  $MY$ . Thus

$$MY \sim N_n(X\beta, \sigma^2 M).$$

- b) Suppose  $X$  has full rank  $p$ . Then  $\beta$  is estimable and the unique least squares (and UMVUE) of  $\beta$  is  $\hat{\beta} = (X'X)^{-1}X'Y$ . We have

$$\begin{aligned} E(\hat{\beta}) &= E((X'X)^{-1}X'Y) \\ &= (X'X)^{-1}X'E(Y) = (X'X)^{-1}X'(X\beta) \\ &= (X'X)^{-1}(X'X)\beta \\ &= \beta. \end{aligned}$$

$$\begin{aligned} \text{Cov}(\hat{\beta}) &= ((X'X)^{-1}X') \text{Cov}(Y) ((X'X)^{-1}X')' \\ &= (X'X)^{-1}X'(\sigma^2 I)X(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1}(X'X)(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1}. \end{aligned}$$

# Distribution Theory

Thus

$$\hat{\beta} \sim N_p(\beta, \sigma^2(X'X)^{-1}) .$$

The sampling distribution for the estimator of  $\sigma^2$  is obtained as follows. We know that  $\frac{Y'(I-M)Y}{n-r}$  is the UMVUE of  $\sigma^2$ . Since  $Y \sim N_n(X\beta, \sigma^2 I)$ , and  $I - M$  is an orthogonal projection operator of rank  $n - r$ , it follows by an earlier theorem that

$$\frac{1}{\sigma^2} (Y'(I - M)Y) \sim \chi^2(n - r, \gamma) ,$$

where

$$\gamma = \frac{(X\beta)'(I - M)(X\beta)}{2\sigma^2} = \frac{\beta' X(I - M)X\beta}{2\sigma^2} = 0 ,$$

Since  $(I - M)X = 0$ .

Thus

$$\frac{1}{\sigma^2} (Y'(I - M)Y) \sim \chi^2(n - r) .$$

We can also write

$$Y'(I - M)Y \sim \sigma^2 \chi^2(n - r) .$$

# Distribution Theory

Now consider the linear model

$$Y = X\beta + \epsilon$$

where  $\epsilon \sim N_n(0, \sigma^2 V)$  where  $V$  is a known positive definite matrix. We have the following results:

- i)  $\rho' A Y$  is the unique UMVUE of  $\rho' X \beta$ , where  $A = X(X' V^{-1} X)^{-1} X' V^{-1}$ .
- ii)  $\rho' A Y \sim N_1(\rho' X \beta, \sigma^2 \rho' A V A' \rho)$ .
- iii)  $P' A Y$  is the unique UMVUE of  $P' X \beta$ , where  $P$  is an  $n \times s$  matrix of constants. Also,

$$P' A Y \sim N_s(P' X \beta, \sigma^2 P' A V A' P) .$$

- iii)  $\frac{Y'(I-A)' V^{-1} (I-A) Y}{n-r}$  is the unique UMVUE of  $\sigma^2$ .
- iv) Any weighted least squares estimate of  $\beta$  is an MLE of  $\beta$ .
- v) If  $X$  has full rank  $p$ , then the UMVUE of  $\beta$  is

$$\hat{\beta} = (X' V^{-1} X)^{-1} X' V^{-1} Y ,$$

and

$$\hat{\beta} \sim N_p(\beta, \sigma^2 (X' V^{-1} X)^{-1}) .$$

# Hypothesis Testing

## Hypothesis Testing

Consider the linear model

$$Y = X\beta + \epsilon \quad (1)$$

where  $\epsilon \sim N_n(0, \sigma^2 I)$ .

In hypothesis testing terminology, we call  $C(X)$  the estimation space and  $C(X)^\perp$  the error space.

Since  $E(Y) = X\beta$ , model (1) specifies that  $E(Y) \in C(X)$  and  $\text{Cov}(Y) = \sigma^2 I$ .

We are interested in testing a linear model against a reduced linear model. That is, we are interested in testing nested linear models. This is the only hypothesis testing situation we will consider. We will develop the null and alternative hypotheses in terms of vector spaces.

# Hypothesis Testing

All hypothesis testing in linear models reduces to specifying a constraint on the estimation space  $C(X)$ . We usually start with the “full” model in (1) and we wish to reduce this model somehow. That is, we want to know if a simpler, more parsimonious model is acceptable. Thus, we consider the reduced model

$$Y = X_0 \gamma_0 + \epsilon \tag{2}$$

where  $\epsilon \sim N_n(0, \sigma^2 I)$  and

$$C(X_0) \subset C(X).$$

We assume that the full model in (1) is correct, so that if the model in (2) is correct, then so is the model in (1). Since  $C(X_0) \subset C(X)$ , we say that the model in (2) is nested in the model in (1). Model (2) specifies that  $E(Y) \in C(X_0)$ .

# Hypothesis Testing

## Examples of nested models

### One-way ANOVA

The “full” model for one-way ANOVA is

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij},$$

where  $\mu$  denotes the grand mean and  $\alpha_i$  is the  $i$ th treatment effect,  $j = 1, \dots, n_i$ ,  $i = 1, \dots, t$ . We often wish to test the hypothesis of no treatment effect, that is,  $H_0 : \alpha_1 = \dots = \alpha_t$ . Thus the reduced model becomes

$$Y_{ij} = \mu + \epsilon_{ij}.$$

## Multiple linear regression

Suppose the full model is

$$Y = X_1\beta_1 + X_2\beta_2 + \epsilon$$

where  $X_1$  is  $n \times r$ ,  $X_2$  is  $n \times (p - r)$ ,  $\beta_1$  is  $r \times 1$  and  $\beta_2$  is  $(p - r) \times 1$ . Often we are interested in testing the hypothesis  $H_0 : \beta_1 = 0$ , so that the reduced model becomes

$$Y = X_2\beta_2 + \epsilon.$$

In this example the  $X$  matrix for the full model is  $X = (X_1, X_2)$  and the  $X$  matrix for the reduced model is  $X_0 = X_2$ .

# Hypothesis Testing

We can now express the null and alternative hypotheses for testing model (2) against model (1) in terms of  $E(Y)$ . The null hypothesis specifies  $E(Y) \in C(X_0)$ . To make the null and alternative disjoint hypotheses, the alternative specifies that  $E(Y) \in C(X)$  and  $E(Y) \notin C(X_0)$ . But this is precisely the same hypothesis as  $H_a : E(Y) \in C(X) \cap C(X_0)^c$ . Summarizing, we have

$$\begin{aligned}H_0 &: E(Y) \in C(X_0) \\H_a &: E(Y) \in C(X) \cap C(X_0)^c.\end{aligned}$$

We now give a heuristic argument for the  $F$  test for testing model (2) against model (1).

Let  $M = X(X'X)^{-1}X'$  denote the orthogonal projection operator onto  $C(X)$  and  $M_0 = X_0(X_0'X_0)^{-1}X_0'$  denotes the orthogonal projection operator onto  $C(X_0)$ . Since  $C(M_0) \subset C(M)$ , we note that  $M - M_0$  is also an orthogonal projection operator.  $C(M - M_0)$  is precisely the subspace within  $M$  that is orthogonal to  $M_0$  (See page 100).

C(M<sub>0</sub>)

C(M)

# Hypothesis Testing

Under the full model in (1), the unique UMVUE of  $\mu = E(Y)$  is  $MY$ , and under model (2), the unique UMVUE of  $\mu = E(Y)$  is  $M_0 Y$ .

If model (2) is correct, then  $MY$  and  $M_0 Y$  are estimates of the same quantity  $E(Y)$ , since the validity of model (2) implies the validity of model (1).

Thus, if model (2) is true,  $MY - M_0 Y = (M - M_0)Y$  should be small in some sense.

On the other hand, a large difference between  $MY$  and  $M_0 Y$  suggests that  $MY$  and  $M_0 Y$  are NOT estimating the same quantity. If  $M_0 Y$  is not estimating  $E(Y)$ , then model (2) cannot be correct because model (2) implies that  $M_0 Y$  is an estimate of  $E(Y)$ .

The decision about whether model (2) is correct hinges on whether the vector  $(M - M_0)Y$  is large. An obvious measure of size is the squared length of this vector given by

$$\|(M - M_0)Y\|^2 = Y'(M - M_0)Y .$$

# Hypothesis Testing

Why are you adjusting for sizes of the column space?

If we adjust for the relative sizes of the subspaces  $C(M)$  and  $C(M_0)$ , we divide  $\|(M - M_0)Y\|^2$  by  $r(M - M_0)$ . Since  $Y$  is random, our measure of size is

$$E \left( \frac{\|(M - M_0)Y\|^2}{r(M - M_0)} \right) = E \left( \frac{Y'(M - M_0)Y}{r(M - M_0)} \right). \quad (3)$$

At this point, we need some idea of how large this measure will be when model (2) is correct and when it is not correct. Thus, we need to compute the expectation of (3) above under both of these scenarios. Using our formula for the expectation of a quadratic form, we have:

- i) if model (2) is NOT true

$$\begin{aligned} & E \left( \frac{Y'(M - M_0)Y}{r(M - M_0)} \right) \\ &= tr \left( \frac{\sigma^2(M - M_0)}{r(M - M_0)} \right) + \frac{(X\beta)'(M - M_0)(X\beta)}{r(M - M_0)} \\ &= \sigma^2 \frac{r(M - M_0)}{r(M - M_0)} + \frac{\|(I - M_0)X\beta\|^2}{r(M - M_0)} \\ &= \sigma^2 + \frac{\|(I - M_0)X\beta\|^2}{r(M - M_0)}. \end{aligned}$$

# Hypothesis Testing

ii) If model (2) is TRUE then  $X\beta$  is replaced by  $X_0\gamma_0$  in i) and we have  $(I - M_0)X_0\gamma_0 = 0$  since  $(I - M_0)X_0 = 0$ . Thus if model (2) is true

$$\frac{\|(I - M_0)X\beta\|^2}{r(M - M_0)} = 0$$

and the formula reduces to

$$E \left( \frac{Y'(M - M_0)Y}{r(M - M_0)} \right) = \sigma^2 . \quad (4)$$

iii) Equation (4) tells us what the expected value of this quadratic form should be when model (2) is the correct model. Thus, if model (2) were correct, the quadratic form should be close to  $\sigma^2$ , and thus the ratio

$$\left( \frac{Y'(M - M_0)Y}{r(M - M_0)\sigma^2} \right) \approx 1 .$$

Thus if the ratio above is much larger than 1, this would indicate that model (2) is not correct. If  $\sigma^2$  was known, then we could compute this ratio, judge the size and be finished. But typically  $\sigma^2$  is not known in practice, and thus needs to be estimated.

# Hypothesis Testing

Since we always assume the full model to be “true”, we estimate  $\sigma^2$  from the full model. The estimate is thus the UMVUE of  $\sigma^2$  based on the full model. We have

$$\hat{\sigma}^2 = MSE = \frac{\|(I - M)Y\|^2}{r(I - M)} = \frac{Y'(I - M)Y}{r(I - M)}$$

Replacing  $\sigma^2$  by this estimate in equation (4) above, get the test statistic  $F$  for the null hypothesis  $H_0 : E(Y) \in C(X_0)$ . We have

$$F = \frac{Y'(M - M_0)Y}{MSE \ r(M - M_0)}$$

Letting  $r = r(X)$  and  $r_0 = r(X_0)$ ,  $r_0 < r$ , we have  $r(I - M) = n - r$  and  $r(M - M_0) = r - r_0$ .

# Hypothesis Testing

The term

$$\|(I - M_0)X\beta\|^2$$

is crucial in evaluating the behavior of the test statistic when model (2) is not correct. Actually, we can recognize this term as part of the noncentrality parameter. Specifically, the noncentrality parameter for the test above is

$$\gamma = \frac{\|(I - M_0)X\beta\|^2}{2\sigma^2} = \frac{\|(M - M_0)X\beta\|^2}{2\sigma^2}$$

and thus the size of the noncentrality parameter plays a major role in deciding on the validity of model (2). If  $\gamma$  is large, this implies that model (2) is not correct and if  $\gamma = 0$ , this implies that model (2) is correct. We are now led to the following theorem.

# Hypothesis Testing

## Theorem

Consider the usual linear model

$$Y = X\beta + \epsilon,$$

where  $\epsilon \sim N_n(0, \sigma^2 I)$ . Consider the reduced model

$$Y = X_0\gamma_0 + \epsilon,$$

where  $C(X_0) \subset C(X)$ . We wish to test the hypothesis

$$H_0 : E(Y) \in C(X_0)$$

$$H_a : E(Y) \in C(X) \cap C(X_0)^c.$$

Let  $M_0 = X_0(X_0'X_0)^{-}X_0$  be the orthogonal projection operator onto  $C(X_0)$  and  $M = X(X'X)^{-}X'$  is the orthogonal projection operator onto  $C(X)$ . Further, let  $X$  be  $n \times p$  with  $r = r(X)$  and  $r_0 = r(X_0)$ ,  $r_0 < r$ . Then, under  $H_a$ ,

$$F = \frac{\|(M - M_0)Y\|^2 / (r - r_0)}{\|(I - M)Y\|^2 / (n - r)} \sim F(r - r_0, n - r, \gamma)$$

where  $\gamma = \frac{\|(I - M_0)X\beta\|^2}{2\sigma^2} = \frac{\|(M - M_0)X\beta\|^2}{2\sigma^2}$ .

# Hypothesis Testing

If model (2) is assumed correct, then  $\gamma = 0$  and

$$F = \frac{\|(M - M_0)Y\|^2/(r - r_0)}{\|(I - M)Y\|^2/(n - r)} \sim F(r - r_0, n - r) .$$

Thus an  $\alpha$  level test of the hypothesis

$$\begin{aligned} H_0 & : E(Y) \in C(X_0) \\ H_a & : E(Y) \in C(X) \cap C(X_0)^c \end{aligned}$$

rejects  $H_0$  if

$$F = \frac{\|(M - M_0)Y\|^2/(r - r_0)}{\|(I - M)Y\|^2/(n - r)} > F(1 - \alpha, r - r_0, n - r) ,$$

where  $F(1 - \alpha, r - r_0, n - r)$  is the  $(1 - \alpha) \times 100\%$  percentile of the  $F$  distribution with  $(r - r_0, n - r)$  degrees of freedom.

# Hypothesis Testing

## Proof

To prove this theorem, we need to show that the quadratic form in the numerator is noncentral chi-square, the denominator quadratic form is central chi-square and the two quadratic forms are independent.

1) Recall the theorem on quadratic forms that states that if

$Y \sim N_n(\mu, \sigma^2 I)$ , then  $\frac{1}{\sigma^2}(Y'MY) \sim \chi^2(r, \gamma)$  if and only if  $M$  is an orthogonal projection operator of rank  $r$ , and  $\gamma = \frac{\|M\mu\|^2}{2\sigma^2} = \frac{\mu'M\mu}{2\sigma^2}$ .

Here, by the theorem, we have

$$Y'(M - M_0)Y \sim \sigma^2 \chi^2(r - r_0, \gamma),$$

where  $\gamma = \frac{\|(M - M_0)X\beta\|^2}{2\sigma^2}$ . For the denominator quadratic form, we have  $Y'(I - M)Y \sim \sigma^2 \chi^2(n - r)$ . The denominator quadratic form is a central chi-square since its noncentrality parameter equals  $\frac{\|(I - M)X\beta\|^2}{2\sigma^2} = 0$  since  $(I - M)X\beta = 0$ .

# Hypothesis Testing

2) To prove independence of the quadratic forms, we use our theorems on independence of quadratic forms. it suffices to show that

$(M - M_0)(I - M) = 0$ . We see that

$$\begin{aligned}(M - M_0)(I - M) &= M - M^2 - M_0 + M_0 M \\ &= M - M - M_0 + M_0 = 0.\end{aligned}$$

The test above is called an *F* test since the test statistic *F* has an *F* distribution. The *F* test given here is very general in that it applies in any hypothesis testing situation in which we have nested models. We note here that  $\|(M - M_0)Y\|^2 = Y'(M - M_0)Y$  and

$\|(I - M)Y\|^2 = Y'(I - M)Y$ . Also, notes that

$r(I - M) = r(I) - r(M) = n - r$ , and

$r(M - M_0) = r(M) - r(M_0) = r - r_0$ .

# Hypothesis Testing

## Testing Linear Parametric Functions

Now we consider tests of hypotheses with linear constraints on the parameter vector  $\beta$ . Consider the usual linear model

$$Y = X\beta + \epsilon,$$

where  $\epsilon \sim N_n(0, \sigma^2 I)$ . Now suppose we want to test a hypothesis concerning an estimable function  $\Lambda' \beta$ , where  $\Lambda' = P'X$ , and  $P'$  is an  $s \times n$  matrix of constants. That is, we want to test  $H_0 : P'X\beta = 0$ . More formally, the hypotheses are

$$H_0 : E(Y) \in C(X) \text{ and } P'X\beta = 0$$

$$H_a : E(Y) \in C(X) \text{ and } P'X\beta \neq 0$$

## Examples

- a) Suppose  $\beta_1 - \beta_2$  is estimable and we wish to test  $H_0 : \beta_1 - \beta_2 = 0$ . Here,  $s = 1$  and  $\lambda' = (1, -1, 0, \dots, 0)$ .

# Hypothesis Testing

b) Suppose  $\begin{pmatrix} \beta_1 + \beta_3 \\ \beta_2 \end{pmatrix}$  is estimable and we wanted to test the hypothesis

$$H_0 : \beta_1 + \beta_3 = 0 \text{ and } \beta_2 = 0 .$$

Here,

$$\Lambda' = \begin{pmatrix} 1 & 0 & 1 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & \dots & 0 \end{pmatrix}_{2 \times p}$$

The condition  $P'X\beta = 0$  imposes a linear constraint on  $\beta$ . We need to find the reduced model  $X_0$  that corresponds to this linear constraint. Since  $E(Y) = X\beta$ , notice that the condition  $P'X\beta = 0$  imposes a restriction on  $E(Y)$ . This restriction is  $E(Y) \in \mathcal{N}(P') = C(P)^\perp$ . Thus the null hypothesis may now be stated as

$$H_0 : E(Y) \in C(X) \cap C(P)^\perp . \quad (5)$$

# Hypothesis Testing

If we can find a matrix  $X_0$  so that  $C(X_0) = C(X) \cap C(P)^\perp$ , then we can use the  $F$  test derived earlier for testing nested linear models. We note here that the choice of  $X_0$  corresponding to (5) is NOT unique since  $P$  is not unique.

Let  $M$  denote the orthogonal projection operator onto  $C(X)$ . We can decompose  $P$  into  $P = MP + (I - M)P$ , and thus

$$\begin{aligned} P'X\beta &= P'MX\beta + P'(I - M)X\beta \\ &= P'MX\beta \end{aligned}$$

Since  $P'X\beta = P'MX\beta$ , Thus  $P'X\beta = 0$  if and only if  $P'MX\beta = 0$  and therefore  $E(Y) \perp C(P)$  if and only if  $E(Y) \perp C(MP)$ . Thus we can state the null hypothesis as

$$H_0 : E(Y) \in C(X) \cap C(MP)^\perp$$

$C(MP)$  is a subspace in  $C(M)$  obtained by projecting  $P$  onto  $C(M)$ . We now have the following choices for  $X_0$  stated in the next two theorems.

Let  $M_{MP}$  denote the orthogonal projection operator onto  $C(MP)$ . By definition,  $M_{MP}$  is given by

$$M_{MP} = MP(P'MP)^{-}P'M$$

We now have the following theorem

# Hypothesis Testing

## Theorem

$$C((I - M_{MP})X) = C(X) \cap C(MP)^\perp$$

This is easily shown by showing that the two subspaces are contained in one another.

## Proof

1)  $\Rightarrow$

We want to show  $C((I - M_{MP})X) \subset C(X) \cap C(MP)^\perp$ . Let  $v \in C((I - M_{MP})X)$ . Then  $v = (I - M_{MP})Xz$  for some  $z \in R^p$ . But  $(I - M_{MP})Xz = Xz - M_{MP}Xz$ . Now  $Xz \in C(X)$  and  $M_{MP}Xz \in C(X)$  by definition of column space and since  $M_{MP}$  is an orthogonal projection operator in  $C(X)$ . Thus by definition of subspace, the difference of two vectors in a subspace, results in a vector in the same subspace, and thus  $v \in C(X)$ . To show that  $v \in C(MP)^\perp$ , notice that  $I - M_{MP}$  is the orthogonal projection operator onto  $C(MP)^\perp$ , and thus by definition  $(I - M_{MP})Xz$  is a vector in  $C(MP)^\perp$  for any  $z \in R^p$ . Thus  $v \in C(MP)^\perp$ , and thus  $v \in C(X) \cap C(MP)^\perp$ .

# Hypothesis Testing

2)  $\Leftarrow$

Now we need to show  $C(X) \cap C(MP)^\perp \subset C((I - M_{MP})X)$ . Let  $v \in C(X) \cap C(M_{MP})^\perp$ . Then  $v = Xw$  and  $v = (I - M_{MP})z$  for some  $w$  and  $z$ . Every vector in  $C((I - M_{MP})X)$  is of the form  $y = (I - M_{MP})Xw = (I - M_{MP})v = v$  since  $v \in C(I - M_{MP})$ , thus  $v \in C((I - M_{MP})X)$ . And thus  $C((I - M_{MP})X) = C(X) \cap C(MP)^\perp$

# Hypothesis Testing

This theorem implies that one choice of  $X_0$  is

$$X_0 = (I - M_{MP})X .$$

Since  $C(X) = C(M)$ , we can replace  $C(X)$  with  $C(M)$  in the theorem above to obtain

$$C(X) \cap C(MP)^\perp = C(M) \cap C(MP)^\perp = c(M - M_{MP}) .$$

And thus another choice for  $X_0$  is  $X_0 = M - M_{MP}$ . This choice of  $X_0$  is the most common one and will be the one we use.

We note here that  $X_0$  is not unique and need not have the same dimension as other  $X_0$ 's. For example,  $X_0 = (I - M_{MP})X$  is an  $n \times p$  matrix and  $X_0 = M - M_{MP}$  is an  $n \times n$  matrix. Thus a test of the reduced model for the hypothesis

$$H_0 : E(Y) \in C(X) \cap C(MP)^\perp$$

has numerator sum of squares given by

$$\begin{aligned} Y'(M - M_0)Y &= Y'(M - (M - M_{MP}))Y' \\ &= Y'M_{MP}Y \\ &= \|M_{MP}Y\|^2 . \end{aligned}$$

# Hypothesis Testing

Thus, the  $F$  test is given by

$$F = \frac{\|M_{MP}Y\|^2/r(M_{MP})}{\|(I-M)Y\|^2/r(I-M)} \sim F(r(M_{MP}), r(I-M), \gamma),$$

where  $\gamma = \frac{\|M_{MP}X\beta\|^2}{2\sigma^2}$ . Under the null hypothesis  $\gamma = 0$ , and the test statistic has a central  $F$  distribution.

Recall that the original hypothesis was  $H_0 : \Lambda'\beta = 0$ , where  $\Lambda' = P'X$ . We can write the  $F$  statistic in terms of  $\Lambda$  and  $\hat{\beta}$ , where  $\hat{\beta}$  is an MLE of  $\beta$ . We have

$$\begin{aligned} Y'M_{MP}Y &= Y'MP(P'MP)^{-}P'MY \\ &= \hat{\beta}'\Lambda(P'X(X'X)^{-}X'P)^{-}\Lambda'\hat{\beta} \\ &= \hat{\beta}'\Lambda(\Lambda'(X'X)^{-}\Lambda)^{-}\Lambda'\hat{\beta} \end{aligned}$$

We also note here that  $r(M_{MP}) = r(\Lambda)$ , and thus the  $F$  statistic becomes

$$F = \frac{\hat{\beta}'\Lambda(\Lambda'(X'X)^{-}\Lambda)^{-}\Lambda'\hat{\beta}/r(\Lambda)}{MSE}$$

where  $MSE = \frac{\|(I-M)Y\|^2}{r(I-M)}$ . The noncentrality parameter can be written as

$$\gamma = \frac{\beta'\Lambda(\Lambda'(X'X)^{-}\Lambda)^{-}\Lambda'\beta}{2\sigma^2}.$$

# Hypothesis Testing

We note here that

$$\text{Cov}(\Lambda' \hat{\beta}) = \sigma^2 \Lambda' (X' X)^{-1} \Lambda .$$

we can also write the test statistic as

$$F = \frac{(\Lambda' \hat{\beta})' (\text{Cov}(\Lambda' \hat{\beta}))^{-1} (\Lambda' \hat{\beta})}{r(\Lambda)} .$$

Thus

$$\frac{(\Lambda' \hat{\beta})' (\text{Cov}(\Lambda' \hat{\beta}))^{-1} (\Lambda' \hat{\beta})}{r(\Lambda)} \sim F(r(\Lambda), r(I - M), \gamma)$$

where

$$\gamma = \frac{\beta' \Lambda (\Lambda' (X' X)^{-1} \Lambda)^{-1} \Lambda' \beta}{2\sigma^2} .$$

The above form of the  $F$  test tells us that in order to obtain the  $F$  test of the hypothesis  $H_0 : \Lambda' \beta = 0$ , we need

- i) The UMVUE of  $\Lambda' \beta$ , denoted  $\Lambda' \hat{\beta}$ .
- ii) The  $\text{Cov}(\Lambda' \hat{\beta})$ .
- iii)  $r(\Lambda)$ .
- iv)  $MSE = \frac{\|(I - M)Y\|^2}{r(I - M)}$ .

# Hypothesis Testing

A special case of the general  $F$  test developed above is testing the hypothesis

$$H_0 : \lambda' \beta = 0$$

where  $\lambda' = \rho' X$ ,  $\rho \in R^n$ . The  $F$  test for this hypothesis is

$$F = \frac{(\lambda' \hat{\beta})^2}{MSE \lambda'(X'X)^{-1} \lambda} \sim F(1, r(I - M), \gamma)$$

where

$$\gamma = \frac{(\lambda' \beta)^2}{2\sigma^2 \lambda'(X'X)^{-1} \lambda} .$$

# Hypothesis Testing

## Generalized Hypothesis Test Procedure

Suppose we wish to test

$$H_0 : \Lambda' \beta = d \quad (6)$$

where  $\Lambda' = P'X$ , where  $P'$  is  $s \times n$ ,  $X$  is  $n \times p$ , and  $d$  is a known  $s \times 1$  vector. We want to derive the general  $F$  test for this hypothesis.

Since  $d$  may be non-zero, we need to do a little trickery to express the null hypothesis in terms of the column space that contains  $E(Y)$ . The null hypothesis stated in (6) describes a flat, which is not a subspace. We have to translate this flat back to the origin so that we can write the null hypothesis in terms of subspaces.

To this end, let  $b$  be any solution to the equation  $\Lambda' \beta = d$ . Thus  $b$  is a  $p \times 1$  vector. Then

$$P' X b = d ,$$

and the null hypothesis in (6) can be written as

$$H_0 : P' X \beta = P' X b ,$$

or

$$H_0 : P'(X\beta - Xb) = 0 . \quad (7)$$

# Hypothesis Testing

From the formulation in (7), we can write the reduced model as

$$Y = X\beta + \epsilon \text{ and } P'(X\beta - Xb) = 0 . \quad (8)$$

Letting  $\beta^* = \beta - b$ , we can rewrite (8) as

$$Y - Xb = X\beta^* + \epsilon \text{ and } P'X\beta^* = 0 . \quad (9)$$

Now we can write (9) in terms of subspaces. Thus (9) can be written as

$$H_0 : E(Y - Xb) \in C(X) \text{ and } E(Y - Xb) \perp C(P) , \quad (10)$$

which finally simplifies to

$$H_0 : E(Y - Xb) \in C(X) \cap C(MP)^\perp . \quad (11)$$

Thus for the hypothesis in (11), we can now write the reduced model as

$$Y - Xb = X_0\gamma_0 + \epsilon ,$$

where

$$X_0 = M - M_{MP} ,$$

$M$  is the orthogonal projection operator onto  $C(X)$ , and  $M_{MP}$  is the orthogonal projection operator onto  $C(MP)$ .

# Hypothesis Testing

The  $F$  statistic for testing (11) can now be written as

$$\begin{aligned} F &= \frac{\|M_{MP}(Y - Xb)\|^2 / r(M_{MP})}{\|(I - M)(Y - Xb)\|^2 / r(I - M)} \\ &= \frac{(Y - Xb)' M_{MP} (Y - Xb) / r(M_{MP})}{(Y - Xb)' (I - M) (Y - Xb) / r(I - M)}. \end{aligned}$$

We note here that the denominator of the  $F$  test does NOT depend on  $b$ . To see this, we have

$$\begin{aligned} &(Y - Xb)' (I - M) (Y - Xb) \\ &= Y' (I - M) Y - b' X' (I - M) \\ &\quad - (I - M) X b + b' X' (I - M) X b \\ &= Y' (I - M) Y \end{aligned}$$

since  $(I - M)X = 0$ .

# Hypothesis Testing

Also the numerator of the  $F$  test does not depend on  $b$  since

$$\begin{aligned} & (Y - Xb)' M_{MP} (Y - Xb) \\ = & (Y - Xb)' MP (P' MP)^{-1} P' M (Y - Xb) \\ = & (P' MY - P' MXb)' (P' MP)^{-1} (P' MY - P' MXb) \\ = & (\Lambda' \hat{\beta} - P' X\beta)' (P' X(X' X)^{-1} X' P)^{-1} (\Lambda' \hat{\beta} - P' X\beta) \\ = & (\Lambda' \hat{\beta} - d)' (\Lambda' (X' X)^{-1} \Lambda)^{-1} (\Lambda' \hat{\beta} - d) \end{aligned} \tag{12}$$

Thus, the  $F$  test is invariant with respect to the choice of  $b$ .

We can write the  $F$  test as

$$\begin{aligned} F &= \frac{\|M_{MP}(Y - Xb)\|^2 / r(M_{MP})}{\|(I - M)Y\|^2 / r(I - M)} \\ &= \frac{(Y - Xb)' M_{MP} (Y - Xb) / r(M_{MP})}{Y' (I - M) Y / r(I - M)} \\ &\sim F(r(M_{MP}), r(I - M), \gamma) \end{aligned}$$

where

$$\begin{aligned} \gamma &= \frac{\|M_{MP}(X\beta - Xb)\|^2}{2\sigma^2} \\ &= \frac{(X\beta - Xb)' M_{MP} (X\beta - Xb)}{2\sigma^2} \end{aligned}$$

# Hypothesis Testing

We can also write the  $F$  test in terms of  $\Lambda$  and  $\hat{\beta}$ . Using (12), we have

$$\begin{aligned} F &= \frac{(\Lambda' \hat{\beta} - d)' (\Lambda' (X' X)^{-1} \Lambda)^{-1} (\Lambda' \hat{\beta} - d)}{r(\Lambda) MSE} \\ &\sim F(r(\Lambda), r(I - M), \gamma) \end{aligned}$$

where

$$MSE = \frac{Y'(I - M)Y}{r(I - M)}$$

and

$$\gamma = \frac{(\Lambda' \beta - d)' (\Lambda' (X' X)^{-1} \Lambda)^{-1} (\Lambda' \beta - d)}{2\sigma^2}$$

# Hypothesis Testing

## Example

Consider the linear model

$$Y = X\beta + \epsilon$$

where  $Y = (Y_1, Y_2, Y_3)'$ ,  $\beta = (\beta_1, \beta_2)'$ ,

$$X = \begin{pmatrix} 1 & 2 \\ 1 & 2 \\ 1 & 2 \end{pmatrix}$$

and  $\epsilon \sim N_3(0, \sigma^2 I)$ . Suppose we wish to test

$$H_0 : \beta_1 + 2\beta_2 = 3 .$$

We can express this hypothesis as

$$H_0 : \rho'(X\beta - Xb) = 0$$

where  $\rho' = (1/3, 1/3, 1/3)$ , and  $b$  is any solution to  $\rho'X\beta = d$ . Solving this equation for  $b$ , we have

$$\rho'Xb = (1/3, 1/3, 1/3) \begin{pmatrix} 1 & 2 \\ 1 & 2 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = 3 .$$

# Hypothesis Testing

This equation implies that  $b_1 + 2b_2 = 3$ , so  $b_1 = 1$  and  $b_2 = 1$  is a solution.

Thus  $b = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$  is a solution. Thus we can write the null hypothesis as

$$H_0 : \rho'(X\beta - Xb) = 0$$

where  $\rho' = (1/3, 1/3, 1/3)$  and  $b = (1, 1)'$ . We note here that  $\rho$  was chosen so that  $\rho \in C(X)$ , and thus  $M\rho = \rho = (1/3, 1/3, 1/3)'$ . Thus

$$\begin{aligned} M_{M\rho} &= (M\rho)(\rho'M\rho)^{-1}(\rho'M) \\ &= \rho(\rho'\rho)^{-1}\rho' \\ &= \frac{\rho\rho'}{\rho'\rho} \\ &= \frac{1}{3} \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}. \end{aligned}$$

We see here that  $r(M_{M\rho}) = r(\rho) = 1$ . We also note that

$$Xb = \begin{pmatrix} 1 & 2 \\ 1 & 2 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = 3J_3$$

where  $J_3 = (1, 1, 1)'$ .

# Hypothesis Testing

Thus the numerator of the  $F$  test can now be written as

$$\begin{aligned} & \frac{(Y - 3J_3)' \rho \rho' (Y - 3J_3) / r(\rho)}{\rho' \rho} \\ = & \frac{(\rho' Y - 3\rho' J_3)(\rho' Y - 3\rho' J_3) / 1}{1/3} \\ = & 3(\rho' Y - 3\rho' J_3)^2 \\ = & 3(\bar{Y} - 3)^2 \end{aligned}$$

where  $\bar{Y} = \frac{1}{3} \sum_{i=1}^3 Y_i$ . We note here since  $C(X) = S \left\{ \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \right\}$ , we have

$$M = \frac{1}{3} \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix},$$

and  $r(M) = 1$ . Thus

$$\begin{aligned} MSE &= \frac{Y'(I - M)Y}{r(I - M)} \\ &= \frac{1}{2} \sum_{i=1}^3 (Y_i - \bar{Y})^2. \end{aligned}$$

# Hypothesis Testing

Thus the  $F$  test can be written as

$$\begin{aligned} F &= \frac{3(\bar{Y} - 3)^2}{\sum_{i=1}^3(Y_i - \bar{Y})^2/2} \\ &\sim F(1, 2, \gamma) \end{aligned}$$

where

$$\begin{aligned} \gamma &= \frac{3(E(\bar{Y}) - 3)^2}{2\sigma^2} \\ &= \frac{3(\beta_1 + 2\beta_2 - 3)^2}{2\sigma^2}. \end{aligned}$$

We note here that

$$\begin{aligned} E(\bar{Y}) &= \frac{1}{3} \sum_{i=1}^3 E(Y_i) \\ &= \frac{1}{3} \sum_{i=1}^3 (\beta_1 + 2\beta_2) \\ &= \beta_1 + 2\beta_2. \end{aligned}$$

# Hypothesis Testing

## Example

Consider the two-way ANOVA model without interaction. This model is given by

$$Y_{ijk} = \mu + \alpha_i + \eta_j + \epsilon_{ijk},$$

for  $i = 1, \dots, a$ ,  $j = 1, \dots, b$ , and  $k = 1, \dots, N$ . Suppose we wish to test

$$H_0 : \sum_{i=1}^a \lambda_i \alpha_i = 4 \text{ and } \sum_{j=1}^b \gamma_j \eta_j = 7,$$

where  $\sum_{i=1}^a \lambda_i = 0$  and  $\sum_{j=1}^b \gamma_j = 0$ . We can write this model in the form  $Y = X\beta + \epsilon$ , (see page 63 of Christensen), where

$\beta = (\mu, \alpha_1, \dots, \alpha_a, \eta_1, \dots, \eta_b)'$ . We have

$$\Lambda' = \begin{pmatrix} 0 & \lambda_1 & \dots & \lambda_a & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 & \gamma_1 & \dots & \gamma_b \end{pmatrix},$$

$$d = \begin{pmatrix} 4 \\ 7 \end{pmatrix}.$$

We also have

$$\Lambda' \hat{\beta} = P' M Y = \begin{pmatrix} \sum_{i=1}^a \lambda_i \bar{y}_{i..} \\ \sum_{j=1}^b \gamma_j \bar{y}_{.j} \end{pmatrix}.$$

# Hypothesis Testing

This result will be proved when we get to Chapter 4. We also have

$$\begin{aligned}\text{Cov}(\Lambda' \hat{\beta}) &= \sigma^2 \Lambda' (X' X)^{-1} \Lambda \\ &= \sigma^2 \begin{pmatrix} \sum_{i=1}^a \frac{\lambda_i^2}{bN} & 0 \\ 0 & \sum_{j=1}^b \frac{\gamma_j^2}{aN} \end{pmatrix}.\end{aligned}$$

Since  $r(\Lambda) = 2$ , the  $F$  statistic is

$$F = \frac{(\Lambda' \hat{\beta} - d)' \hat{\text{Cov}}(\Lambda' \hat{\beta} - d)(\Lambda' \hat{\beta} - d)}{2} \sim F(2, r(I - M), \gamma)$$

where

$$\gamma = \frac{(\Lambda' \beta - d)' (\Lambda' (X' X)^{-1} \Lambda)^{-1} (\Lambda' \beta - d)}{2\sigma^2}.$$

# Hypothesis Testing

## Breaking a sum of squares into independent components

We now present the general theory of breaking up a sum of squares into independent (orthogonal) components. Thus, we want to decompose a quadratic form into a sum of independent quadratic forms, where each quadratic form has one degree of freedom. This decomposition is especially useful in ANOVA models, when testing single degree of freedom contrasts.

To motivate the idea of breaking up sums of squares, we consider the two way ANOVA table without interaction. This model is given by

$$Y_{ijk} = \mu + \alpha_i + \eta_j + \epsilon_{ijk},$$

$i = 1, \dots, a$ ,  $j = 1, \dots, b$ , and  $k = 1, \dots, N$ . Let  $n = abN$ .

# Hypothesis Testing

The ANOVA table is given by

Source	DF	SS	MS
Mean	1	$Y' \left( \frac{J^n}{n} \right) Y$	$Y' \left( \frac{J^n}{n} \right) Y$
Treatments( $\alpha$ )	$a - 1$	$Y' M_\alpha Y$	$\frac{Y' M_\alpha Y}{a-1}$
Treatments( $\eta$ )	$b - 1$	$Y' M_\eta Y$	$\frac{Y' M_\eta Y}{b-1}$
Error	$n - a - b + 1$	$Y'(I - M)Y$	$\frac{Y'(I - M)Y}{n-a-b+1}$
Total	$n$	$Y' Y$	

where  $J_n^n = JJ'$  and  $J = (1, \dots, 1)'$  is the  $n \times 1$  vector of one's. In the table above,  $M_\alpha$  is the orthogonal projection operator onto the column space of  $X_\alpha$ , where  $X_\alpha$  is the design matrix corresponding to the model

$$Y_{ijk} = \mu + \alpha_i + \epsilon_{ijk},$$

and so on. In the ANOVA setting, we often wish to test single degree of freedom contrasts such as

$$\sum_{i=1}^a \lambda_i \alpha_i = 0.$$

# Hypothesis Testing

In order to test such a contrast, we need to break up the  $\alpha$  treatment sums of squares into  $a - 1$  separate components, each having 1 degree of freedom. That is, the quadratic form  $Y' M_\alpha Y$  must be decomposed into

$$Y' M_\alpha Y = \sum_{i=1}^{a-1} Y' M_i Y$$

where each  $M_i$  has rank 1 and  $M_i M_j = 0$  for  $i \neq j$ . Thus in terms of subspaces, we decompose  $C(M_\alpha)$  into a sum of  $a - 1$  orthogonal subspaces each of dimension 1. Thus

$$C(M_\alpha) = C(M_1) + C(M_2) + \dots + C(M_{a-1}).$$

# Hypothesis Testing

We now describe a general procedure for decomposing a subspace into a sum of 1 dimensional orthogonal subspaces.

Consider the linear model

$$Y = X\beta + \epsilon$$

where  $\epsilon \sim N_n(0, \sigma^2 I)$ . Let  $M = X(X'X)^{-}X'$  denote the orthogonal projection onto  $C(X)$ . Let  $M_T$  be any orthogonal projection operator with the property  $C(M_T) \subset C(M)$ . Then  $M_T$  defines a statistic

$$\begin{aligned} F &= \frac{\|M_T Y\|^2 / r(M_T)}{\|(I - M)Y\|^2 / r(I - M)} \\ &= \frac{Y' M_T Y / r(M_T)}{Y' (I - M)Y / r(I - M)} \end{aligned}$$

for testing the reduced model

$$Y = (M - M_T)\gamma_0 + \epsilon .$$

Here,  $C(M - M_T)$  is the estimation space under  $H_0$  and  $C(M_T)$  is the “test” space. The error space is  $C(I - (M - M_T))$ .

# Hypothesis Testing

Suppose  $r(M_T) = r$ , and we want to decompose  $C(M_T)$  into a sum of  $r$  orthogonal subspaces

$$C(M_T) = C(M_1) + \dots + C(M_r)$$

where each  $M_i$  is an orthogonal projection operator of rank 1 and  $M_i M_j = 0$  for  $i \neq j$ .

Suppose  $R = (R_1, \dots, R_r)$  is an orthonormal basis for  $C(M_T)$ , where  $R_i$  is an  $n \times 1$  vector. We have  $r(R_i) = 1$  since  $R_i$  is a vector and  $R'_i R_i = 1$  and  $R'_i R_j = 0$  for  $i \neq j$ . We know by a previous theorem that the orthogonal projection operator onto  $C(M_T)$  is

$$\begin{aligned} M_T &= RR' \\ &= (R_1, \dots, R_r) \begin{pmatrix} R'_1 \\ \vdots \\ R'_r \end{pmatrix} \\ &= \sum_{i=1}^r R_i R'_i \end{aligned}$$

# Hypothesis Testing

Now let  $M_i = R_i R_i'$ . Then by definition,  $M_i$  is an orthogonal projection operator and  $M_i M_j = 0$  for  $i \neq j$ . Thus, the quadratic forms  $Y' M_i Y$  and  $Y' M_j Y$  are independent for each  $i \neq j$ . Since  $M_T = \sum_{i=1}^r M_i$ , we have

$$Y' M_T Y = Y' \left( \sum_{i=1}^r M_i \right) Y = \sum_{i=1}^r Y' M_i Y .$$

Thus, we have

$$\begin{aligned} F &= \frac{\|M_i Y\|^2}{\|(I - M)Y\|^2 / r(I - M)} \\ &= \frac{Y' M_i Y}{Y'(I - M)Y / r(I - M)} \sim F(1, r(I - M), \gamma) , \end{aligned}$$

where  $\gamma = \frac{\|M_i X \beta\|^2}{2\sigma^2}$ .

In one-way ANOVA,  $Y' M_T Y$  corresponds to the treatment sums of squares, while the  $Y' M_i Y$ 's correspond to the sums of squares for a set of orthogonal contrasts.

# Hypothesis Testing

Let us now consider the correspondence between the hypothesis tested using  $Y'M_T Y$  and that using the  $Y'M_i Y$ 's. Since  $M_T$  and the  $M_i$ 's are positive semidefinite, we have

$$0 = \|M_T X \beta\|^2 = \sum_{i=1}^r \|M_i X \beta\|^2$$

if and only if  $\|M_i X \beta\|^2 = 0$  for all  $i = 1, \dots, r$ . Thus, the null hypothesis that corresponds to  $M_T$  is true if and only if the null hypothesis corresponding to ALL of the  $M_i$ 's is true. Equivalently, if the null hypothesis corresponding to  $M_T$  is NOT true, we have

$$0 < \|M_T X \beta\|^2 = \sum_{i=1}^r \|M_i X \beta\|^2.$$

Again, since  $M_T$  and the  $M_i$ 's are positive semidefinite, this occurs if and only if at least one of the  $\|M_i X \beta\|^2 > 0$ . Thus the null hypothesis corresponding to  $M_T$  is NOT true if and only if AT LEAST ONE of the hypotheses corresponding to the  $M_i$ 's is NOT true.

In terms of one-way ANOVA, these results correspond to stating that the alternative hypothesis of a treatment effect is true if and only if at least one contrast in a set of orthogonal contrasts is not 0.

# Hypothesis Testing

## Confidence Regions

Consider the problem of finding a confidence region for an estimable vector  $\Lambda' \beta$ , where  $\Lambda' = P'X$ .

- a) For example, suppose  $\beta_1 - \beta_2$  is estimable and we wish to construct a 95% confidence interval for it. In this case,  $\lambda' = (1, -1, 0, \dots, 0)$ .
- b) Suppose  $\begin{pmatrix} \beta_1 + \beta_3 \\ \beta_2 \end{pmatrix}$  is estimable and we wanted to find a 95% joint confidence region for these parameters. Here,

$$\Lambda' = \begin{pmatrix} 1 & 0 & 1 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & \dots & 0 \end{pmatrix}.$$

Let  $M_{MP}$  be the orthogonal projection operator onto  $C(MP)$ . Using the theory developed for tests of hypotheses, we have

$$\frac{(Y - X\beta)' M_{MP} (Y - X\beta) / r(M_{MP})}{(Y - X\beta)' (I - M) (Y - X\beta) / r(I - M)} \sim F(r(M_{MP}), r(I - M))$$

# Hypothesis Testing

We note here that the noncentrality parameter is 0 since  $E(Y) = X\beta$ . Moreover,

$$(Y - X\beta)'(I - M)(Y - X\beta) = Y'(I - M)Y$$

so that the denominator above equals MSE. We can write

$$\begin{aligned} & (Y - X\beta)'M_{MP}(Y - X\beta) \\ &= (\Lambda'\hat{\beta} - \Lambda'\beta)'(\Lambda'(X'X)^{-}\Lambda)^{-}(\Lambda'\hat{\beta} - \Lambda'\beta) \end{aligned}$$

Thus, a  $(1 - \alpha) \times 100\%$  confidence region for  $\Lambda'\beta$  is

$$\left\{ \beta : \frac{(\Lambda'\hat{\beta} - \Lambda'\beta)'(\Lambda'(X'X)^{-}\Lambda)^{-}(\Lambda'\hat{\beta} - \Lambda'\beta)/r(\Lambda)}{MSE} \leq c_\alpha \right\}$$

where  $c_\alpha = F(1 - \alpha, r(\Lambda), r(I - M))$  is the upper  $(1 - \alpha) \times 100\%$  point of a central  $F$  distribution with degrees of freedom  $(r(\Lambda), r(I - M))$ . We note here that the confidence region is an  $s$  dimensional ellipsoid, where  $\Lambda'$  is  $s \times p$ .

# Hypothesis Testing

## Special cases

- 1) Suppose  $X$  has rank  $p$  so that  $\beta$  is estimable. In this case,  $\Lambda = I_{p \times p}$ . A  $(1 - \alpha) \times 100\%$  confidence region for  $\beta$  is

$$\left\{ \beta : \frac{(\hat{\beta} - \beta)'(X'X)(\hat{\beta} - \beta)}{p \text{ MSE}} \leq F(1 - \alpha, p, n - p) \right\}$$

- 2) Suppose  $X$  is of full rank  $p$ , and suppose  $Y_f$  is an  $m \times 1$  vector of future responses, taken at covariate values  $X_f$ , where  $X_f$  is  $m \times p$ . Then

$$Y_f = X_f \beta + \epsilon$$

where  $\epsilon \sim N_n(0, \sigma^2 I)$ . A  $(1 - \alpha) \times 100\%$  prediction region for  $Y_f$  is

$$\left\{ Y_f : \frac{(\hat{Y}_f - Y_f)'(X_f(X'X)^{-1}X_f' + I)^{-1}(\hat{Y}_f - Y_f)}{m \text{ MSE}} \leq c_\alpha \right\}$$

where  $c_\alpha = F(1 - \alpha, m, n - p)$ ,  $\hat{Y}_f = X_f \hat{\beta} = X_f(X'X)^{-1}X'Y$  and  $MSE = Y'(I - M)Y/r(I - M)$ .

# Hypothesis Testing

## Hypothesis tests for weighted least squares

Consider the model

$$Y = X\beta + \epsilon \quad (13)$$

where  $\epsilon \sim N_n(0, \sigma^2 V)$ , where  $V$  is a known positive definite matrix. The transformed model is given by

$$Q^{-1}Y = Q^{-1}X\beta + Q^{-1}\epsilon \quad (14)$$

where  $V = QQ'$ , and  $Q$  is nonsingular.

Now consider testing

$$Y = X_0\gamma_0 + \epsilon \quad (15)$$

where  $\epsilon \sim N_n(0, \sigma^2 V)$  and

$$C(X_0) \subset C(X).$$

Consider the transformed model

$$Q^{-1}Y = Q^{-1}X_0\gamma_0 + Q^{-1}\epsilon \quad (16)$$

We test (13) against (15) by testing (14) against (16). To test (14) against (16), we must show that

$$C(Q^{-1}X_0) \subset C(Q^{-1}X).$$

That is, the column space of the transformed reduced model is still contained in the column space of the transformed full model.

# Hypothesis Testing

## Theorem

If  $C(X_0) \subset C(X)$ , and  $Q$  is nonsingular, then

$$C(Q^{-1}X_0) \subset C(Q^{-1}X).$$

## Proof

Suppose  $X_0$  is  $n \times q$ , and  $C(X_0) \subset C(X)$ . Then there exists a  $G$  so that  $G$  is  $p \times q$  and

$$X_0 = XG.$$

If  $v \in C(Q^{-1}X_0)$ , then  $v = Q^{-1}X_0d$  for some  $d$ . Substituting for  $X_0$  gives  $v = Q^{-1}XGd$ , so  $v$  is a linear combination of the columns of  $Q^{-1}X$ , so that  $C(Q^{-1}X_0) \subset C(Q^{-1}X)$ .

For model (14), recall that

$$MSE = \frac{\|Q^{-1}(I - A)Y\|^2}{n - r(X)} = \frac{Y'(I - A)'V^{-1}(I - A)Y}{n - r(X)}$$

Now define

$$A_0 = X_0(X_0'V^{-1}X_0)^{-1}X_0'V^{-1}.$$

$A_0$  is a projection operator onto  $C(X_0)$ . We now have the following result.

# Hypothesis Testing

## Theorem

To test (14) against (16), the test statistic is

$$\begin{aligned} F &= \frac{Y'(A - A_0)'V^{-1}(A - A_0)Y / (r(X) - r(X_0))}{MSE} \\ &\sim F(r(X) - r(X_0), n - r(X), \gamma) \end{aligned}$$

where

$$\gamma = \frac{\beta'X'(A - A_0)'V^{-1}(A - A_0)X\beta}{2\sigma^2}.$$

$\gamma = 0$  if and only if  $E(Y) \in C(X_0)$ , that is, if  $H_0$  is true.

For the weighted least squares model, suppose we wish to test

$$H_0 : \Delta'\beta = 0$$

where  $\Delta'\beta = P'X\beta$  is an estimable vector. The  $F$  statistic for this hypothesis is given by

$$F = \frac{\hat{\beta}'\Delta(\Delta'(X'V^{-1}X)^{-}\Delta)^{-}\Delta'\hat{\beta}/r(\Delta)}{MSE} \sim F(r(\Delta), n - r(X), \gamma)$$

where

$$\gamma = \frac{\beta'\Delta(\Delta'(X'V^{-1}X)^{-}\Delta)^{-}\Delta'\beta}{2\sigma^2},$$

$MSE = \frac{Y'(I-A)'V^{-1}(I-A)Y}{n-r(X)}$ , and  $\Delta\hat{\beta}$  is the unique UMVUE of  $\Delta\beta$ .

# Hypothesis Testing

Writing  $\Delta'\beta = P'X\beta$ , we can rewrite the  $F$  test above as

$$\begin{aligned} F &= \frac{Y'A'P(P'X(X'V^{-1}X)^{-1}X'P)^{-1}P'AY/r(\Lambda)}{MSE} \\ &\sim F(r(A'P), n - r(X), \gamma) \end{aligned}$$

where

$$\gamma = \frac{\beta'X'P(P'X(X'V^{-1}X)^{-1}X'P)^{-1}P'X\beta}{2\sigma^2}.$$

# Likelihood Ratio Tests

## Likelihood Ratio Tests

The  $F$  test for testing nested linear models is equivalent to the likelihood ratio test (LRT). We now give the general definition of LRT.

### Defn

Let  $\Theta$  denote the parameter space, and let  $\Theta_0 \subset \Theta$ . Let  $\theta$  be a vector in  $\Theta$ , and let  $y$  denote the data. The likelihood ratio test (LRT) for testing

$$H_0 : \theta \in \Theta_0$$

$$H_a : \theta \in \Theta_0^c$$

is given by

$$\lambda(y) = \frac{\sup_{\Theta_0} L(\theta|y)}{\sup_{\Theta} L(\theta|y)}$$

The LRT is any test that has a rejection region of the form  $\{y : \lambda(y) \leq c\}$ , where  $c$  is any number satisfying  $0 \leq c \leq 1$ .

# Likelihood Ratio Tests

## Example

Consider the linear model

$$Y = X\beta + \epsilon,$$

where  $\epsilon \sim N_n(0, \sigma^2 I)$ . Let  $C(X_0) \subset C(X)$  and suppose we wish to test

$$\begin{aligned} H_0 & : E(Y) \in C(X_0) \\ H_a & : E(Y) \in C(X) \cap C(X_0)^c. \end{aligned}$$

The model under  $H_0$  is  $Y = X_0\gamma_0 + \epsilon$ . We want to derive the likelihood ratio test for this hypothesis. The likelihood function under  $H_0$  is given by

$$L(\gamma_0, \sigma | Y) = \sigma^{-n} \exp \left\{ \frac{-1}{2\sigma^2} (Y - X_0\gamma_0)'(Y - X_0\gamma_0) \right\}.$$

To get the numerator of the LRT, we need to maximize this likelihood with respect to  $(\gamma_0, \sigma)$ . From previous results, we know that the maximizer of  $\gamma_0$  satisfies

$$X_0 \hat{\gamma}_0 = M_0 Y$$

where  $M_0 = X_0(X_0'X_0)^{-1}X_0'$ . Let  $\hat{\sigma}_0^2$  denote the maximizer of  $\sigma^2$  under  $H_0$ . We know from previous results that

$$\hat{\sigma}_0^2 = \frac{Y'(I - M_0)Y}{n}.$$

# Likelihood Ratio Tests

To compute the denominator of the LRT, we compute the supremum of the likelihood over the entire parameter space, i.e., under the full model  $Y = X\beta + \epsilon$ . Thus, the likelihood function under  $H_a$  is

$$L(\beta, \sigma | Y) = \sigma^{-n} \exp \left\{ \frac{-1}{2\sigma^2} (Y - X\beta)'(Y - X\beta) \right\}$$

for which  $X\hat{\beta} = MY$  and

$$\hat{\sigma}^2 = \frac{Y'(I - M)Y}{n}.$$

# Likelihood Ratio Tests

Thus,

$$\begin{aligned}\lambda(y) &= \frac{\sup_{(\gamma_0, \sigma)} L(\gamma_0, \sigma | Y)}{\sup_{(\beta, \sigma)} L(\beta, \sigma | Y)} \\ &= \frac{(\hat{\sigma}_0^2)^{-n/2} \exp \left\{ \frac{-1}{2\hat{\sigma}_0^2} (Y - X_0 \hat{\gamma}_0)'(Y - X_0 \hat{\gamma}_0) \right\}}{(\hat{\sigma}^2)^{-n/2} \exp \left\{ \frac{-1}{2\hat{\sigma}^2} (Y - X \hat{\beta})'(Y - X \hat{\beta}) \right\}} \\ &= \left( \frac{Y'(I - M)Y}{Y'(I - M_0)Y} \right)^{n/2} \frac{\exp \left\{ \frac{-1}{2\hat{\sigma}_0^2} (Y - M_0 Y)'(Y - M_0 Y) \right\}}{\exp \left\{ \frac{-1}{2\hat{\sigma}^2} (Y - M Y)'(Y - M Y) \right\}} \\ &= \left( \frac{Y'(I - M)Y}{Y'(I - M_0)Y} \right)^{n/2} \times \\ &\quad \frac{\exp \left\{ \frac{-n}{2Y'(I - M_0)Y} (Y - M_0 Y)'(Y - M_0 Y) \right\}}{\exp \left\{ \frac{-n}{2Y'(I - M)Y} (Y - M Y)'(Y - M Y) \right\}} \\ &= \left( \frac{Y'(I - M)Y}{Y'(I - M_0)Y} \right)^{n/2} \frac{\exp \{-n/2\}}{\exp \{-n/2\}} \\ &= \left( \frac{Y'(I - M)Y}{Y'(I - M_0)Y} \right)^{n/2}\end{aligned}$$

# Likelihood Ratio Tests

Thus, we reject  $H_0$  if

$$\left( \frac{Y'(I - M)Y}{Y'(I - M_0)Y} \right)^{n/2} \leq c$$

Taking both sides to the  $2/n$  power, This is equivalent to rejecting  $H_0$  when

$$\frac{Y'(I - M)Y}{Y'(I - M_0)Y} \leq c_1 . \quad (1)$$

Now write  $I - M_0 = (I - M) + (M - M_0)$  so that

$$\begin{aligned} Y'(I - M_0)Y &= Y'[(I - M) + (M - M_0)]Y \\ &= Y'(I - M)Y + Y'(M - M_0)Y. \end{aligned}$$

Thus (1) becomes

$$\frac{Y'(I - M)Y}{Y'(I - M)Y + Y'(M - M_0)Y} \leq c_1 \quad (2)$$

Now (2) is equivalent to

$$\frac{Y'(I - M)Y + Y'(M - M_0)Y}{Y'(I - M)Y} \geq c_2 \quad (3)$$

where  $c_2 = 1/c_1$ .

# Likelihood Ratio Tests

We can write (3) as

$$1 + \frac{Y'(M - M_0)Y}{Y'(I - M)Y} \geq c_2 ,$$

which is equivalent to

$$\frac{Y'(M - M_0)Y}{Y'(I - M)Y} \geq c_3$$

where  $c_3 = c_2 - 1$ . Finally, the expression above is equivalent to

$$\frac{Y'(M - M_0)Y/r(M - M_0)}{Y'(I - M)Y/r(I - M)} \geq c_4 \quad (4)$$

We see that the likelihood ratio test is equivalent to the  $F$  test, since the statistic in (4) is precisely the  $F$  statistic derived earlier.

Note: please do recommended problem 3.9.2

# One-Way ANOVA

## One-Way ANOVA

We want to examine the theory for the one-way ANOVA model. A one-way ANOVA model can be written as

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \quad (5)$$

where  $i = 1, \dots, t$ ,  $j = 1, \dots, n_i$ . We will assume that the  $\epsilon_{ij}$ 's are *i.i.d.* and  $\epsilon_{ij} \sim N_n(0, \sigma^2)$ . We let  $n = \sum_{i=1}^t n_i$ .

Our main goal here will be to test hypotheses and construct confidence regions for estimable functions.

We can write the model in (5) as

$$Y = X\beta + \epsilon$$

where  $\epsilon \sim N_n(0, \sigma^2 I)$ . We use the convention that we let the last most subscript change the fastest when writing ANOVA models in matrix form.

# One-Way ANOVA

We want to construct the orthogonal projection operator onto  $C(X)$ , denoted as usual by  $M$ .

The design matrix for the model in (5) is given by

$$X = (J, X_1, \dots, X_t)$$

where  $J$  is the  $n \times 1$  vector of ones and  $X_k$  is an  $n \times 1$  vector of zeroes and ones. We can write  $X_k$  as

$$X_k = (t_{ij})$$

where  $t_{ij} = \delta_{ik}$  with

$$\delta_{ik} = \begin{cases} 0 & \text{if } i \neq k \\ 1 & \text{if } i = k \end{cases}$$

Thus, if the observation in the  $ijth$  row got the  $kth$  treatment, the  $ijth$  row of  $X_k$  is 1, otherwise it is 0. Thus  $X_k$  has exactly  $n_k$  ones and  $n - n_k$  zeroes.

Note that  $X$  is not of full rank since

$$J = X_1 + \dots + X_t .$$

However, the  $X_k$ 's are orthogonal,  $k = 1, \dots, t$ .

# One-Way ANOVA

Let

$$Z = (X_1, \dots, X_t)$$

Thus,  $Z$  is of full rank  $t$  since the  $X_k$ 's are orthogonal. Moreover, it is clear that

$$C(Z) = C(X),$$

and thus,

$$M = Z(Z'Z)^{-1}Z'.$$

Also, since  $r(Z) = t$ , we have  $r(M) = t$ . Hence

$$Z'Z = \text{diag}(n_1, \dots, n_t).$$

To see this, note that the  $ij$ th element of  $Z'Z$  is  $X_i'X_j = 0$  and the  $i$ th diagonal element of  $Z'Z$  is  $X_i'X_i = n_i$ .

# One-Way ANOVA

Now

$$\begin{aligned} Z(Z'Z)^{-1}Z' &= (X_1, \dots, X_t) \text{diag}(n_1^{-1}, \dots, n_t^{-1}) \begin{pmatrix} X'_1 \\ \vdots \\ X'_t \end{pmatrix} \\ &= (n_1^{-1}X_1, \dots, n_t^{-1}X_t) \begin{pmatrix} X'_1 \\ \vdots \\ X'_t \end{pmatrix} \\ &= n_1^{-1}X_1X'_1 + n_2^{-1}X_2X'_2 + \dots + n_t^{-1}X_tX'_t \\ &= \begin{pmatrix} n_1^{-1}J_{n_1}^{n_1} & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \end{pmatrix} + \dots + \begin{pmatrix} 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ 0 & \dots & 0 & n_t^{-1}J_{n_t}^{n_t} \end{pmatrix} \\ &= \begin{pmatrix} n_1^{-1}J_{n_1}^{n_1} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & n_t^{-1}J_{n_t}^{n_t} \end{pmatrix}_{n \times n} \end{aligned}$$

where

$$J_{n_i}^{n_i} = \begin{pmatrix} 1 & \dots & 1 \\ \vdots & \vdots & \vdots \\ 1 & \dots & 1 \end{pmatrix}_{n_i \times n_i}$$

is the  $n_i \times n_i$  matrix of ones.

# One-Way ANOVA

Thus  $M$  is an  $n \times n$  block diagonal matrix, with the  $i$ th block being  $n_i^{-1} J_{n_i}^{n_i}$ .

Thus,

$$M = \text{blkdiag}(n_1^{-1} J_{n_1}^{n_1}, \dots, n_t^{-1} J_{n_t}^{n_t}) .$$

Thus,  $M$  represents the orthogonal projection operator for the “full” model.

Suppose we wish to test no treatment effect. The null hypothesis for no treatment effect may be written as

$$H_0 : \alpha_1 = \alpha_2 = \dots = \alpha_t .$$

The model under the null hypothesis is given by

$$Y_{ij} = \mu + \epsilon_{ij} .$$

The design matrix for this model is  $X_0 = J$ , and therefore, the orthogonal projection operator onto  $C(J)$  is

$$M_\mu = \frac{1}{n} JJ' = \frac{1}{n} J_n^n .$$

The orthogonal projection operator for the treatments subspace can now be found by subtraction. Thus

$$M_\alpha = M - M_\mu .$$

# One-Way ANOVA

Thus we see that we have the decomposition

$$\begin{aligned} M &= M_\mu + M_\alpha \\ &= M_\mu + (M - M_\mu) . \end{aligned}$$

Moreover,  $C(M_\alpha)$  and  $C(M_\mu)$  are orthogonal. Thus

$$\begin{aligned} M_\mu M_\alpha &= M_\alpha M_\mu \\ &= (M - M_\mu) M_\mu \\ &= MM_\mu - M_\mu \\ &= M_\mu - M_\mu = 0 . \end{aligned}$$

We also have  $r(M_\mu) = 1$  and  $r(M) = t$ , so that

$r(M_\alpha) = r(M) - r(M_\mu) = t - 1$ . As usual, the error space is  $C(I - M)$ , and  $r(I - M) = n - t$ . We can now write the one-way ANOVA table as

Source	DF	SS	MS
Mean	1	$Y'(n^{-1}J_n^n)Y$	$Y'M_\mu Y$
Treatments( $\alpha$ )	$t - 1$	$Y'M_\alpha Y$	$\frac{Y'M_\alpha Y}{t-1}$
Error	$n - t$	$Y'(I - M)Y$	$\frac{Y'(I - M)Y}{n-t}$
Total	$n$	$Y'Y$	

# One-Way ANOVA

We notice here that analysis of variance is a decomposition of  $R^n$  into a sum of orthogonal subspaces. The degrees of freedom in the ANOVA table correspond to the dimensions of the space. For one-way ANOVA, we have

$$\begin{aligned} R^n &= C(M) + C(I - M) \\ &= C(M_\mu) + C(M_\alpha) + C(I - M) \end{aligned}$$

The first two terms in the last equation are a decomposition of the estimation space  $C(M)$  into sum of two orthogonal subspaces, and the last term is the error space. We can also write

$$I_{n \times n} = M_\mu + M_\alpha + (I - M) .$$

Thus, the one-way ANOVA model sheds light on decompositions for higher way ANOVA. The basic ideas are the same. For  $k$ -way ANOVA, the estimation space  $C(M)$  can be decomposed in a similar form. For example, for a two-way ANOVA model with interaction and replication, we have

$$Y_{ijk} = \mu + \alpha_i + \eta_j + \gamma_{ij} + \epsilon_{ijk} .$$

# One-Way ANOVA

Here, the estimation space can be decomposed into a sum of four orthogonal subspaces. We have

$$C(M) = C(M_\mu) + C(M_\alpha) + C(M_\eta) + C(M_\gamma)$$

where  $C(M_\alpha)$  is the  $\alpha$  treatment space,  $M_\eta$  is the  $\eta$  treatment space and  $M_\gamma$  is the interaction space. Shortly, we see how to construct these orthogonal projection operators.

The expected mean squares are the expectations of the quadratic forms in an ANOVA table divided by their degrees of freedom. For the one-way ANOVA, We have

$$E(Y' M_\mu Y) = \sigma^2 + \|M_\mu X \beta\|^2 ,$$

$$E\left(\frac{Y' M_\alpha Y}{t-1}\right) = \sigma^2 + \frac{\|M_\alpha X \beta\|^2}{t-1} ,$$

and

$$E\left(\frac{Y'(I-M)Y}{n-t}\right) = \sigma^2 .$$

Expected mean squares are useful quantities in random effects and mixed models, because they determine the numerator for the  $F$  test.

# One-Way ANOVA

## Estimation of Parameters

Consider the one-way ANOVA model

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij},$$

where the  $\epsilon_{ij}$ 's are *i.i.d.*  $N(0, \sigma^2)$  random variables. For the one-way ANOVA model,

$$E(Y_{ij}) = \mu + \alpha_i.$$

Since  $E(Y)$  is ALWAYS estimable for ANY linear model, we have  $\mu + \alpha_i$  is estimable for all  $i = 1, \dots, t$ . Writing  $Y$  as  $Y = (Y_{11}, Y_{12}, \dots, Y_{tn_t})'$ , we know that the BLUE of  $E(Y)$  is  $MY$ . This is given by

$$MY = \begin{pmatrix} J_{n_1} \bar{Y}_{1\cdot} \\ J_{n_2} \bar{Y}_{2\cdot} \\ \vdots \\ J_{n_t} \bar{Y}_{t\cdot} \end{pmatrix}$$

where  $J_{n_i}$  is the  $n_i \times 1$  vector of ones.

# One-Way ANOVA

## Remarks

- 1) We will see shortly that the estimable parameters in a one-way ANOVA model are  $\mu + \alpha_i$ , and any contrasts, i.e., functions of the form  $\sum_{i=1}^t \lambda_i \alpha_i$ , where  $\sum_{i=1}^t \lambda_i = 0$ . All other parameters such as  $\mu$ ,  $\alpha_i$  are not estimable.
- 2) Frequently, we know that in ANOVA, “side conditions” are imposed to get “estimates” of nonestimable quantities such as the  $\alpha_i$ ’s. For example, we have all seen

$$\hat{\alpha}_i = \bar{Y}_{i\cdot} - \bar{Y}_{..} .$$

The side conditions needed to produce such estimates are

$$\sum_{i=1}^t n_i \alpha_i = 0 .$$

Side conditions remove the arbitrariness in the estimation of the parameter vector  $\beta = (\mu, \alpha_1, \dots, \alpha_t)'$ . They do so by imposing a nonestimable constraint on the parameters. Initially, the one way ANOVA is overparameterized since  $X$  has less than full rank. The nonestimable constraint is chosen to make  $X$  full rank.

Fundamentally, one choice is as good as any other. For this reason, side conditions are silly. One should not try to estimate functions that are not estimable.

# One-Way ANOVA

## Estimation and Testing of Contrasts

### Defn

A contrast in a one-way ANOVA is a function

$$\sum_{i=1}^t \lambda_i \alpha_i$$

where

$$\sum_{i=1}^t \lambda_i = 0.$$

Write  $\lambda = (0, \lambda_1, \dots, \lambda_t)'$ . If  $\lambda' \beta$  is estimable, then we know that  $\lambda' = \rho' X$ .

The BLUE of  $\lambda' \beta$  is  $\rho' M Y$ . The vector  $M \rho$  is always unique, and for one-way ANOVA, it has the form

$$M \rho = (t_{ij})$$

where  $t_{ij} = \lambda_i / n_i$ . Thus

$$\rho' = \left( \frac{\lambda_1}{n_1} J'_{n_1}, \frac{\lambda_2}{n_2} J'_{n_2}, \dots, \frac{\lambda_t}{n_t} J'_{n_t} \right).$$

Contrasts of the  $\alpha_i$ 's are the set of all estimable functions of the  $\alpha_i$ 's that do not involve  $\mu$ . Since  $J$  is the column of  $X$  associated with  $\mu$ ,  $\rho' X \beta$  does not involve the parameter  $\mu$  if and only if  $\rho' J = 0$ . We are led to the following theorem.

# One-Way ANOVA

## Theorem

$\rho'X\beta$  is a contrast if and only if  $\rho'J = 0$ .

Note that  $\rho'J = 0$  implies that the

$$\sum_{i=1}^n \rho_i = 0 .$$

Contrasts are the estimable functions that impose a constraint on  $C(M_\alpha)$ . To see this, recall that  $\rho'X\beta = 0$  implies  $E(Y) \in C(X) \cap C(M\rho)^\perp$ . By definition,  $\rho'X\beta$  puts a constraint on  $C(M_\alpha)$  if  $M\rho \in C(M_\alpha)$ . We are led to the following theorem.

## Theorem

$\rho'X\beta$  is a contrast if and only if  $M\rho \in C(M_\alpha)$ .

## Proof

since  $J \in C(X)$ ,  $\rho'X\beta$  is a contrast if and only if  $0 = \rho'J = \rho'MJ$ .  $C(M_\alpha)$  is the subspace that is everything in  $C(X)$  that is orthogonal to  $J$ . Thus  $\rho'MJ = 0$  if and only if  $M\rho \in C(M_\alpha)$ .

# One-Way ANOVA

We can now characterize  $C(M_\alpha)$ .

Theorem

$$C(M_\alpha) = \left\{ \rho : \rho = (t_{ij}), t_{ij} = \lambda_i / n_i, \sum_{i=1}^t \lambda_i = 0 \right\} .$$

This theorem says that set of all possible contrasts makes up  $C(M_\alpha)$ .  
Now suppose in the one-way ANOVA, we wanted to test the contrast

$$H_0 : \lambda' \beta = 0$$

where  $\lambda' = (0, \lambda_1, \dots, \lambda_t)$ , and  $\sum_{i=1}^t \lambda_i = 0$ . The BLUE of  $\lambda' \beta$  is

$$\rho' M Y = \sum_{i=1}^t \lambda_i \bar{Y}_i .$$

From previous results, we know the  $F$  test is

$$F = \frac{(\rho' M Y)' (\rho' M \rho)^{-1} (\rho' M Y)}{MSE} .$$

# One-Way ANOVA

This can be simplified as follows:

$$\begin{aligned}(\rho' M \rho)^{-1} &= \left( (M \rho)' (M \rho) \right)^{-1} \\&= \left( \sum_{i=1}^t \sum_{j=1}^{n_i} \lambda_i^2 / n_i \right)^{-1} \\&= \left( \sum_{i=1}^t \lambda_i^2 / n_i \right)^{-1}.\end{aligned}$$

Also,

$$\begin{aligned}(\rho' M Y)' (\rho' M Y) &= (Y' M \rho) (\rho' M Y) \\&= \left( \sum_{i=1}^t \lambda_i \bar{Y}_{i.} \right)^2\end{aligned}$$

Thus, the  $F$  statistic is

$$F = \frac{\left( \sum_{i=1}^t \lambda_i \bar{Y}_{i.} \right)^2}{MSE \left( \sum_{i=1}^t \lambda_i^2 / n_i \right)} \sim F(1, n - t, \gamma)$$

where

$$\gamma = \frac{\left( \sum_{i=1}^t \lambda_i (\mu + \alpha_i) \right)^2}{2\sigma^2 \left( \sum_{i=1}^t \lambda_i^2 / n_i \right)}.$$

Under  $H_0$ ,  $\gamma = 0$ , so that  $F \sim F(1, n - t)$ .

# One-Way ANOVA

## Defn

Consider the One-way ANOVA model

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

where  $j = 1, \dots, n_i$ , and  $i = 1, \dots, t$ . Consider two contrasts  $\lambda'_1 \beta$  and  $\lambda'_2 \beta$ , where  $\lambda'_j = (0, \lambda_{j1}, \dots, \lambda_{jt})$ ,  $j = 1, 2$ . Then  $\lambda'_1 \beta$  and  $\lambda'_2 \beta$  are said to be orthogonal contrasts if

$$\sum_{k=1}^t \frac{\lambda_{1k} \lambda_{2k}}{n_k} = 0 .$$

In one-way ANOVA, the only way to break up the treatment subspace  $C(M_\alpha)$  into a sum of  $t - 1$  orthogonal subspaces is to find  $t - 1$  mutually orthogonal contrasts. Each contrast  $\lambda'_i \beta$ ,  $i = 1, \dots, t - 1$  will have orthogonal projection operator  $M_i$ , where

$$M_\alpha = \sum_{i=1}^{t-1} M_i .$$

# One-Way ANOVA

Write  $\lambda_i = \rho'_i X$ , where  $\lambda'_i = (0, \lambda_{i1}, \dots, \lambda_{it})$ . Recall that  $M\rho_i$  has the structure

$$M\rho_i = (t_{ik})$$

where  $t_{jk} = \lambda_{ij}/n_j$ . Thus

$$\begin{aligned} 0 &= \rho'_i M\rho_j \\ &= (M\rho_i)'(M\rho_j) \\ &= \sum_{k=1}^t \sum_{l=1}^{n_k} \left( \frac{\lambda_{ik}}{n_k} \right) \left( \frac{\lambda_{jk}}{n_k} \right) \\ &= \sum_{k=1}^t \frac{\lambda_{ik}\lambda_{jk}}{n_k}. \end{aligned}$$

Thus, for any set of contrasts,  $\sum_{j=1}^t \lambda_{ij}\alpha_j$ ,  $i = 1, \dots, t - 1$  for which

$$\sum_{k=1}^t \frac{\lambda_{ik}\lambda_{jk}}{n_k} = 0 \quad \text{for all } i \neq j,$$

implies that we have a set of  $t - 1$  mutually orthogonal contrasts. From the result above we see that an equivalent condition that two contrasts  $\lambda'_i\beta$  and  $\lambda'_j\beta$  are orthogonal is

$$\rho'_i M\rho_j = 0.$$

# One-Way ANOVA

Since  $M_\alpha = M - M_\mu$ , we have

$$\rho' MY = \rho'(M_\alpha + M_\mu)Y = \rho' M_\alpha Y, \text{ and}$$

$$\rho' M\rho = \rho' M_\alpha \rho.$$

Since  $\rho' M\rho = \rho' M_\alpha \rho$ , this implies that two contrasts are orthogonal if and only if

$$\rho_i' M_\alpha \rho_j = 0.$$

Estimates and quantities needed for tests of hypotheses depend only on the orthogonal projection operator  $M_\alpha$ . Moreover, the condition required for contrasts to give an orthogonal breakdown of the treatment sums of squares is that they must be orthogonal. That is

$$\begin{aligned} 0 &= \sum_{k=1}^t \frac{\lambda_{ik}\lambda_{jk}}{n_k} \\ &= \rho_i' M \rho_j \\ &= \rho_i' M_\alpha \rho_j. \end{aligned}$$

# One-Way ANOVA

We note here that the orthogonal projection operator for a contrast  $\rho_i'X\beta$  is given by

$$\begin{aligned}M_i &= \frac{(M\rho_i)(M\rho_i)'}{\rho_i'M\rho_i} \\&= \frac{M\rho_i\rho_i'M}{\rho_i'M\rho_i} \\&= \frac{\rho_i\rho_i'}{\rho_i'\rho_i}, \quad \text{since } \rho_i \in C(M).\end{aligned}$$

# One-Way ANOVA

## Theorem

If we choose

$$\rho' = \left( \frac{\lambda_1}{n_1} J'_{n_1}, \frac{\lambda_2}{n_2} J'_{n_2}, \dots, \frac{\lambda_t}{n_t} J'_{n_t} \right), \quad (6)$$

then

$$\rho \in C(X)$$

so that

$$M\rho = M_\alpha \rho = \rho.$$

We note that this theorem has a lot of implications. For example, if  $\rho$  is of the form in (6), then

- i)  $\rho' M Y = \rho' Y = \sum_{i=1}^t \lambda_i \bar{Y}_i ..$
- ii)  $\rho' M \rho = \rho' \rho = \sum_{i=1}^t \lambda_i^2 / n_i.$
- iii) The orthogonal projection operator  $M_i$  for a contrast  $\rho'_i X \beta$  is

$$M_i = \frac{\rho_i \rho'_i}{\rho'_i \rho_i}$$

where  $\rho_i$  has the form in (6).

# One-Way ANOVA

- iv) The  $F$  test for the hypothesis  $H_0 : \rho_i' X \beta = 0$  takes the form

$$\begin{aligned} F &= \frac{\|M_i Y\|^2}{MSE} \\ &= \frac{(\rho_i' Y)^2}{(\rho_i' \rho_i) MSE} \end{aligned}$$

where  $MSE = \|(I - M)Y\|^2 / (n - t)$ .

# Multifactor Analysis of Variance

## Multifactor Analysis of Variance

We want to examine the theory of two-way and higher way ANOVA. Let us first consider the two-way balanced ANOVA model without interaction. This model can be written as

$$Y_{ijk} = \mu + \alpha_i + \eta_j + \epsilon_{ijk} \quad (7)$$

for  $i = 1, \dots, a$ ,  $j = 1, \dots, b$ , and  $k = 1, \dots, N$ . We have a total of  $n = abN$  observations. In this model, we have two factors (treatments), and there are a total of  $ab$  treatment combinations. We have  $N$  observations per treatment combination. These are sometimes referred to as replicates. Thus we have an equal number of replicates  $N$  per treatment combination. We can write model (7) in the form  $Y = X\beta + \epsilon$ . For example, when  $a = 3$ ,  $b = 2$ , and  $N = 4$ , we have the design matrix given on page 144 of Christensen.

In general, we can write the design matrix of the model in (7) as

$$X = (J, X_1, \dots, X_a, X_{a+1}, \dots, X_{a+b})_{n \times (a+b+1)}$$

where  $X_r = (t_{ijk})$ , where  $t_{ijk} = \delta_{ir}$ ,  $r = 1, \dots, a$ ,  $X_s = (t_{ijk})$ , where  $t_{ijk} = \delta_{j(s-a)}$ , for  $s = a+1, \dots, a+b$ . Thus each  $X_i$  is an  $n \times 1$  vector. Here, we define  $\delta_{gh} = 1$  if  $g = h$  and 0 otherwise.

# Two-Way ANOVA

The  $\beta$  vector is given by

$$\beta = (\mu, \alpha_1, \dots, \alpha_a, \eta_1, \dots, \eta_b)'$$

which is an  $(a + b + 1) \times 1$  vector.

We want to break up the estimation space into a sum of orthogonal subspaces, with each subspace corresponding to one of the effects. That is we want to write

$$C(M) = C(M_\mu) + C(M_\alpha) + C(M_\eta) .$$

Define

$$Z = (J, Z_1, \dots, Z_a, Z_{a+1}, \dots, Z_{a+b})$$

where

$$Z_r = X_r - \frac{X'_r J}{J' J} J \quad r = 1, \dots, a + b .$$

Thus, we have used Gram-Schmidt on the  $X$ 's to eliminate the effect of  $\mu$  from the remaining columns of  $X$ . Thus  $Z_r \perp J$  for  $r = 1, \dots, a + b$ . Since  $J' J = abN = n$ , we have  $X'_r J = bN$  for  $r = 1, \dots, a$ , and  $X'_s J = aN$  for  $s = a + 1, \dots, a + b$ . Thus we have

$$Z_r = X_r - \frac{1}{a} J \quad r = 1, \dots, a$$

and

$$Z_s = X_s - \frac{1}{b} J \quad s = a + 1, \dots, a + b .$$

# Two-Way ANOVA

Now we have

$$C(X) = C(J, X_1, \dots, X_{a+b}) = C(J, Z_1, \dots, Z_{a+b}) = C(Z),$$

$$C(J, X_1, \dots, X_a) = C(J, Z_1, \dots, Z_a)$$

and

$$C(J, X_{a+1}, \dots, X_{a+b}) = C(J, Z_{a+1}, \dots, Z_{a+b}).$$

We can see that  $J \perp Z_r$ ,  $r = 1, \dots, a+b$ . Also

$$C(Z_1, \dots, Z_a) \perp C(Z_{a+1}, \dots, Z_{a+b}). \quad (8)$$

To see (8), we observe that for  $r = 1, \dots, a$ ,  $s = a+1, \dots, a+b$ ,

$$\begin{aligned} Z'_s Z_r &= \sum_{ijk} (\delta_{j(s-a)} - \frac{1}{b})(\delta_{ir} - \frac{1}{a}) \\ &= \sum_{ijk} \delta_{j(s-a)} \delta_{ir} - \sum_{ijk} \delta_{j(s-a)} \frac{1}{a} - \sum_{ijk} \delta_{ir} \frac{1}{b} + \frac{abN}{ab} \\ &= N - \frac{aN}{a} - \frac{bN}{b} + N = 0 \end{aligned}$$

# Two-Way ANOVA

Thus, we have decomposed  $C(X) = C(M)$  into three orthogonal subspaces.

- i)  $C(M_\mu) = C(J)$
- ii)  $C(M_\alpha) = C(Z_1, \dots, Z_a)$
- iii)  $C(M_\eta) = C(Z_{a+1}, \dots, Z_{a+b})$ .

Letting  $Z_\alpha = (Z_1, \dots, Z_a)$  and  $Z_\eta = Z_{a+1}, \dots, Z_{a+b}$ , we have

$$M_\alpha = Z_\alpha (Z'_\alpha Z_\alpha)^{-1} Z'_\alpha$$

$$M_\eta = Z_\eta (Z'_\eta Z_\eta)^{-1} Z'_\eta .$$

Thus

$$M = M_\mu + M_\alpha + M_\eta$$

# Two-Way ANOVA

Note that  $r(Z_\alpha) = a - 1$  since  $\sum_{i=1}^a Z_i = J - J = 0$ , thus  $Z_\alpha$  has  $a - 1$  linearly independent columns. Also  $r(Z_\eta) = b - 1$ , since  $\sum_{j=a+1}^{a+b} Z_i = J - J = 0$ , thus  $Z_\eta$  has  $b - 1$  linearly independent columns. Thus,  $r(M_\alpha) = a - 1$  and  $r(M_\eta) = b - 1$ . We note here that  $C(J, X_1, \dots, X_a)$  is the column space for the one-way ANOVA model

$$Y_{ijk} = \mu + \alpha_i + \epsilon_{ijk}$$

where there are  $bN$  observations per treatment, and we have  $a$  treatments. Also,  $C(J, X_{a+1}, \dots, X_{a+b})$  is the column space for the one-way ANOVA model

$$Y_{ijk} = \mu + \eta_j + \epsilon_{ijk}$$

where there are  $aN$  observations per treatment and there are  $b$  treatments. The point is that a two-way ANOVA can be viewed as two one-way ANOVA's. That is, one can obtain the orthogonal projection operators  $M_\alpha$  and  $M_\eta$  based on two one-way ANOVA models.

# Two-Way ANOVA

We can write the ANOVA table for the two-way model in (7) as

Source	DF	SS	MS
Mean	1	$Y' \left( \frac{J\bar{n}}{n} \right) Y$	$Y' \left( \frac{J\bar{n}}{n} \right) Y$
Treatments( $\alpha$ )	$a - 1$	$Y' M_\alpha Y$	$\frac{Y' M_\alpha Y}{a-1}$
Treatments( $\eta$ )	$b - 1$	$Y' M_\eta Y$	$\frac{Y' M_\eta Y}{b-1}$
Error	$n - a - b + 1$	$Y'(I - M)Y$	$\frac{Y'(I - M)Y}{n-a-b+1}$
Total	$n$	$Y' Y$	

The expected mean squares are given by

$$E(Y' M_\mu Y) = \sigma^2 + n(\mu + \bar{\alpha}_. + \bar{\eta}_.)^2$$

$$E\left(\frac{Y' M_\alpha Y}{a-1}\right) = \sigma^2 + \frac{bN}{a-1} \sum_{i=1}^a (\alpha_i - \bar{\alpha}_.)^2$$

$$E\left(\frac{Y' M_\eta Y}{b-1}\right) = \sigma^2 + \frac{aN}{b-1} \sum_{j=1}^b (\eta_j - \bar{\eta}_.)^2$$

$$E\left(\frac{Y'(I - M)Y}{n-a-b+1}\right) = \sigma^2 .$$

## Contrasts for Two-way ANOVA

Estimation and testing of contrasts in two-way ANOVA is done in exactly the same way as in one-way ANOVA. This will follow from the fact that a contrast in the  $\alpha_i$ 's involves a constraint on  $C(M_\alpha)$  and  $C(M_\alpha)$  does not involve the  $\eta_j$ 's.

# Two-Way Anova

## Theorem

Let  $\lambda' \beta$  be estimable and let  $\lambda' = \rho' X$ . Then  $\lambda' \beta$  is a contrast in the  $\alpha_i$ 's if and only if

$$M\rho = M_\alpha \rho.$$

In this case, the estimate of  $\lambda' \beta$  is  $\rho' M_\alpha Y$ , which is the estimate from the one-way ANOVA ignoring the  $\eta_j$ 's.

Suppose we have a contrast

$$\sum_{i=1}^a c_i \alpha_i$$

where  $\sum_{i=1}^a c_i = 0$ . Consider writing this contrast in the form  $\rho' X \beta$ , where

$$\rho' = \frac{1}{bN} (J'_{bN} c_1, J'_{bN} c_2, \dots, J'_{bN} c_a).$$

Then,

- i)  $M\rho = M_\alpha \rho = \rho$  since  $\rho \in C(M_\alpha)$ .
- ii) The estimate of the contrast is  $\rho' M Y = \rho' M_\alpha Y = \rho' Y = \sum_{i=1}^a c_i \bar{Y}_i$ .
- iii)  $\text{Var}(\rho' M Y) = \sigma^2 \rho' M \rho = \sigma^2 \rho' \rho = \sigma^2 \frac{1}{bN} \sum_{i=1}^a c_i^2$ .
- iv) The  $F$  statistic for testing  $H_0 : \sum_{i=1}^a c_i \alpha_i = 0$  is

$$F = \frac{(\rho' Y)^2}{(\rho' \rho) MSE}$$

## Two-Way ANOVA

To get orthogonal contrasts  $\rho_1' X \beta = \sum_{i=1}^a c_{1i} \alpha_i$  and  $\rho_2' X \beta = \sum_{i=1}^a c_{2i} \alpha_i$ , the requirement is

$$\rho_1' M_\alpha \rho_2 = 0 \Leftrightarrow \sum_{i=1}^a c_{1i} c_{2i} = 0.$$

Using the exact same techniques, similar results can be derived for contrasts in the  $\eta_j$ 's.

# Balanced Two-way ANOVA with interaction

## Balanced Two-way ANOVA with Interaction

Consider the model

$$Y_{ijk} = \mu + \alpha_i + \eta_j + \gamma_{ij} + \epsilon_{ijk} \quad (1)$$

where  $i = 1, \dots, a$ ,  $j = 1, \dots, b$ , and  $k = 1, \dots, N$ . We assume the  $\epsilon_{ijk}$ 's are i.i.d.  $N(0, \sigma^2)$  random variables. The term  $\gamma_{ij}$  is called an interaction term. When the interaction term is constant for all  $(i, j)$ , then the model is additive in the treatment effects  $\alpha_i$  and  $\eta_j$ . To better understand what an interaction term is, we consider the means model formulation. The means model for a two-way ANOVA is given by

$$Y_{ijk} = \mu_{ij} + \epsilon_{ijk} .$$

The  $\gamma_{ij}$ 's are constant for all  $(i, j)$  (i.e., no interaction) if and only if any of the following conditions hold.

- i)  $\mu_{ij} = a_i + b_j$ .
- ii)  $\mu_{ij} - \mu_{i'j}$  is independent of  $j$  for all  $i$  and  $i'$ . This means that the differences in the means are constant across rows.
- iii)  $\mu_{ij} - \mu_{ij'}$  is independent of  $i$  for all  $j$  and  $j'$ . This means that the differences in the means is constant across columns.
- iv)  $\mu_{ij} - \mu_{ij'} - \mu_{i'j} + \mu_{i'j'}$  is constant for all  $(i, i', j, j')$ .

# Balanced Two-way ANOVA with interaction

To get the overparameterized model in (1), we model  $\mu_{ij}$  as follows.

$$\begin{aligned}\mu_{ij} &= \bar{\mu}_{..} + (\bar{\mu}_{i.} - \bar{\mu}_{..}) + (\bar{\mu}_{.j} - \bar{\mu}_{..}) + \\ &\quad (\mu_{ij} - \bar{\mu}_{i.} - \bar{\mu}_{.j} + \bar{\mu}_{..}) \\ &= \mu + \alpha_i + \eta_j + \gamma_{ij}\end{aligned}$$

Note here that

$$\mu_{ij} - \bar{\mu}_{i.} - \bar{\mu}_{.j} + \bar{\mu}_{..} = \mu_{ij} - (\bar{\mu}_{i.} - \bar{\mu}_{..}) - (\bar{\mu}_{.j} - \bar{\mu}_{..}) - \bar{\mu}_{..}$$

We can write the model in (1) in the form  $Y = X\beta + \epsilon$ , where

$$X = (J, X_1, \dots, X_a, X_{a+1}, \dots, X_{a+b}, X_{a+b+1}, \dots, X_{a+b+ab})$$

and

$$\beta = (\mu, \alpha_1, \dots, \alpha_a, \eta_1, \dots, \eta_b, \gamma_{11}, \dots, \gamma_{ab})'$$

$X$  is an  $n \times (1 + a + b + ab)$  matrix,  $\beta$  is an  $(1 + a + b + ab) \times 1$  vector, and  $n = abN$ .

## Balanced Two-way ANOVA with interaction

The columns of  $X$  that correspond to the  $\gamma_{ij}$ 's are the vectors

$X_{a+b+1}, \dots, X_{a+b+ab}$ . We can reindex these as

$X_{(1,1)}, X_{(1,2)}, \dots, X_{(1,b)}, X_{(2,1)}, \dots, X_{(a,b)}$ , and thus  $X_{(i,j)}$  is the column corresponding to  $\gamma_{ij}$ . We can write

$$X_{(r,s)} = (t_{ijk}) \quad \text{where} \quad t_{ijk} = \delta_{(i,j)(r,s)}$$

and  $\delta_{(i,j)(r,s)} = 1$  if  $(i,j) = (r,s)$  and 0 otherwise. We can now see that

$$J = \sum_{i=1}^a \sum_{j=1}^b X_{(i,j)}$$

$$X_r = \sum_{j=1}^b X_{(r,j)} , \quad r = 1, \dots, a$$

$$X_s = \sum_{i=1}^a X_{(i,s-a)} , \quad s = a+1, \dots, a+b .$$

These equations imply that

$$\begin{aligned} C(X) &= C(X_{(1,1)}, X_{(1,2)}, \dots, X_{(a,b)}) \\ &= C(X_{a+b+1}, \dots, X_{a+b+ab}) . \end{aligned}$$

Thus the columns of  $X$  responsible for  $\gamma_{ij}$  actually make up  $C(X)$ .

## Balanced Two-way ANOVA with interaction

Again, it is our goal to break up the estimation space into a sum of four orthogonal subspaces. That is, we want to write

$$C(M) = C(M_\mu) + C(M_\alpha) + C(M_\eta) + C(M_\gamma).$$

Since these spaces are orthogonal, we can obtain  $M_\alpha$  and  $M_\eta$  from the one-way ANOVA models. We then find  $M_\gamma$  by subtraction. That is

$$M_\gamma = M - M_\mu - M_\alpha - M_\eta.$$

Note here that  $M$  can be obtained from  $X_{(1,1)}, X_{(1,2)}, \dots, X_{(a,b)}$ . These  $ab$  vectors are linearly independent. If we let  $Z_\gamma = (X_{(1,1)}, X_{(1,2)}, \dots, X_{(a,b)})$ , then  $Z_\gamma$  is an  $n \times ab$  full rank matrix. Therefore,

$$M = Z_\gamma (Z_\gamma' Z_\gamma)^{-1} Z_\gamma' .$$

Since  $Z_\gamma$  has rank  $ab$ ,  $M$  has rank  $ab$ . It can now easily be shown that  $M$  is a block diagonal matrix of with each block consisting of  $\frac{1}{N} J_N^N$ , and there are  $ab$  such blocks. Thus

$$M = \text{Blkdiag} \left( N^{-1} J_N^N \right) .$$

# Balanced Two-way ANOVA with interaction

The column space

$$C(M - M_\mu - M_\alpha - M_\eta)$$

will be called the column space for the interaction subspace. This interaction subspace has dimension

$$\begin{aligned} & r(M - M_\mu - M_\alpha - M_\eta) \\ = & r(M) - r(M_\mu) - r(M_\alpha) - r(M_\eta) \\ = & ab - 1 - (a - 1) - (b - 1) \\ = & (a - 1)(b - 1). \end{aligned}$$

The quadratic form (i.e., sums of squares) for the interaction effect is given by

$$Y'(M - M_\mu - M_\alpha - M_\eta)Y$$

and its expected value is

$$\begin{aligned} & E(Y'(M - M_\mu - M_\alpha - M_\eta)Y) \\ = & \sigma^2(a - 1)(b - 1) + \beta' X'(M - M_\mu - M_\alpha - M_\eta)X\beta \\ = & \sigma^2(a - 1)(b - 1) + N \sum_{i=1}^a \sum_{j=1}^b (\gamma_{ij} - \bar{\gamma}_{i\cdot} - \bar{\gamma}_{\cdot j} + \bar{\gamma}_{\dots})^2. \end{aligned}$$

## Balanced Two-way ANOVA with interaction

The expected values of  $Y' M_\alpha Y$  and  $Y' M_\eta Y$  are different from those found in the main effects (no interaction) model. It is easily shown that

$$E(Y' M_\alpha Y) = \sigma^2(a - 1) + bN \sum_{i=1}^a (\alpha_i + \bar{\gamma}_{i\cdot} - \bar{\alpha}_{\cdot\cdot} - \bar{\gamma}_{\cdot\cdot})^2$$

which now depends on the  $\gamma_{ij}$ 's. Recall that for the no interaction model that

$$E(Y' M_\alpha Y) = \sigma^2(a - 1) + bN \sum_{i=1}^a (\alpha_i - \bar{\alpha}_{\cdot\cdot})^2.$$

This difference in expectations implies that the  $F$  test for testing no  $\alpha$  treatment effect is NOT a test that the  $\alpha_i$ 's are all equal. Rather it is a test of the hypothesis

$$H_0 : \alpha_1 + \bar{\gamma}_{1\cdot} = \dots = \alpha_a + \bar{\gamma}_{a\cdot} .$$

## Balanced Two-way ANOVA with interaction

Thus, in the two-way ANOVA model with interaction, the hypothesis of no  $\alpha$  treatment effect is a test of the hypothesis that all the  $\alpha_i + \bar{\gamma}_i$ 's are equal. The  $F$  test for this hypothesis is given by

$$F = \frac{\|M_\alpha Y\|^2 / r(M_\alpha)}{MSE} ,$$

where as usual  $MSE = \|(I - M)Y\|^2 / r(I - M)$ . Similarly, the  $F$  test for no  $\eta$  treatment effect is a test of the hypothesis

$$H_0 : \eta_1 + \bar{\gamma}_{.1} = \dots = \eta_b + \bar{\gamma}_{.b} ,$$

and the  $F$  test is given by

$$F = \frac{\|M_\eta Y\|^2 / r(M_\eta)}{MSE} .$$

# Balanced Two-way ANOVA with interaction

Since

$$C(X) = C(X_{a+b+1}, \dots, X_{a+b+ab}),$$

it follows that ALL estimable functions of the parameters are functions of the  $\gamma_{ij}$ 's. Thus, ANY estimable function MUST involve the  $\gamma_{ij}$ 's in some way. Thus

- i)  $\sum_{i=1}^a \lambda_i \alpha_i$ ,  $\sum_{i=1}^a \lambda_i = 0$ , is NOT estimable.
- ii)  $\sum_{j=1}^b \lambda_j \eta_j$ ,  $\sum_{j=1}^b \lambda_j = 0$ , is NOT estimable.

It turns out that the class of all estimable function for the two-way ANOVA model with interaction is

- i)  $E(Y_{ijk}) = \mu + \alpha_i + \eta_j + \gamma_{ij}.$
- ii)  $\alpha_i + \bar{\gamma}_i - \bar{\alpha}_. - \bar{\gamma}_..$
- iii)  $\eta_j + \bar{\gamma}_{.j} - \bar{\eta}_. - \bar{\gamma}_{...}$
- iv) All contrasts in the  $\alpha$  space.
- v) All contrasts in the  $\eta$  space.
- vi) All contrasts in the interaction space.

## Balanced Two-way ANOVA with interaction

To see that any estimable function must involve the  $\gamma_{ij}$ 's, note that if  $\lambda'\beta$  is not a function of the  $\gamma_{ij}$ 's, but is estimable, then

$$\rho' X_i = 0$$

for  $i = a + b + 1, \dots, a + b + ab$ . This implies that  $\rho' X = 0$ , since

$$C(X) = C(X_{a+b+1}, \dots, X_{a+b+ab}) .$$

This would thus imply that the estimable function is identically equal to 0.

# Balanced Two-way ANOVA with interaction

## Contrasts for Two-way ANOVA with interaction

We first consider constructing contrasts for the  $\alpha$  space. We can put a constraint on  $M_\alpha$  by choosing a function  $\lambda' \beta$  such that  $\lambda' = \rho' X$  and  $M_\alpha \rho = M \rho$ . Such a constraint NO longer defines a contrast involving only the  $\alpha_i$ 's. To examine contrasts in the  $\alpha$  space, we need to examine the nature of

$$\lambda' \beta = \rho' X \beta = \rho' M X \beta = \rho' M_\alpha X \beta.$$

$M_\alpha X \beta$  is an  $n \times 1$  vector whose elements are of the form

$$\alpha_i + \bar{\gamma}_{i\cdot} - \bar{\alpha}_{\cdot} - \bar{\gamma}_{\dots} .$$

Now  $\rho' M_\alpha X \beta$  will be a contrast in these terms, or equivalently a contrast in

$$\alpha_i + \bar{\gamma}_{i\cdot} .$$

A contrast in terms of  $\alpha_i + \bar{\gamma}_{i\cdot}$  will be called a contrast in the  $\alpha$  space.

Similarly, a contrast in terms of

$$\eta_j + \bar{\gamma}_{\cdot j}$$

will be called a contrast in the  $\eta$  space.

## Examples

- i)  $\alpha_1 + \bar{\gamma}_{1\cdot} - (\alpha_2 + \bar{\gamma}_{2\cdot})$  is an contrast in the  $\alpha$  space.
- ii)  $\eta_1 + \bar{\gamma}_{\cdot 1} - (\eta_2 + \bar{\gamma}_{\cdot 2})$  is a contrast in the  $\eta$  space.

# Balanced Two-way ANOVA with interaction

## Contrasts in the interaction space

To find the set of contrasts in the interaction space, we need to know how to put a constraint on  $M_\gamma = M - M_\mu - M_\alpha - M_\eta$ . The hypothesis

$$H_0 : \rho' X \beta = 0$$

puts a constraint on the interaction space if and only if

$$M\rho = M_\gamma\rho.$$

Thus,  $\rho' X \beta = 0$ , implies

$$\rho \in C(M_\mu + M_\alpha + M_\eta)^\perp ,$$

which implies that

$$\rho' X_i = 0$$

for  $i = 0, \dots, a+b$ , where  $X_0 = J$ .

There are two properties which characterize contrasts in the interaction space.  
These are

- i)  $M\rho = \rho$ .
- ii)  $\rho' X_i = 0, \quad i = 0, \dots, a+b$ .

Property (ii) ensures that  $M\rho$  is in the interaction space, and property (i) implies that  $\rho$  is in the interaction space.

# Balanced Two-way ANOVA with interaction

To construct a contrast in the interaction space, we combine contrasts in the  $\alpha$  space and the  $\eta$  space. To do this, we need to establish the notion of a Kronecker product.

## Defn

Suppose  $A$  is an  $r \times c$  matrix and  $B$  is an  $s \times d$  matrix. The Kronecker product of  $A$  and  $B$ , written,  $A \otimes B$  is an  $rs \times cd$  matrix of the form

$$A \otimes B = (a_{ij}B)$$

where  $a_{ij}$  is the  $ij$ th element of  $A$ .

## Example

Suppose  $A = \begin{pmatrix} 1 & 4 \\ 2 & 5 \end{pmatrix}$  and  $B = \begin{pmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{pmatrix}$ . Then

$$A \otimes B = \begin{pmatrix} 1 & 2 & 4 & 8 \\ 3 & 4 & 12 & 16 \\ 5 & 6 & 20 & 24 \\ 2 & 4 & 5 & 10 \\ 6 & 8 & 15 & 20 \\ 10 & 12 & 25 & 30 \end{pmatrix}.$$

We are now led to the following theorem.

# Balanced Two-way ANOVA with interaction

## Theorem

Let  $c' = (c_1, \dots, c_a)_{1 \times a}$  denote a set of contrast coefficients for the  $\alpha$  space and let  $d' = (d_1, \dots, d_b)_{1 \times b}$  denote a set of contrast coefficients for the  $\eta$  space. We have  $\sum_{i=1}^a c_i = 0$  and  $\sum_{i=1}^b d_i = 0$ . Then a set of contrast coefficients for the interaction space are of the form

$$c' \otimes d'.$$

Thus, a corresponding  $\rho$  vector for a contrast in the interaction space is of the form

$$\begin{aligned}\rho' &= \frac{1}{N}(c' \otimes d') \otimes J'_N \\ &= \frac{1}{N}(c_1 d_1 J'_N, c_1 d_2 J'_N, \dots, c_1 d_b J'_N, c_2 d_1 J'_N, \dots, c_a d_b J'_N)\end{aligned}$$

where  $J_N$  is the  $N \times 1$  vector of ones. Thus, to obtain a contrast in the interaction space, we take the Kronecker product of contrasts in the  $\alpha$  space and  $\eta$  space. However, this theorem does NOT characterize the set of all possible contrasts in the interaction space. It just tells us how to construct a contrast in the interaction space.

## Balanced Two-way ANOVA with interaction

### Example

Consider example 7.2.2 on page 154 of Christensen. Here  $c' = (1, 2, -3)$  and  $d' = (1, -1)$ . Thus

$$\begin{aligned}c' \otimes d' &= (c_1d_1, c_1d_2, c_2d_1, c_2d_2, c_3d_1, c_3d_2) \\&= (1, -1, 2, -2, -3, 3)\end{aligned}$$

Since  $N = 4$ , we have

$$\begin{aligned}\rho' &= \frac{1}{4}(c' \otimes d') \otimes J'_4 \\&= \frac{1}{4}(1J'_4, -1J'_4, 2J'_4, -2J'_4, -3J'_4, 3J'_4)\end{aligned}$$

We now give a theorem that characterizes all possible contrasts in the interaction space.

# Balanced Two-way ANOVA with interaction

## Theorem

Suppose  $Q$  is an  $a \times b$  matrix such that  $J_a' Q = 0$  and  $Q J_b = 0$ . (i.e.,  $\bar{q}_{i\cdot} = 0$  and  $\bar{q}_{\cdot j} = 0$ ). Thus  $Q$  is a matrix with each row and each column adding to zero. Then ALL contrasts in the interaction space have  $\rho$  of the form

$$\rho' = (\rho_{ijk}) \quad (2)$$

where  $\rho_{ijk} = q_{ij}/N$  where  $\bar{q}_{i\cdot} = 0$  and  $\bar{q}_{\cdot j} = 0$ . That is

$$\rho' = \left( q_{11} \otimes \frac{J_N'}{N}, q_{12} \otimes \frac{J_N'}{N}, \dots, q_{ab} \otimes \frac{J_N'}{N} \right).$$

This class of  $\rho$ 's satisfy

- i)  $\rho \in C(X)$  and
- ii)  $\rho' X_i = 0$  for  $i = 0, \dots, a+b$ .

Thus, the interaction space is precisely the set of all vectors with the form of (2).

# Three- or higher-way balanced ANOVA

## Three or higher way balanced ANOVA

We apply the same ideas as before to decompose  $C(X)$ . For example, the three-way ANOVA model with all possible interactions takes the form

$$\begin{aligned} Y_{ijkl} &= \mu + \alpha_i + \eta_j + \gamma_k + (\alpha\eta)_{ij} + (\alpha\gamma)_{ik} \\ &\quad + (\eta\gamma)_{jk} + (\alpha\eta\gamma)_{ijk} + \epsilon_{ijkl} \end{aligned}$$

where  $i = 1, \dots, a$ ,  $j = 1, \dots, b$ ,  $k = 1, \dots, c$ , and  $l = 1, \dots, N$ .

In this model we have three main effects, three two-way interactions, and one three-way interaction. We can decompose  $C(X)$  into a sum of eight orthogonal subspaces. That is

$$\begin{aligned} C(M) &= C(M_\mu) + C(M_\alpha) + C(M_\eta) + C(M_\gamma) \\ &\quad + C(M_{\alpha\eta}) + C(M_{\alpha\gamma}) + C(M_{\eta\gamma}) + C(M_{\alpha\eta\gamma}) \end{aligned}$$

The same procedures as before apply in obtaining the orthogonal projection operators onto the various spaces.

# A Unified Approach to Balanced ANOVA Models

## A Unified Approach to Balanced ANOVA Models

We can develop a unified approach to obtaining orthogonal projection operators in arbitrary balanced  $k$ -way ANOVA models by exploiting the structure of the design matrix. The structure of the design matrix can be easily examined using Kronecker products. Therefore, before we proceed further, we need to establish some more properties of Kronecker products.

Recall the definition of a Kronecker product.

$$A \otimes B = (a_{ij}B).$$

### Defn

Suppose  $A$  is an  $n \times p$  matrix, written as  $A = (a_1, \dots, a_p)$ , where  $a_i$  is the  $i$ th column of  $A$ . Then

$$\text{Vec}(A) = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{pmatrix}.$$

Thus  $\text{Vec}(A)$  is an  $np \times 1$  vector. The  $\text{Vec}$  operator stacks the columns of  $A$  into one long vector.

## Properties of Kronecker Products

- 1)  $A \otimes 0 = 0 \otimes A = 0.$
- 2)  $(A_1 + A_2) \otimes B = A_1 \otimes B + A_2 \otimes B.$
- 3)  $B \otimes (A_1 + A_2) = B \otimes A_1 + B \otimes A_2.$
- 4)  $(\alpha A) \otimes (\beta B) = (\alpha\beta)(A \otimes B),$  where  $\alpha$  and  $\beta$  are scalars.
- 5)  $(A_1 A_2) \otimes (B_1 B_2) = (A_1 \otimes B_1)(A_2 \otimes B_2).$
- 6)  $(A \otimes B)^- = A^- \otimes B^-.$
- 7)  $A \otimes (B \otimes C) = (A \otimes B) \otimes C.$
- 8)  $(A \otimes B)' = A' \otimes B'.$
- 9)  $|A \otimes B| = |A|^m |B|^n,$  where  $A_{n \times n}$  and  $B_{m \times m}.$
- 10)  $tr(A \otimes B) = tr(A)tr(B).$

## Properties of Kronecker Products (cont'd)

- 11) The eigenvalues of  $A \otimes B$  are  $\lambda_{ia}^m \lambda_{jb}^n$ ,  $i = 1, \dots, n$ ,  $j = 1, \dots, m$ , where  $\lambda_{ia}$ ,  $i = 1, \dots, n$  are the eigenvalues of  $A$ , and  $\lambda_{jb}$ ,  $j = 1, \dots, m$  are the eigenvalues of  $B$ , and  $A_{n \times n}$  and  $B_{m \times m}$  are symmetric. The eigenvector of  $\lambda_{ia}^m \lambda_{jb}^n$  is  $v_{ia} \otimes v_{jb}$ , where  $v_{ia}$  is the eigenvector of  $\lambda_{ia}$  and  $v_{jb}$  is the eigenvector of  $\lambda_{jb}$ .
- 12)  $\text{Vec}(ABC) = (A \otimes C')\text{Vec}(B)$ .

# A Unified Approach to Balanced ANOVA Models

Consider the balanced two-way ANOVA model with interaction. This model is given by

$$Y_{ijk} = \mu + \alpha_i + \eta_j + \gamma_{ij} + \epsilon_{ijk},$$

where  $i = 1, \dots, a$ ,  $j = 1, \dots, b$ ,  $k = 1, \dots, N$ , and  $n = abN$ .

We want to write

$$C(M) = C(M_\mu) + C(M_\alpha) + C(M_\eta) + C(M_\gamma)$$

and be able to compute the orthogonal projection operators in an easy and unified way.

We can represent each subspace making up  $C(M)$  in terms of Kronecker products. Once we do this, we can easily obtain the orthogonal projection operator for that space.

Notation: Let  $s$  be an arbitrary index. Define  $J_s$  as the  $s \times 1$  vector of ones,  $P_s = \frac{1}{s} J_s J_s'$  and  $Q_s = I_s - P_s$ , where  $I_s$  is the  $s \times s$  identity matrix. Thus,  $P_s$  is the orthogonal projection operator onto  $C(J_s)$  and  $Q_s$  is the orthogonal projection operator onto  $C(J_s)^\perp$ .

# A Unified Approach to Balanced ANOVA Models

## Facts

- 1) Recall that the orthogonal projector operator onto  $C(A)$  is always given by  $A(A'A)^{-}A'$ .
- 2) If  $M$  is an orthogonal projection operator, then  $M^{-} = M$ .

## Kronecker Product forms for the Orthogonal

### Projection Operators

- 1) Computing  $M_{\mu}$ .

We can write  $J_n = J_a \otimes J_b \otimes J_N$ , so that  $M_{\mu}$  is the orthogonal projection operator onto  $C(J_a \otimes J_b \otimes J_N)$ . Thus by Fact 1 above, we have

$$\begin{aligned}M_{\mu} &= (J_a \otimes J_b \otimes J_N) \left( (J'_a \otimes J'_b \otimes J'_N)(J_a \otimes J_b \otimes J_N) \right)^{-} (J'_a \otimes J'_b \otimes J'_N) \\&= (J_a \otimes J_b \otimes J_N)(J'_a J_a \otimes J'_b J_b \otimes J'_N J_N)^{-} (J'_a \otimes J'_b \otimes J'_N) \\&= (J_a \otimes J_b \otimes J_N)(abN)^{-} (J'_a \otimes J'_b \otimes J'_N) \\&= \frac{1}{a} J_a J'_a \otimes \frac{1}{b} J_b J'_b \otimes \frac{1}{N} J_N J'_N \\&= P_a \otimes P_b \otimes P_N.\end{aligned}$$

# A Unified Approach to Balanced ANOVA Models

- 2) Computing  $M_\alpha$ . The  $\alpha$  space is  $C(Q_a \otimes J_b \otimes J_N)$ . Thus

$$\begin{aligned} M_\alpha &= (Q_a \otimes J_b \otimes J_N) \left( (Q'_a \otimes J'_b \otimes J'_N)(Q_a \otimes J_b \otimes J_N) \right)^{-1} (Q'_a \otimes J'_b \otimes J'_N) \\ &= (Q_a \otimes J_b \otimes J_N)(Q'_a Q_a \otimes J'_b J_b \otimes J'_N J_N)^{-1} (Q_a \otimes J_b \otimes J_N) \\ &= (Q_a \otimes J_b \otimes J_N)(Q_a^- \otimes b^- \otimes N^-)(Q'_a \otimes J'_b \otimes J'_N) \\ &= Q_a \otimes P_b \otimes P_N. \end{aligned}$$

- 3) Computing  $M_\eta$ . The  $\eta$  space is  $C(J_a \otimes Q_b \otimes J_N)$ . Similar derivations obtain

$$M_\eta = P_a \otimes Q_b \otimes P_N.$$

- 4) Computing  $M_\gamma$ . The interaction space is given by  $C(Q_a \otimes Q_b \otimes J_N)$ . This yields

$$M_\gamma = Q_a \otimes Q_b \otimes P_N .$$

# A Unified Approach to Balanced ANOVA Models

Now  $M = M_\mu + M_\alpha + M_\eta + M_\gamma$ . In Kronecker product notation, these orthogonal projection operators are very easy to sum. For example, note that

$$\begin{aligned}M_\alpha + M_\eta &= (Q_a \otimes P_b \otimes P_N) + (P_a \otimes Q_b \otimes P_N) \\&= (Q_a \otimes P_b + P_a \otimes Q_b) \otimes P_N.\end{aligned}$$

Using the properties of Kronecker products, it can easily be shown that

$$M = I_a \otimes I_b \otimes P_N.$$

The error space is  $C(I - M)$  and

$$\begin{aligned}I - M &= I_{abN} - M \\&= (I_a \otimes I_b \otimes I_N) - (I_a \otimes I_b \otimes P_N) \\&= (I_a \otimes I_b) \otimes (I_N - P_N) \\&= I_a \otimes I_b \otimes Q_N.\end{aligned}$$

Observe that

$$\begin{aligned}M + I - M &= (I_a \otimes I_b \otimes P_N) + I_a \otimes I_b \otimes Q_N \\&= (I_a \otimes I_b) \otimes (P_N + Q_N) \\&= (I_a \otimes I_b) \otimes I_N \\&= I_a \otimes I_b \otimes I_N \\&= I_n.\end{aligned}$$

# A Unified Approach to Balanced ANOVA Models

We can summarize the subspaces and the orthogonal projection operators for the two-way ANOVA model as follows.

Effect	Subspace	Orthogonal Projection Operator
$\mu$	$C(J_a \otimes J_b \otimes J_N)$	$P_a \otimes P_b \otimes P_N$
$\alpha$	$C(Q_a \otimes J_b \otimes J_N)$	$Q_a \otimes P_b \otimes P_N$
$\eta$	$C(J_a \otimes Q_b \otimes J_N)$	$P_a \otimes Q_b \otimes P_N$
$\gamma$	$C(Q_a \otimes Q_b \otimes J_N)$	$Q_a \otimes Q_b \otimes P_N$
Error	$C(I - M)$	$I_a \otimes I_b \otimes Q_N$
Total		$I_a \otimes I_b \otimes I_N$

## Exercise

Consider the three-way ANOVA model

$$\begin{aligned} Y_{ijkl} &= \mu + \alpha_i + \eta_j + \gamma_k + (\alpha\eta)_{ij} + (\alpha\gamma)_{ik} + (\eta\gamma)_{jk} \\ &\quad + (\alpha\eta\gamma)_{ijk} + \epsilon_{ijkl} \end{aligned}$$

where  $i = 1, \dots, a$ ,  $j = 1, \dots, b$ ,  $k = 1, \dots, c$ , and  $l = 1, \dots, N$ .

- Write out the subspaces and all orthogonal projection operators corresponding to each term in the ANOVA model completely in terms of Kronecker products.
- Find the simplest expression for  $M_\mu + M_\alpha + M_\eta$ .

# Unbalanced Two-way ANOVA

## Unbalanced Two-way ANOVA

When we have an unequal number of replicates per cell, the estimation space  $C(X)$  CANNOT be decomposed into a sum of orthogonal subspaces, in general. There are special cases under which  $C(X)$  can be decomposed into a sum of orthogonal subspaces when we have an unequal number of replicates. We first discuss these special cases.

Consider the model

$$Y_{ijk} = \mu + \alpha_i + \eta_j + \epsilon_{ijk} \quad (3)$$

where  $i = 1, \dots, a$ ,  $j = 1, \dots, b$ ,  $k = 1, \dots, n_{ij}$ .

### Defn

We say that a model has proportional numbers if for  $i, i' = 1, \dots, a$ , and  $j, j' = 1, \dots, b$ ,

$$\frac{n_{ij}}{n_{ij'}} = \frac{n_{i'j}}{n_{i'j'}} .$$

It turns out that orthogonality of subspaces is preserved if we have proportional numbers. That is,

$$C(X) = M_\mu + M_\alpha + M_\eta ,$$

where  $M_\mu$ ,  $M_\alpha$  and  $M_\eta$  are all mutually orthogonal. The following theorem gives a an equivalent condition for proportional numbers.

# Unbalanced Two-way ANOVA

## Theorem

Suppose the  $n_{rs}$ 's satisfy the proportional numbers definition. Then for any  $r = 1, \dots, a$ , and  $s = 1, \dots, b$ ,

$$n_{rs} = \frac{n_{r\cdot} n_{\cdot s}}{n_{..}}$$

where

$$n_{r\cdot} = \sum_{j=1}^b n_{rj}$$

$$n_{\cdot s} = \sum_{i=1}^a n_{is}$$

and

$$n_{..} = \sum_{i=1}^a \sum_{j=1}^b n_{ij} .$$

This theorem says that proportional numbers is equivalent to writing any  $n_{rs}$  as a product of the  $r$ th row sample size total times the  $s$ th column sample size total divided by total sample size.

# Unbalanced Two-way ANOVA

## Proof

Using the definition of proportional numbers, we have

$$\frac{n_{rs}}{n_{rj}} = \frac{n_{is}}{n_{ij}} ,$$

which implies

$$n_{ij} n_{rs} = n_{rj} n_{is} .$$

Now summing both sides of the equation over  $i$  and  $j$ , we have

$$\sum_{i=1}^a \sum_{j=1}^b n_{ij} n_{rs} = \sum_{i=1}^a \sum_{j=1}^b n_{rj} n_{is} \quad (4)$$

The left side of (4) reduces to

$$n_{rs} \sum_{i=1}^a \sum_{j=1}^b n_{ij} = n_{rs} n_{..}$$

The right side of (4) reduces to

$$\sum_{i=1}^a \sum_{j=1}^b n_{rj} n_{is} = \left( \sum_{j=1}^b n_{rj} \right) \left( \sum_{i=1}^a n_{is} \right) = n_r n_s .$$

# Unbalanced Two-way ANOVA

Thus we have

$$n_{rs} n_{..} = n_r n_s$$

which implies

$$n_{rs} = \frac{n_r n_s}{n_{..}}$$

We can write the design matrix of model (3) as

$$X = (J, X_1, \dots, X_a, X_{a+1}, \dots, X_{a+b})$$

where

$$X_r = (t_{ijk})$$

where  $r = 1, \dots, a$ ,  $t_{ijk} = \delta_{ir}$ , and  $\delta_{ir} = 1$  if  $i = r$ , 0 otherwise. Also,

$$X_{a+s} = (u_{ijk})$$

$s = 1, \dots, b$ ,  $u_{ijk} = \delta_{js}$ , and  $\delta_{js} = 1$  if  $j = s$  and 0 otherwise.

# Unbalanced Two-way ANOVA

With proportional numbers, we can construct orthogonal subspaces as follows.  
As before, define

$$Z_r = X_r - \frac{n_{r..}}{n_{..}} J \quad r = 1, \dots, a$$

$$Z_{a+s} = X_s - \frac{n_{.s}}{n_{..}} J \quad s = 1, \dots, b .$$

We can see that for  $r = 1, \dots, a$ , and  $s = 1, \dots, b$ ,

$$\begin{aligned} Z'_{a+s} Z_r &= (X_r - \frac{n_{r..}}{n_{..}} J)' (X_s - \frac{n_{.s}}{n_{..}} J) \\ &= X'_r X_s - \frac{n_{r..}}{n_{..}} (J' X_s) - \frac{n_{.s}}{n_{..}} (X'_r J) \\ &\quad + \frac{n_{r..} n_{.s}}{n_{..}^2} (J' J) \\ &= n_{rs} - n_{.s} \frac{n_{r..}}{n_{..}} - n_{r..} \frac{n_{.s}}{n_{..}} + n_{..} \frac{n_{r..} n_{.s}}{n_{..}^2} \\ &= n_{rs} - \frac{n_{.s} n_{r..}}{n_{..}} - \frac{n_{r..} n_{.s}}{n_{..}} + \frac{n_{r..} n_{.s}}{n_{..}} \\ &= n_{rs} - \frac{n_{r..} n_{.s}}{n_{..}} \\ &= \frac{n_{r..} n_{.s}}{n_{..}} - \frac{n_{r..} n_{.s}}{n_{..}} = 0 . \end{aligned}$$

# Unbalanced Two-way ANOVA

Similar results can be obtained if an interaction term is included in the model with proportional numbers.

## The General Case

If we don't have proportional numbers, then  $C(X)$  cannot be decomposed into a sum of orthogonal subspaces with each subspace corresponding to a term in the model. In this case, the sum of squares depend on which terms have been included the model. That is, the sums of squares depend on the order of the inclusion of the effects. Let  $R(\alpha | \mu, \eta)$  denote the sums of squares of the  $\alpha$  treatment given that  $\mu$  and  $\eta$  are in the model, and so forth. In this general case,

$$R(\alpha | \mu, \eta) \neq R(\alpha | \mu)$$

and

$$R(\eta | \mu, \alpha) \neq R(\eta | \mu)$$

In the general case, we analyze the unbalanced ANOVA model using the general theory of linear models. That is, we write the model as

$$Y = X\beta + \epsilon$$

and compute sums of squares of various models by finding the appropriate  $X_0$  matrix, where  $C(X_0) \subset C(X)$ . Thus, we do tests and inference using the general nested versus full framework that we developed earlier.

# Unbalanced Two-way ANOVA

For example, consider the two-way unbalanced ANOVA model with interaction, given by

$$Y_{ijk} = \mu + \alpha_i + \eta_j + \gamma_{ij} + \epsilon_{ijk}$$

where  $i = 1, \dots, a$ ,  $j = 1, \dots, b$ , and  $k = 1, \dots, n_{ij}$ . Also assume that the  $\epsilon_{ijk}$ 's are *i.i.d.*  $N(0, \sigma^2)$  random variables. Suppose we wanted to test the hypothesis of no interaction. That is, we want to test

$$H_0 : \gamma_{ij} = \text{constant for all } (i, j)$$

$$H_a : \gamma_{ij} \neq \text{constant for at least one pair } (i, j)$$

To conduct this test in the unbalanced case, we rewrite the hypotheses as

$$H_0 : E(Y) \in C(X_0)$$

$$H_a : E(Y) \in C(X) \cap C(X_0)^c$$

where  $C(X_0) \subset C(X)$ , and  $X_0$  corresponds to the matrix of the reduced model of no interaction. Thus the  $F$  test is

$$F = \frac{\|(M - M_0)Y\|^2 / r(M - M_0)}{\|(I - M)Y\|^2 / r(I - M)}$$

as before.

# Unbalanced Two-way ANOVA

As a specific example, consider  $a = 2$ ,  $b = 3$ ,  $n_{11} = 2$ ,  $n_{12} = 1$ ,  $n_{13} = 3$ ,  $n_{21} = 1$ ,  $n_{22} = 2$ , and  $n_{23} = 2$ . It is clear that we don't have proportional numbers here. The total sample size is  $n = \sum_{i=1}^a \sum_{j=1}^b n_{ij} = 11$ . We can write the two-way ANOVA model with interaction as

$$Y = X\beta + \epsilon$$

where

$$Y = \begin{pmatrix} Y_{111} \\ Y_{112} \\ Y_{121} \\ Y_{131} \\ Y_{132} \\ Y_{133} \\ Y_{211} \\ Y_{221} \\ Y_{222} \\ Y_{231} \\ Y_{232} \end{pmatrix}_{11 \times 1}, \beta = \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \eta_1 \\ \eta_2 \\ \eta_3 \\ \gamma_{11} \\ \gamma_{12} \\ \gamma_{13} \\ \gamma_{21} \\ \gamma_{22} \\ \gamma_{23} \end{pmatrix}_{12 \times 1}$$

# Unbalanced Two-way ANOVA

$$X = \begin{pmatrix} 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}_{11 \times 12}$$

# Unbalanced Two-way ANOVA

The  $X_0$  matrix corresponding to the no interaction model is

$$X_0 = \begin{pmatrix} 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 \end{pmatrix}_{11 \times 6}.$$

# Experimental Design Models

## Experimental Design Models

We want to discuss some basic ANOVA models that are heavily used in experimental design. These are

- i) Completely randomized design (CRD)
- ii) Randomized Complete Block design (RCB)
- iii) Latin Square Design
- iv) Factorial designs

### Completely Randomized Design (CRD)

This is perhaps the simplest possible experimental design. In this design, we have homogeneous experimental units. If we have  $t$  treatments, we randomly divide the experimental units into  $t$  groups, and a treatment is applied to each experimental unit in a group. The standard model for a CRD is the one-way ANOVA model given by

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

$i = 1, \dots, t$ , and  $j = 1, \dots, n_i$ . Here the  $\alpha_i$ 's represent the treatment effects. We assume the  $\epsilon_{ij}$ 's are i.i.d.  $N(0, \sigma^2)$  random variables. Our main goal in CRD design is to compare the  $t$  treatments.

## Blocking

Blocking is one of the most important concepts in experimental design. We block to reduce variability so that differences between treatments can be better assessed. Blocking consists of grouping homogeneous experimental units into blocks, and then randomly applying the treatments to the units in each block. Blocking factors include days, physical characteristics such as height or weight, gender, and so forth. Blocking is a central issue in randomized complete block designs.

# Experimental Design Models

## Randomized Complete Block Design (RCB)

The standard model for an RCB is the two-way ANOVA model without interaction model given by

$$Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}$$

where  $i = 1, \dots, a$ ,  $j = 1, \dots, b$ , and the  $\epsilon_{ij}$ 's are *i.i.d.*  $N(0, \sigma^2)$  random variables. The  $\alpha_i$ 's represent the treatment effects and the  $\beta_j$ 's represent the block effects. Our main goal for this design is to compare the  $a$  treatments, after adjusting for the blocks. We typically do not have an interest in making inferences about the blocks, nor do we usually include an interaction term for this type of model.

The randomization for an RCB proceeds as follows. Experimental units are first arranged into blocks, then the treatments are randomly assigned to the experimental units within a block. The idea of blocking is that we remove extraneous sources of variability so that treatment differences can be better assessed. The RCB design is known as a variance reduction design.

# Experimental Design Models

## Latin Square Designs

The Latin square design allows for two different blocking factors. Consider the following example.

### Example

- i) Suppose we want to compare four treatments  $A$ ,  $B$ ,  $C$ , and  $D$ .
- ii) We have four different hospitals (blocking factor 1).
- iii) We have four different machines for treatment administration at each hospital.

Machines

Hospitals	M1	M2	M3	M4
1	C	D	B	A
2	B	C	A	D
3	A	B	D	C
4	D	A	C	B

We note that each treatment occurs only once in each row and each column. The rows and columns represent the blocking factors. In this example, the Latin square allows us to remove the variation due to hospitals, machine types, and treatments. The Latin square is also a variance reduction design.

# Experimental Design Models

Latin square designs arise often in agricultural experiments. Often, a fertility gradient exists on plots of land where crops are planted. Typically, there is an East-West, North-South fertility gradient.

The model for the Latin square design is

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + \epsilon_{ijk}$$

where  $i = 1, \dots, a$ ,  $j = 1, \dots, a$ , and  $k = f(i, j)$ , where for each  $i$ ,  $f(i, j)$  is a one-to-one function of  $\{1, \dots, a\}$ . Here  $\alpha_i$  represents the  $i$ th row effect,  $\beta_j$  represents the  $j$ th column effect, and  $\gamma_k$  represents the  $k$ th treatment effect.

This model can be written in the form

$$Y = X\beta + \epsilon$$

Where  $X$ ,  $Y$ , and  $\beta$  are defined in the usual way. For the special case  $a = 4$ , Christensen, p. 187 gives the design matrix  $X$ , and  $\beta$ .

# Experimental Design Models

From the Latin square design, we can decompose  $C(X)$  into a sum of four orthogonal subspaces. Specifically, we can write

$$C(M) = C(M_\mu) + C(M_\alpha) + C(M_\beta) + C(M_\gamma).$$

To construct these column spaces, we use the same old trick of defining the  $Z$ 's. Write

$$X = (J, X_1, \dots, X_a, X_{a+1}, \dots, X_{2a}, X_{2a+1}, \dots, X_{3a}).$$

Using the  $\delta$  notation, we write

$$X_r = (u_{ijk}) \ , r = 1, \dots, a$$

where  $u_{ijk} = \delta_{ir}$ ,  $\delta_{ir} = 1$  if  $i = r$  and 0 otherwise. Also, write

$$X_{a+s} = (u_{ijk}) \ , s = 1, \dots, a$$

where  $u_{ijk} = \delta_{js}$ ,  $\delta_{js} = 1$  if  $j = s$  and 0 otherwise. Finally, write

$$X_{2a+t} = (u_{ijk}) \ , t = 1, \dots, a$$

where  $u_{ijk} = \delta_{kt}$ ,  $\delta_{kt} = 1$  if  $k = t$  and 0 otherwise.

# Experimental Design Models

Now define  $Z_0 = J$ ,

$$\begin{aligned} Z_i &= X_i - \left( \frac{X_i' J}{J' J} \right) J \\ &= X_i - \frac{a}{a^2} J \\ &= X_i - \frac{1}{a} J \end{aligned}$$

for  $i = 1, \dots, 3a$ . Using the same arguments as before, it can be shown that the four column spaces  $C(Z_0)$ ,  $C(Z_1, \dots, Z_a)$ ,  $C(Z_{a+1}, \dots, Z_{2a})$ , and  $C(Z_{2a+1}, \dots, Z_{3a})$  are all orthogonal. Let

$$Z_\alpha = (Z_1, \dots, Z_a)$$

$$Z_\beta = (Z_{a+1}, \dots, Z_{2a})$$

and

$$Z_\gamma = (Z_{2a+1}, \dots, Z_{3a}) .$$

We have  $r(Z_\alpha) = r(Z_\beta) = r(Z_\gamma) = a - 1$ , and

$$M_\alpha = Z_\alpha (Z_\alpha' Z_\alpha)^{-1} Z_\alpha' ,$$

$$M_\beta = Z_\beta (Z_\beta' Z_\beta)^{-1} Z_\beta'$$

and

$$M_\gamma = Z_\gamma (Z_\gamma' Z_\gamma)^{-1} Z_\gamma' .$$

# Experimental Design Models

As usual,  $M_\mu = \frac{1}{n}JJ'$ , where  $n = a^2$ . The ANOVA table for an  $a \times a$  Latin square is given on page 188 of Christensen. The sums of squares for each effect can be written as  $Y'M_\mu Y$ ,  $Y'M_\alpha Y$ ,  $Y'M_\beta Y$ , and  $Y'M_\gamma Y$ .

# Experimental Design Models

## Factorial Treatment Structures

Factorial designs arise when we have two or more factors (treatments) in an experiment, and we wish to construct all possible treatment combinations. For example, suppose we have two factors  $A$  and  $B$ , where  $A$  has  $a$  levels and  $B$  has  $b$  levels. Then there are  $ab$  different treatment combinations. Factorial designs arise often in agricultural and chemical experiments.

## Example

Suppose an experiment is to be conducted examining the effects of fertilizer on potato yields. Of interest are two kinds of fertilizer, a nitrogen based fertilizer (factor  $A$ , at 2 levels) and a phosphate based fertilizer (factor  $B$ , at 3 levels). Thus there are six possible treatment combinations of the two fertilizers. See Example 8.4.1, p.189 of Christensen for more details.

# Analysis of Covariance

## Analysis of Covariance

Analysis of covariance models are generalizations of ANOVA models in that the design matrix contains both qualitative and quantitative explanatory variables. The quantitative variable are referred to as regression variables, covariates, or concomitant variables. There are essentially two goals in analysis of covariance (ANCOVA).

- i) Compare treatments
- ii) Inference on the regression coefficients corresponding to the covariates.

The primary goal is still to compare treatments. The concomitant variables are intended to serve as blocking factors to sharpen the analysis so that differences between treatments can be better assessed. Thus, concomitant variables are introduced to reduce variability. Thus, ANCOVA can be viewed as a variance reduction design, for which the concomitant variable is quantitative and often continuous. The blocking variable in RCB and Latin square is a qualitative or discrete variable.

Two very important applications of ANCOVA are

- i) missing data
- ii) balanced incomplete block designs (BIBD).

# Analysis of Covariance

An example of the scalar form of the ANCOVA model was given in problem 1.1, page 3. This model was written as

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma x_{ijk} + \epsilon_{ijk}.$$

Here, we have two qualitative factors and one concomitant variable  $x_{ijk}$ .  $\gamma$  is the regression coefficient of the concomitant variable.

The general ANCOVA model can be written in matrix notation as

$$\begin{aligned} Y &= (X, Z) \begin{pmatrix} \beta \\ \gamma \end{pmatrix} + \epsilon \\ &= X\beta + Z\gamma + \epsilon \end{aligned} \tag{5}$$

where  $X$  is an  $n \times p$  experimental design matrix for an ANOVA, and  $Z$  is an  $n \times s$  matrix of concomitant variables. To test hypotheses, we assume that

$$\epsilon \sim N_n(0, \sigma^2 I)$$

so that

$$Y \sim N_n(X\beta + Z\gamma, \sigma^2 I).$$

As we mentioned above, the concomitant variables  $Z$  are introduced only to sharpen the analysis, and thus the inference on the ANOVA is done AFTER the regression fit. The regression coefficients for the concomitant variables are tested after the ANOVA tests.

# Analysis of Covariance

## Estimation of $\gamma$

We do not need distributional assumptions to estimate  $\gamma$ . To estimate  $\gamma$ , we only need

$$E(\epsilon) = 0 \quad (\text{i.e., } E(Y) = X\beta + Z\gamma)$$

and

$$\text{Cov}(\epsilon) = \sigma^2 I .$$

To estimate  $\gamma$ , we break up model (5) into two orthogonal parts. Since

$$E(Y) = X\beta + Z\gamma ,$$

write

$$\begin{aligned} X\beta + Z\gamma &= X\beta + MZ\gamma + (I - M)Z\gamma \\ &= (X, MZ) \begin{pmatrix} \beta \\ \gamma \end{pmatrix} + (I - M)Z\gamma , \end{aligned}$$

where  $M = X(X'X)^{-1}X'$ . Since  $C(X) = C(X, MZ)$ , the first part of the model is overparameterized. We reparameterize from  $(\beta, \gamma)$  to  $\delta$ , and write

$$\begin{aligned} E(Y) &= X\delta + (I - M)Z\gamma \\ &= (X, (I - M)Z) \begin{pmatrix} \delta \\ \gamma \end{pmatrix} . \end{aligned}$$

# Analysis of Covariance

Now the two parts  $(X, (I - M)Z)$  are orthogonal since

$$X'(I - M)Z = (X' - X'M)Z = (X' - X')Z = 0.$$

Since,  $(X, (I - M)Z)$  are orthogonal, we can do the estimation of  $\delta$  and  $\gamma$  separately. We justify this in the following theorem.

## Theorem

Consider the linear model

$$Y = X_1\beta_1 + X_2\beta_2 + \epsilon \quad (6)$$

where  $E(\epsilon) = 0$ ,  $\text{Cov}(\epsilon) = \sigma^2 I$ ,  $X_1$  is  $n \times p_1$  of rank  $r_1$  and  $X_2$  is  $n \times p_2$  of rank  $r_2$ . Suppose

$$C(X_1) \perp C(X_2)$$

so that  $X_1'X_2 = 0_{p_1 \times p_2}$  and  $X_2'X_1 = 0_{p_2 \times p_1}$ . Then the least squares estimates of  $\beta_1$  and  $\beta_2$  satisfy

$$X_1\hat{\beta}_1 = M_1 Y$$

and

$$X_2\hat{\beta}_2 = M_2 Y$$

where

$$M_1 = X_1(X_1'X_1)^{-}X_1'$$

and

$$M_2 = X_2(X_2'X_2)^{-}X_2'.$$

# Analysis of Covariance

## Proof

Write  $X = (X_1, X_2)$  and  $\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$ , so that the linear model in (6) can be written as

$$Y = X\beta + \epsilon. \quad (7)$$

We know that the least squares solution to  $\beta$  for the model in (7) is given by

$$X\hat{\beta} = MY.$$

In terms of  $X = (X_1, X_2)$ ,

$$\begin{aligned} M &= X(X'X)^{-1}X' \\ &= (X_1, X_2) \begin{pmatrix} X'_1 X_1 & X'_1 X_2 \\ X'_2 X_1 & X'_2 X_2 \end{pmatrix}^{-1} \begin{pmatrix} X'_1 \\ X'_2 \end{pmatrix} \\ &= X_1(X'_1 X_1)^{-1} X'_1 + X_2(X'_2 X_2)^{-1} X'_2 \\ &= M_1 + M_2. \end{aligned}$$

Thus  $MY = (M_1 + M_2)Y = M_1 Y + M_2 Y$ . Now

$$X\hat{\beta} = (X_1, X_2) \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = X_1\hat{\beta}_1 + X_2\hat{\beta}_2.$$

Thus

$$X_1\hat{\beta}_1 + X_2\hat{\beta}_2 = M_1 Y + M_2 Y. \quad (8)$$

# Analysis of Covariance

Now multiplying both sides by  $X_2'$ , we get

$$X_2' X_1 \hat{\beta}_1 + X_2' X_2 \hat{\beta}_2 = X_2' M_1 Y + X_2' M_2 Y$$

which implies

$$X_2' X_2 \hat{\beta}_2 = X_2' Y .$$

since  $X_2' X_1 = 0$ . These are the normal equations for  $\hat{\beta}_2$ , which we know are equivalent to

$$X_2 \hat{\beta}_2 = M_2 Y .$$

Multiplying both sides of (8) by  $X_1'$  leads to

$$X_1 \hat{\beta}_1 = M_1 Y .$$

This completes the proof.

# Analysis of Covariance

The normal equations for  $(\delta, \gamma)$  are given by

$$(X, (I - M)Z)'(X, (I - M)Z) \begin{pmatrix} \delta \\ \gamma \end{pmatrix} = (X, (I - M)Z)'Y,$$

Carrying through the multiplication and noting that  $X'(I - M)Z = 0$ , the normal equations reduce to

$$\begin{pmatrix} X'X & 0 \\ 0 & Z'(I - M)Z \end{pmatrix} \begin{pmatrix} \delta \\ \gamma \end{pmatrix} = \begin{pmatrix} X'Y \\ Z'(I - M)Y \end{pmatrix}$$

Thus, a least squares estimate of  $\gamma$  is

$$\hat{\gamma} = (Z'(I - M)Z)^{-1}Z'(I - M)Y.$$

Since  $Z$  consists of concomitant variables,  $Z'(I - M)Z$  will be typically nonsingular. In this case, the unique least squares estimate of  $\gamma$  is

$$\hat{\gamma} = (Z'(I - M)Z)^{-1}Z'(I - M)Y.$$

Since  $X$  is the design matrix for an ANOVA,  $\delta$  (or  $\beta$ ) is usually not estimable. Thus we will concern ourselves with the estimation of  $X\delta$  (or  $X\beta$ ). The least squares equations for  $\delta$  are

$$X\hat{\delta} = MY.$$

# Analysis of Covariance

Thus, if we want the least squares estimate of  $X\beta$ , we note that

$$X\delta + (I - M)Z\gamma = X\beta + MZ\gamma + (I - M)Z\gamma$$

so that

$$X\hat{\delta} + (I - M)Z\hat{\gamma} = X\hat{\beta} + MZ\hat{\gamma} + (I - M)Z\hat{\gamma}.$$

subtracting  $(I - M)Z\hat{\gamma}$  from both sides gives

$$X\hat{\delta} = X\hat{\beta} + MZ\hat{\gamma}.$$

Therefore,

$$\begin{aligned} X\hat{\beta} &= X\hat{\delta} - MZ\hat{\gamma} \\ &= MY - MZ(Z'(I - M)Z)^{-1}Z'(I - M)Y \\ &= M(Y - Z\hat{\gamma}) \end{aligned}$$

# Analysis of Covariance

Notice that if  $Z'(I - M)Z$  is singular, then  $X\beta$  is not estimable. Note that

$$\begin{aligned}E(X\hat{\beta}) &= E(MY - MZ\hat{\gamma}) \\&= ME(Y) - MZE(\hat{\gamma}) \\&= M(X\beta + Z\gamma) - MZE(\hat{\gamma}) \\&= X\beta + MZ\gamma - MZ(Z'(I - M)Z)^{-1}Z'(I - M)(X\beta + Z\gamma) \\&= X\beta + MZ\gamma - MZ(Z'(I - M)Z)^{-1}Z'(I - M)Z\gamma\end{aligned}$$

If  $Z'(I - M)Z$  is nonsingular, then of course  $\gamma$  is estimable. Also, in this case,  $X\beta$  is estimable. In particular, if  $Z'(I - M)Z$  is nonsingular, then any function of  $\beta$  that's estimable in the ANOVA model is also estimable in the ANCOVA model.

When  $Z'(I - M)Z$  is singular, we need to characterize the estimable functions of  $\gamma$  and  $\beta$ . We are led to the following theorem.

# Analysis of Covariance

## Theorem

$\xi' \gamma$  is estimable if and only if  $\xi' = \rho'(I - M)Z$  for some vector  $\rho \in R^n$ .

This theorem says that the estimable functions of  $\gamma$  are those that are linear functions of  $(I - M)Z$ .

## Proof

If  $\xi' \gamma$  is estimable, then there exists a  $\rho$  such that

$$\xi' \gamma = \rho'(X, Z) \begin{pmatrix} \beta \\ \gamma \end{pmatrix}$$

so that

$$\rho'(X, Z) = (0, \xi')$$

and  $\rho' X = 0$ . Therefore,  $\xi' = \rho' Z = \rho'(I - M)Z$ . Conversely, if  $\xi' = \rho'(I - M)Z$ , then

$$\rho'(I - M)(X, Z) \begin{pmatrix} \beta \\ \gamma \end{pmatrix} = \xi' \gamma .$$

# Analysis of Covariance

## Corollary

The unique BLUE of  $\rho'(I - M)Z\gamma$  is

$$\rho' M_{(I-M)Z} Y$$

where

$$M_{(I-M)Z} = (I - M)Z(Z'(I - M)Z)^{-1}Z'(I - M)$$

is the orthogonal projection operator onto  $C((I - M)Z)$ .

Finally, we note that if  $Z'(I - M)Z$  is nonsingular, so that  $(X\beta, \gamma)$  are both estimable, then

$$\text{Cov} \begin{pmatrix} X\hat{\beta} \\ \hat{\gamma} \end{pmatrix} = \sigma^2 \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$$

where

$$A_{11} = M + MZ(Z'(I - M)Z)^{-1}Z'M$$

$$A_{12} = -MZ(Z'(I - M)Z)^{-1}$$

$$A_{21} = A'_{12}$$

and

$$A_{22} = (Z'(I - M)Z)^{-1}.$$

# Analysis of Covariance

## Estimation of $\sigma^2$

Let  $P$  denote the orthogonal projection operator onto  $C(X, Z)$ . Then, the error sum of squares (SSE) for the ANCOVA model is given by

$$SSE = \|(I - P)Y\|^2 = Y'(I - P)Y.$$

We would like to express  $P$  in terms of  $X$  and  $Z$ . We have established that

$$C(X, Z) = C(X, (I - M)Z).$$

Since

$$C(X) \perp C((I - M)Z),$$

the orthogonal projection operator for  $C(X, (I - M)Z)$  is the sum of the orthogonal projection operators onto  $C(X)$  and  $C(X, (I - M)Z)$ , respectively.

Thus

$$P = M + M_{(I-M)Z}$$

where

$$M = X(X'X)^{-}X'$$

and

$$M_{(I-M)Z} = (I - M)Z(Z'(I - M)Z)^{-}Z'(I - M).$$

Thus,

$$SSE = \left\| \left( I - \left[ M + (I - M)Z(Z'(I - M)Z)^{-}Z'(I - M) \right] \right) Y \right\|^2.$$

# Analysis of Covariance

Define the notation

$$E_{AB} = A'(I - M)B$$

where  $A$  and  $B$  are arbitrary matrices. Thus, we can write

$$\begin{aligned}\|(I - P)Y\|^2 &= Y'(I - P)Y \\ &= E_{yy} - E_{yz}E_{zz}^{-}E_{zy}.\end{aligned}$$

To prove this identity, note that

$$\begin{aligned}&Y'(I - P)Y \\ &= Y'(I - M - (I - M)Z(Z'(I - M)Z)^{-}Z'(I - M))Y \\ &= Y'(I - M)Y - Y'(I - M)Z(Z'(I - M)Z)^{-}Z'(I - M)Y \\ &= E_{yy} - E_{yz}E_{zz}^{-}E_{zy}.\end{aligned}$$

Tests of hypotheses for ANCOVA are obtained by considering the reductions in the SSE's for the models being tested. We note here that the primary interest in tests of hypothesis are tests concerning the treatment effects. Hypotheses tests concerning  $\gamma$  are usually not of interest in ANCOVA. If  $X_0$  is such that  $C(X_0) \subset C(X)$ , we may wish to test the reduced model

$$Y = X_0\gamma_0 + Z\gamma + \epsilon$$

against

$$Y = X\beta + Z\gamma + \epsilon.$$

# Analysis of Covariance

We can write the hypotheses as

$$H_0 : E(Y) \in C(X_0, Z)$$

$$H_a : E(Y) \in C(X, Z) \cap (C(X_0, Z))^c$$

The  $F$  test for nested models applies here and is given by

$$\begin{aligned} F &= \frac{\|(P - P_0)Y\|^2 / r(P - P_0)}{\|(I - P)Y\|^2 / r(I - P)} \\ &\sim F(r(P - P_0), r(I - P), \gamma^*) \end{aligned}$$

where

$$\gamma^* = \frac{\|(P - P_0)(X\beta + Z\gamma)\|^2}{2\sigma^2}$$

and  $P_0$  is the orthogonal projection operator onto  $C(X_0, Z)$ . Thus,

$$P_0 = M_0 + (I - M_0)Z(Z'(I - M_0)Z)^{-1}Z'(I - M_0)$$

where

$$M_0 = X_0(X_0'X_0)^{-1}X_0'$$

is the orthogonal projection operator onto  $C(X_0)$ .

# Analysis of Covariance

The  $F$  test above can also be written as

$$\begin{aligned} F &= \frac{Y'(P - P_0)Y/r(P - P_0)}{Y'(I - P)Y/r(I - P)} \\ &= \frac{[Y'(I - P_0)Y - Y'(I - P)Y]/(r(P) - r(P_0))}{Y'(I - P)Y/(n - r(P))} \end{aligned}$$

where  $n$  is the total number of observations. We see that numerator of the  $F$  test can be written as a difference in the error sums of squares between the nested models.

# Analysis of Covariance

## Example

Consider the balanced two-way ANOVA model with no replication, (and hence no interaction), and one covariate. The ANCOVA model can be written as

$$Y_{ij} = \mu + \alpha_i + \eta_j + \gamma z_{ij} + \epsilon_{ij}$$

for  $i = 1, \dots, a$  and  $j = 1, \dots, b$ . We can write this model in the matrix form

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}$$

where  $\mathbf{X}$  is the usual  $n \times (a + b + 1)$  design matrix for a two-way ANOVA without interaction with  $n = abN$ ,  $N = 1$ ,

$$\mathbf{Z} = \begin{pmatrix} z_{11} \\ \vdots \\ z_{1b} \\ z_{21} \\ \vdots \\ z_{ab} \end{pmatrix}_{ab \times 1},$$

# Analysis of Covariance

$\gamma$  is a scalar, and

$$\beta = \begin{pmatrix} \mu \\ \alpha_1 \\ \vdots \\ \alpha_a \\ \eta_1 \\ \vdots \\ \eta_b \end{pmatrix}_{(a+b+1) \times 1} .$$

We can write the sums of squares for this ANCOVA model as follows.

$$E_{yy} = Y'(I - M)Y = \sum_{i=1}^a \sum_{j=1}^b (Y_{ij} - \bar{Y}_{i\cdot} - \bar{Y}_{\cdot j} + \bar{Y}_{\cdot\cdot})^2 ,$$

$$\begin{aligned} E_{yz} & \\ &= Y'(I - M)Z \\ &= \sum_{i=1}^a \sum_{j=1}^b (Y_{ij} - \bar{Y}_{i\cdot} - \bar{Y}_{\cdot j} + \bar{Y}_{\cdot\cdot}) (Z_{ij} - \bar{Z}_{i\cdot} - \bar{Z}_{\cdot j} + \bar{Z}_{\cdot\cdot}) . \end{aligned}$$

# Analysis of Covariance

We note here that since  $Z$  is a vector in this example,

$$Y'(I - M)Z = Z'(I - M)Y, \text{ so that } E_{yz} = E_{zy}$$

Finally, we have

$$E_{zz} = Z'(I - M)Z = \sum_{i=1}^a \sum_{j=1}^b (Z_{ij} - \bar{Z}_{i\cdot} - \bar{Z}_{\cdot j} + \bar{Z}_{\cdot\cdot})^2 ,$$

and

$$SSE = Y'(I - P)Y = E_{yy} - E_{yz} E_{zz}^- E_{zy} .$$

Suppose we are interested in testing

$$H_0 : \gamma = 0$$

$$H_a : \gamma \neq 0 .$$

The reduced model is given by

$$Y_{ij} = \mu + \alpha_i + \eta_j + \epsilon_{ij} .$$

Thus,  $X_0$  is the design matrix for the two-way ANOVA without interaction.

Thus,  $P_0 = X_0(X_0'X_0)^{-1}X_0'$ , and

$$Y'(I - P_0)Y = \sum_{i=1}^a \sum_{j=1}^b (Y_{ij} - \bar{Y}_{i\cdot} - \bar{Y}_{\cdot j} + \bar{Y}_{\cdot\cdot})^2 .$$

# Analysis of Covariance

Note here that  $r(P_0) = 1 + a - 1 + b - 1 = a + b - 1$  and

Thus, the  $F$  test takes the form

$$\begin{aligned} F &= \frac{\|(P - P_0)Y\|^2/r(P - P_0)}{\|(I - P)Y\|^2/r(I - P)} \\ &= \frac{Y'(P - P_0)Y/r(P - P_0)}{Y'(I - P)Y/r(I - P)} \\ &= \frac{[Y'(I - P_0)Y - Y'(I - P)Y]/r(P - P_0)}{Y'(I - P)Y/r(I - P)} \\ &= \frac{(E_{yy} - [E_{yy} - E_{yz}E_{zz}^{-1}E_{zy}])/r(P - P_0)}{Y'(I - P)Y/r(I - P)} \\ &= \frac{E_{yz}E_{zz}^{-1}E_{zy}/1}{Y'(I - P)Y/(n - a - b)} \\ &\sim F(1, n - a - b, \gamma^*) \end{aligned}$$

where  $n = ab$  and

$$\gamma^* = \frac{\|(P - P_0)(X\beta + Z\gamma)\|^2}{2\sigma^2}.$$

# Analysis of Covariance

We note here that

$$r(P_0) = 1 + (a - 1) + (b - 1) = a + b - 1$$

and

$$r(P) = 1 + (a - 1) + (b - 1) + 1 = a + b ,$$

so that

$$\begin{aligned} r(P - P_0) &= r(P) - r(P_0) \\ &= a + b - (a + b - 1) = 1 \end{aligned}$$

and

$$\begin{aligned} r(I - P) &= r(I) - r(P) \\ &= ab - (a + b) \\ &= ab - a - b \\ &= n - a - b . \end{aligned}$$

# Analysis of Covariance

## Sums of Squares for the General ANCOVA Model

Consider the general ANCOVA model

$$Y = X\beta + Z\gamma + \epsilon .$$

We wish to construct the general ANCOVA table. To do so, we need to construct sums of squares for  $Y$ , denoted  $SS_{yy}$ , the sums of squares for  $Z$ , denoted  $SS_{zz}$ , and the sum of cross products of  $Y$  and  $Z$ , denoted  $SS_{yz}$ . To illustrate how this can be done, suppose the ANOVA part of the model is a two-way balanced ANOVA with interaction. We have a total of  $n = abN$  observations. Also,

$$C(X) = C(M_\mu) + C(M_\alpha) + C(M_\eta) + C(M_{\alpha\eta}) .$$

We can write the ANCOVA table as

source	df	$SS_{yy}$	$SS_{yz}$	$SS_{zz}$
$\mu$	1	$Y'M_\mu Y$	$Y'M_\mu Z$	$Z'M_\mu Z$
$\alpha$	$a - 1$	$Y'M_\alpha Y$	$Y'M_\alpha Z$	$Z'M_\alpha Z$
$\eta$	$b - 1$	$Y'M_\eta Y$	$Y'M_\eta Z$	$Z'M_\eta Z$
$\alpha\eta$	$(a - 1)(b - 1)$	$Y'M_{\alpha\eta} Y$	$Y'M_{\alpha\eta} Z$	$Z'M_{\alpha\eta} Z$
Error	$n - ab - r$	$Y'(I - M)Y$	$Y'(I - M)Z$	$Z'(I - M)Z$
Total	$n$	$Y'Y$	$Y'Z$	$Z'Z$

Thus, in ANCOVA, we still partition  $R^n$  into a sum of orthogonal subspaces.

# Analysis of Covariance

To test for no interaction for this model, our hypothesis is

$$H_0 : (\alpha\eta)_{11} = \dots = (\alpha\eta)_{ab} .$$

To construct the  $F$  test, we need  $P_0$  and  $P$ . We have

$$P_0 = M_0 + (I - M_0)Z(Z'(I - M)Z)^{-1}Z'(I - M_0)$$

where

$$M_0 = M - M_{\alpha\eta} .$$

Thus, the numerator sums of squares for the  $F$  test is

$$\begin{aligned} & \| (P - P_0)Y \|^2 \\ &= Y'(P - P_0)Y \\ &= Y'(M - M_{\alpha\eta})Y \\ &+ Y'(I - M + M_{\alpha\eta})Z(Z'(I - M)Z)^{-1}Z'(I - M + M_{\alpha\eta})Y \\ &- Y' [M + (I - M)Z(Z'(I - M)Z)^{-1}Z'(I - M)] Y \\ &= Y'(I - M + M_{\alpha\eta}Z(Z'(I - M)Z)^{-1}Z'(I - M + M_{\alpha\eta}))Y \\ &- Y'(I - M)Z(Z'(I - M)Z)^{-1}Z'(I - M)Y - Y'M_{\alpha\eta}Y \\ &= (Y'(I - M)Z + Y'M_{\alpha\eta}Z) \\ &\times (Z'(I - M)Z + Z'M_{\alpha\eta}Z)^{-1} (Z'(I - M)Y + Z'M_{\alpha\eta}Y) \\ &- Y'(I - M)Z(Z'(I - M)Z)^{-1}Z'(I - M)Y - Y'M_{\alpha\eta}Y \end{aligned}$$

# Analysis of Covariance

The decomposition of  $Y'(I - P)Y$  into terms involving only  $Y$ ,  $M$ , and  $Z$  is similar. Note that all of these terms can be obtained from the ANCOVA table.

# Analysis of Covariance

## Applications of ANCOVA: Missing Data

Suppose some responses are missing from a balanced ANOVA model. We can rewrite the ANOVA model with missing observations in terms of an ANCOVA model. The ANOVA model with missing observations can be written as

$$Y = X\beta + \epsilon \quad (1)$$

where

$$Y = (Y_1, \dots, Y_n)' .$$

Note here that we have  $n$  responses, some of which may be missing. The indices of the components of  $Y$  have single subscripts to make notation simpler. Now suppose  $r$  components of  $Y$  are missing. Without loss of generality, we assume the last  $r$  components of  $Y$  are missing. Write

$$Y = (Y_{n-r}, Y_r)'$$

so that  $Y_{n-r} = (Y_1, \dots, Y_{n-r})'$ .

The model with the missing observations deleted is given by

$$Y_{n-r} = X_{n-r}\beta + \epsilon_{n-r} \quad (2)$$

where  $X_{n-r}$  is the  $(n - r) \times p$  design matrix corresponding to  $Y_{n-r}$ . The model in (2) now corresponds to an unbalanced ANOVA model.

# Analysis of Covariance

We can obtain estimates based on (1) by rewriting (1) in terms of an ANCOVA model. We rewrite (1) in terms of an ANCOVA model as follows. For each missing observation  $Y_i$ , we introduce a covariate vector

$$Z_i = (0, \dots, 0, 1, 0, \dots, 0)'$$

where the 1 is in the  $i$ th position. We will show that estimates based on the ANCOVA formulation are equivalent to estimates based on the model with the missing observations deleted, given by (2).

If the last  $r$  observations are missing, then  $Z$  consists of an  $n \times r$  matrix of covariates, and has the form

$$Z = \begin{pmatrix} 0 \\ I_r \end{pmatrix},$$

where  $0$  is an  $(n - r) \times r$  matrix of zeroes and  $I_r$  is the  $r \times r$  identity matrix. We can write the design matrix  $X$  from model (1) as

$$X = \begin{pmatrix} X_{n-r} \\ X_r \end{pmatrix}$$

where  $X_r$  is the  $r \times p$  matrix whose rows correspond to the missing observations.

# Analysis of Covariance

The ANCOVA formulation of model (1) can now be written as

$$\begin{aligned} Y &= \begin{pmatrix} X_{n-r} \\ X_r \end{pmatrix} \beta + \begin{pmatrix} 0 \\ I_r \end{pmatrix} \gamma + \epsilon \\ &= \begin{pmatrix} X_{n-r} & 0 \\ X_r & I_r \end{pmatrix} \begin{pmatrix} \beta \\ \gamma \end{pmatrix} + \epsilon, \end{aligned} \quad (3)$$

where

$$Y = \begin{pmatrix} Y_{n-r} \\ 0 \end{pmatrix}.$$

We note here that

$$C\left(\begin{pmatrix} X_{n-r} & 0 \\ X_r & I_r \end{pmatrix}\right) = C\left(\begin{pmatrix} X_{n-r} & 0 \\ 0 & I_r \end{pmatrix}\right)$$

since  $C(X_r) \subset C(I_r)$ . Now let  $M_{n-r}$  denote the orthogonal projection operator onto  $C(M_{n-r})$ , and let  $P$  denote the orthogonal projection operator onto

$$C\left(\begin{pmatrix} X_{n-r} & 0 \\ 0 & I_r \end{pmatrix}\right).$$

# Analysis of Covariance

Then,

$$\begin{aligned} P &= \begin{pmatrix} X_{n-r} & 0 \\ 0 & I_r \end{pmatrix} \\ &\times \left[ \left( \begin{pmatrix} X_{n-r} & 0 \\ 0 & I_r \end{pmatrix}' \begin{pmatrix} X_{n-r} & 0 \\ 0 & I_r \end{pmatrix} \right) \right]^{-1} \\ &\times \begin{pmatrix} X_{n-r} & 0 \\ 0 & I_r \end{pmatrix}' \\ &= \begin{pmatrix} M_{n-r} & 0 \\ 0 & I_r \end{pmatrix}. \end{aligned}$$

Now the residual sums of squares based on the ANCOVA formulation is

$$\begin{aligned} &Y'(I - P)Y \\ &= (Y'_{n-r}, 0) \begin{pmatrix} M_{n-r} & 0 \\ 0 & I_r \end{pmatrix} \begin{pmatrix} Y_{n-r} \\ 0 \end{pmatrix} \\ &= Y'_{n-r}(I - M_{n-r})Y_{n-r} \end{aligned}$$

Thus, we have shown that the error sums of squares based on the ANCOVA formulation is equal to the error sums of squares obtained from the model with the missing observations deleted given by (2). Since the model with the missing observations deleted results in an unbalanced ANOVA model, the error sums of squares from the ANCOVA formulation (model (3)) is equal to the error sums of squares based on the unbalanced ANOVA model given in (2).

# Analysis of Covariance

Estimable functions of  $\beta$  will be the same for models (2) and (3). To see this, note that  $E(Y)$  from model (3) is

$$E(Y) = \begin{pmatrix} X_{n-r}\beta \\ X_r\beta + \gamma \end{pmatrix} \quad (4)$$

The estimate of  $E(Y)$  in (4) is thus

$$PY = \begin{pmatrix} M_{n-r}Y_{n-r} \\ 0 \end{pmatrix} .$$

More generally, anything that is estimable in model (3) is also estimable in model (2), and the estimates are identical. To see this, note that any estimable function of  $\beta$  in the model (2) is of the form

$$\rho'_{n-r}X_{n-r}\beta , \quad (5)$$

where  $\rho_{n-r}$  is a vector in  $R^{n-r}$ . Any estimable function of  $(\beta, \gamma)$  for the model in (3) is of the form

$$\rho'(X\beta + Z\gamma) \quad (6)$$

where  $\rho \in R^n$ . Note that (5) and (6) are equal if and only if  $\rho' = (\rho'_{n-r}, 0)$ .

# Analysis of Covariance

To see this we have

$$\begin{aligned}\rho'(X\beta + Z\gamma) &= (\rho'_{n-r}, 0) \left( \begin{pmatrix} X_{n-r} \\ X_r \end{pmatrix} \beta + \begin{pmatrix} 0 \\ I_r \end{pmatrix} \gamma \right) \\ &= (\rho'_{n-r} X_{n-r} \beta + 0) + (0 + 0) \\ &= \rho'_{n-r} X_{n-r} \beta.\end{aligned}$$

Thus, with this form of  $\rho$ , least squares estimates of estimable functions are identical from models (2) and (3) are identical, and are given by

$$\begin{aligned}\rho' PY &= (\rho'_{n-r}, 0) \begin{pmatrix} M_{n-r} & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} Y_{n-r} \\ 0 \end{pmatrix} \\ &= \rho'_{n-r} M_{n-r} Y_{n-r}.\end{aligned}$$

An alternative approach to missing data is based on finding estimates of the missing values and then doing a complete data analysis with those estimates. This procedure produces different estimates, in general, from the procedure in which we delete all of the missing observations. However, if we choose the estimates of the missing values in a certain way, then the analysis based on estimating the missing values will yield estimates identical to the analysis based on deleting the missing observations.

# Analysis of Covariance

The key point in this procedure is how to choose the estimates of the missing values. The estimates of the missing values are chosen so that if the estimates are in fact the real data, then the “correct” error sums of squares is computed by the ANCOVA model. That is, if we estimate the missing values, and then do a complete data analysis with the “filled-in” values, the error sums of squares from the complete data analysis will equal the error sums of squares from the ANCOVA model (3).

The procedure for estimating the missing values is as follows.

- 1) We write down the ANCOVA formulation of model (1). The ANCOVA formulation is

$$Y = X\beta + Z\gamma + \epsilon$$

where

$$X = \begin{pmatrix} X_{n-r} \\ X_r \end{pmatrix}$$

and

$$Z = \begin{pmatrix} 0 \\ I_r \end{pmatrix} .$$

- 2) Assuming  $\gamma$  is estimable, estimate  $\gamma$  from the ANCOVA model. This estimate is given by

$$\hat{\gamma} = (Z'(I - M)Z)^{-1}Z'(I - M)Y .$$

# Analysis of Covariance

- 3) Once  $\hat{\gamma}$  is obtained, construct the vector

$$\begin{aligned} Y^* = Y - Z\hat{\gamma} &= \begin{pmatrix} Y_{n-r} \\ 0 \end{pmatrix} - \begin{pmatrix} 0 \\ I_r \end{pmatrix} \hat{\gamma} \\ &= \begin{pmatrix} Y_{n-r} \\ -\hat{\gamma} \end{pmatrix}. \end{aligned}$$

Thus  $Y^*$  is now a “complete data” vector of responses.

# Analysis of Covariance

- 4) Using  $Y^*$ , fit the complete data balanced ANOVA as usual.

We note here that the missing response values get replaced by  $-\hat{\gamma}$ . To get a better feel of what the estimates of the missing values represent, we recall that

$$\hat{\gamma} = (Z'(I - M)Z)^{-1} Z'(I - M)Y$$

where  $M = X(X'X)^{-1}X'$ . partition  $M$  into

$$M = \begin{pmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{pmatrix}$$

where  $M_{11}$  is  $(n - r) \times r$ ,  $M_{22}$  is  $r \times r$ , and  $M_{12} = M'_{21}$ . We assume  $M_{22}$  is positive definite. Then

$$\begin{aligned} -\hat{\gamma} &= - \left[ (0', I_r) \begin{pmatrix} I_{n-r} - M_{11} & -M_{12} \\ -M_{21} & I_r - M_{22} \end{pmatrix} \begin{pmatrix} 0 \\ I_r \end{pmatrix} \right]^{-1} \\ &\times (0', I_r) \begin{pmatrix} I_{n-r} - M_{11} & -M_{12} \\ -M_{21} & I_r - M_{22} \end{pmatrix} \begin{pmatrix} Y_{n-r} \\ 0 \end{pmatrix} \\ &= (I_r - M_{22})^{-1} M_{21} Y_{n-r}. \end{aligned}$$

# Analysis of Covariance

## Remark

We note that  $-\hat{\gamma}$  corresponds to estimating the missing values by linear functions of  $Y_{n-r}$ . Thus,  $-\hat{\gamma}$  can be viewed as the “predicted” values of the missing observations.

To show that estimating the missing values by  $-\hat{\gamma}$  yields the correct error sums of squares, we compute error sums of squares from the complete data ANOVA model with the filled-in values. Our response vector with the filled-in data is

$$Y^* = Y - Z\hat{\gamma}.$$

The error sums of squares from the complete data ANOVA is thus given by

$$\begin{aligned} & Y'^*(I - M)Y^* \\ &= (Y - Z\hat{\gamma})'(I - M)(Y - Z\hat{\gamma}) \\ &= Y'(I - M)Y - 2\hat{\gamma}Z'(I - M)Y + \hat{\gamma}Z'(I - M)Z\hat{\gamma} \\ &= Y'(I - M)Y - 2Y'(I - M)Z(Z'(I - M)Z)^{-1}Z'(I - M)Y \\ &+ Y'(I - M)Z(Z'(I - M)Z)^{-1}(Z'(I - M)Z) \\ &+ (Z'(I - M)Z)^{-1}Z'(I - M)Y \\ &= Y'(I - M)Y - Y'(I - M)Z(Z'(I - M)Z)^{-1}Z'(I - M)Y \\ &= Y'(I - P)Y \\ &= Y'_{n-r}(I - M_{n-r})Y_{n-r} \end{aligned}$$

# Analysis of Covariance

Thus, the error sums of squares computed from the complete data ANOVA model is the same as the error sums of squares from the ANCOVA formulation. Also, the estimate of  $X\beta$  using the complete data model is

$$\begin{aligned} MY^* &= M(Y - Z\hat{\gamma}) \\ &= My - MZ\hat{\gamma} \\ &= X\hat{\beta}. \end{aligned}$$

Thus, the estimate of  $X\beta$  from the complete data model

$$Y^* = X\beta + \epsilon$$

is the same as the estimate of  $X\beta$  from the ANCOVA model. Also for any estimable function  $\rho'X\beta$ , we have

$$\text{Var}(\rho'X\hat{\beta}) = \sigma^2(\rho'M\rho + \rho'MZ(Z'(I-M)Z)^{-1}Z'M\rho).$$

This the variance obtained by the complete data model and the ANCOVA model.

# Analysis of Covariance

## Remark 1

The procedure of estimating the missing values by  $-\hat{\gamma}$  and then doing a complete data analysis produces identical estimates to the procedure of deleting the missing values and then doing the analysis. In summary, we have the estimates of estimable functions from

$$Y_{n-r} = X_{n-r}\beta + \epsilon$$

are identical to the estimates from

$$Y = X\beta + Z\gamma + \epsilon$$

which are identical to estimates from

$$Y^* = X\beta + \epsilon$$

where  $Y^* = Y - Z\hat{\gamma}$ .

## Remark 2

Note here we have made no mention of the mechanism that generated the missing data. We have assumed in our setup here that the mechanism generating the missing data can be ignored for making inferences about the parameters.

### Remark 3

A more common problem in missing data is missing data in the covariates in the context of regression. We will discuss missing covariates in detail when we reach Chapter 6. Missing covariates does not make sense in ANOVA.

# Analysis of Covariance

## Balanced Incomplete Block Designs (BIB)

The ANCOVA model can be used to develop the analysis of the balanced incomplete block (BIB) design model. The BIB design proceeds as follows.

Suppose a design is to be set up with  $b$  blocks and  $t$  treatments, but the number of treatments that can be observed in any block is  $k$ , where  $k < t$ . If a pair of treatments occur together in the same block a fixed number of times, say  $\lambda$ , then such a design is called a BIB design. If a pair of treatments occur together in the same block a variable number of times, then the design is called an incomplete block design.

### Example

Suppose we have 4 treatments denoted A, B, C, and D, and 4 blocks. Consider the design,

block 1	block 2	block 3	block 4
A	A	A	-
B	B	-	B
C	-	C	C
-	D	D	D

here we have  $t = 4$ ,  $b = 4$ ,  $\lambda = 2$ , and  $k = 3$ .

# Analysis of Covariance

Let  $r$  denote the number of occurrences for each treatment. In our example,  $r = 3$ . Two general identities are

$$rt = bk$$

and

$$(t - 1)\lambda = r(k - 1) .$$

Thus, the total number of observations is  $n = rt = bk$ . The latter identity implies that the number of within block comparisons between any given treatment and the other treatment is fixed.

The balanced incomplete block model can be written as

$$Y_{ij} = \mu + \beta_i + \tau_j + \epsilon_{ij} \tag{7}$$

where the  $\epsilon_{ij}$ 's are *i.i.d.*  $N(0, \sigma^2)$  random variables,  $i = 1, \dots, b$ , and  $j \in D_i$ , where  $D_i$  is a set of indices of the treatments in block  $i$ . We can also write  $j = 1, \dots, t$ , and  $i \in A_j$ , where  $A_j$  is the set of indices of the blocks in which treatment  $j$  occurs.

# Analysis of Covariance

We have  $k$  elements in each set  $D_i$  and  $r$  elements in each set  $A_j$ . We can now write the BIB design as an ANCOVA model. The ANOVA part of the ANCOVA model is a one-way ANOVA model with the grand mean and the blocks as the terms, and the regression part of the ANCOVA model are the columns of the design matrix corresponding to the treatments. We can write the BIB model as

$$Y = (X, Z) \begin{pmatrix} \beta \\ \tau \end{pmatrix} + \epsilon$$

where

$$\beta = (\mu, \beta_1, \dots, \beta_b)'$$

and

$$\tau = (\tau_1, \dots, \tau_t)'.$$

Thus, the ANCOVA model takes the form

$$Y = X\beta + Z\tau + \epsilon \tag{8}$$

where  $X$  is  $bk \times (b+1)$  and  $Z$  is  $bk \times t$ .

Our primary interest in the BIB design is in assessing treatment effects, i.e., the  $\tau$ 's. These are the coefficients of the covariates in the ANCOVA model. By our previous results for ANCOVA, we know that

$$\hat{\tau} = (Z'(I - M)Z)^{-1} Z'(I - M)Y.$$

# Analysis of Covariance

Thus we focus our discussion on the computation of the matrices needed to construct  $\hat{\tau}$ . Write

$$Z = (Z_1, \dots, Z_t)_{bk \times t}$$

where each  $Z_i$  is  $bk \times 1$ . The rows of the  $m$ th column of  $Z$  indicate presence or absence of the  $m$ th treatment. We can write the  $m$ th column as

$$Z_m = (z_{ij,m})$$

where

$$z_{ij,m} = \delta_{jm}$$

and  $\delta_{jm} = 1$  if  $j = m$ , and 0 otherwise. Thus,  $Z_m = 0$  for all rows except the  $r$  rows that correspond to an observation on treatment  $m$ , and those  $r$  rows all equal 1. We first need to obtain  $Z'(I - M)Z = Z'Z - Z'MZ$ . We do this by finding  $Z'_m Z_s$  and  $Z'_m M Z_s = Z'_m M M Z_s$  for all values of  $m$  and  $s$ .

# Analysis of Covariance

First, for  $m = s$ , we have

$$\begin{aligned} Z_m' Z_m &= \sum_i \sum_j (z_{ij,m})^2 \\ &= \sum_{j=1}^t \sum_{i \in A_j} \delta_{jm} \\ &= \sum_{j=1}^t r \delta_{jm} \\ &= r . \end{aligned}$$

Now if  $m \neq s$ , then

$$\begin{aligned} Z_s' Z_m &= \sum_i \sum_j (z_{ij,s})(z_{ij,m}) \\ &= \sum_{j=1}^t \sum_{i \in A_j} \delta_{js} \delta_{jm} \\ &= \sum_{j=1}^t r \delta_{js} \delta_{jm} \\ &= 0 \end{aligned}$$

# Analysis of Covariance

Thus  $Z'Z = rI_{t \times t}$ . To compute  $Z'MZ$ , write

$$M = (v_{ij,i'j'})$$

where

$$v_{ij,i'j'} = \frac{1}{k} \delta_{ii'} .$$

Let

$$MZ_m = (d_{ij,m}) .$$

Then,

$$\begin{aligned} d_{ij,m} &= \sum_{i'j'} v_{ij,i'j'} z_{i'j',m} \\ &= \sum_{j'=1}^t \sum_{i' \in A_j'} \frac{1}{k} \delta_{ii'} \delta_{j'm} \\ &= \sum_{i' \in A_m} \frac{1}{k} \delta_{ii'} \\ &= \frac{1}{k} \delta_i(A_m) , \end{aligned}$$

where  $\delta_i(A_m) = 1$  if  $i \in A_m$  and 0 otherwise.

# Analysis of Covariance

Thus, if treatment  $m$  is in block  $i$ , then all  $k$  of the units in block  $i$  have

$$d_{ij,m} = 1/k .$$

If treatment  $m$  is not in block  $i$ , then all  $k$  of the units in block  $i$  have  $d_{ij,m} = 0$ . Since treatment  $m$  is contained in exactly  $r$  blocks, we have

$$\begin{aligned} Z_m' M M Z_m &= \sum_{ij} (d_{ij,m})^2 \\ &= \sum_{i=1}^b \sum_{j \in D_i} k^{-2} \delta_i(A_m) \\ &= \sum_{i=1}^b \left( \frac{k}{k^2} \right) \delta_i(A_m) \\ &= r/k \end{aligned}$$

# Analysis of Covariance

Thus, since for  $s \neq m$ , there are  $\lambda$  blocks in which both treatments  $s$  and  $m$  are contained,

$$\begin{aligned} Z'_s M M Z_m &= \sum_{ij} (d_{ij,s})(d_{ij,m}) \\ &= \sum_{i=1}^b \sum_{j \in D_i} \frac{1}{k^2} \delta_i(A_s) \delta_i(A_m) \\ &= \sum_{i=1}^b \left( \frac{k}{k^2} \right) \delta_i(A_s) \delta_i(A_m) \\ &= \lambda/k \end{aligned}$$

From this derivation, it follows that  $Z' M Z$  has values  $r/k$  down the diagonal and values of  $\lambda/k$  off the diagonal. Thus,

$$\begin{aligned} Z' M Z &= \frac{1}{k} ((r - \lambda)I + \lambda J_t^t) \\ &= \begin{pmatrix} r/k & \lambda/k & \dots & \dots & \lambda/k \\ \lambda/k & r/k & \lambda/k & \dots & \lambda/k \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \lambda/k & \dots & \dots & \dots & r/k \end{pmatrix} \end{aligned}$$

# Analysis of Covariance

and hence

$$\begin{aligned} Z'(I - M)Z &= rI - k^{-1}((r - \lambda)I + \lambda J_t^t) \\ &= k^{-1}[(r(k - 1) + \lambda)I - \lambda J_t^t] . \end{aligned}$$

Now define

$$W = I - \frac{1}{t} J_t^t . \quad (9)$$

recall that

$$r(k - 1) + \lambda = (t - 1)\lambda + \lambda = \lambda t$$

and this leads to

$$\begin{aligned} Z'(I - M)Z &= \frac{\lambda}{k}(tI - J_t^t) \\ &= \left(\frac{\lambda t}{k}\right) W \end{aligned} \quad (10)$$

We note here that  $W$  is an orthogonal projection operator, so that  $W^- = W$ .

# Analysis of Covariance

Thus,

$$\begin{aligned}(Z'(I - M)Z)^{-1} &= \left(\frac{\lambda t}{k} W\right)^{-1} \\ &= \frac{k}{\lambda t} W^{-1} \\ &= \frac{k}{\lambda t} W.\end{aligned}$$

Now

$$\begin{aligned}Y'(I - M)Z_m &= \sum_{ij} (y_{ij} - \bar{y}_{i.}) z_{ij,m} \\ &= \sum_{i \in A_m} (y_{im} - \bar{y}_{i.}).\end{aligned}$$

Now define

$$Q_m = \sum_{i \in A_m} (y_{im} - \bar{y}_{i.}).$$

$Q_m$  is called the adjusted treatment total for treatment  $m$ . Then

$$Y'(I - M)Z = (Q_1, \dots, Q_t).$$

# Analysis of Covariance

Our primary interest is in estimable functions of  $\tau$ , given by

$$\xi' \tau$$

where

$$\xi' = \rho'(I - M)Z .$$

Thus,

$$\xi' \hat{\tau} = \rho'(I - M)Z(Z'(I - M)Z)^{-} Z'(I - M)Y .$$

We note here that the term  $(I - M)Z(Z'(I - M)Z)^{-}$  can be simplified. Since the columns of  $Z$  are 0's and 1's indicating the presence or absence of treatment effects, we have

$$ZJ_t = J_n ,$$

where  $n = bk$ . Also we have

$$0 = (I - M)J_n = (I - M)ZJ_t . \quad (11)$$

The equation (11), along with equations (9) and (10) imply

$$(I - M)Z(Z'(I - M)Z)^{-} = \frac{k}{\lambda t}(I - M)Z.$$

# Analysis of Covariance

Now we can write

$$\begin{aligned}\xi' \hat{\tau} &= \rho'(I - M)Z(Z'(I - M)Z)^{-}Z'(I - M)Y \\ &= \rho' \frac{k}{\lambda t} \xi'(Q_1, \dots, Q_t)' \\ &= \frac{k}{\lambda t} \sum_{j=1}^t \xi_j Q_j.\end{aligned}$$

The variance of this estimate is given by

$$\begin{aligned}\text{Var}(\xi' \hat{\tau}) &= \sigma^2 \rho'(I - M)Z(Z'(I - M)Z)^{-}Z'(I - M)\rho \\ &= \sigma^2 \frac{k}{\lambda t} \xi' \xi.\end{aligned}$$

The estimate of  $\sigma^2$  is

$$MSE = \frac{SSE}{bk - t - b + 1}$$

where

$$\begin{aligned}SSE &= Y'(I - M)Y - Y'(I - M)Z(Z'(I - M)Z)^{-}Z'(I - M)Y \\ &= Y'(I - M)Y - \frac{k}{\lambda t} Y'(I - M)ZZ'(I - M)Y \\ &= \sum_{ij} (Y_{ij} - \bar{Y}_{i.})^2 - \frac{k}{\lambda t} \sum_{j=1}^t Q_j^2\end{aligned}$$

# Analysis of Covariance

Thus if we wish to test the hypothesis

$$H_0 : \xi' \tau = 0 ,$$

the  $F$  test is given by

$$\begin{aligned} F &= \frac{(\xi' \hat{\tau})^2}{MSE \frac{k}{\lambda_t} \xi' \xi} \\ &\sim F(1, bk - t - b + 1, \gamma) \end{aligned}$$

where

$$\gamma = \frac{(\xi' \tau)^2}{2\sigma^2 \xi' \xi \frac{k}{\lambda_t}} .$$

Let  $n = bk = rt$ , and let

$$B_i = \sum_{j \in D_i} Y_{ij}$$

denote the  $i$ th block total,  $i = 1, \dots, b$ . Let

$$G = \sum_{ij} Y_{ij}$$

denote the grand total.

# Analysis of Covariance

Then, the ANOVA table for the BIB design can be written as

Source	df	SS
Mean	1	$\frac{1}{n} \mathbf{Y}' \mathbf{J}_n^n \mathbf{Y}$
blocks (ignoring treatments)	$b - 1$	$\frac{1}{k} \sum_{i=1}^b B_i^2 - G^2/n$
treatments (adjusted for blocks)	$t - 1$	$\frac{k}{\lambda t} \sum_{j=1}^t Q_j^2$
Error	$n - t - b + 1$	SSE
Total	$n$	$\mathbf{Y}' \mathbf{Y}$

## Remarks about BIB's

1) For a BIB, we must have

- i)  $n = bk = rt$
- ii)  $\lambda(t - 1) = r(k - 1)$

Conditions i) and ii) imply that  $b \geq t$ . This is called Fisher's inequality.  
Thus for a BIB we need  $b \geq t$ .

2) The conditions in 1) are necessary but not sufficient for a BIB to exist.  
Thus, a BIB need not exist for some sets of parameters that satisfy 1).

# Analysis of Covariance

## Incomplete Block Designs

For the general incomplete block design, any treatment arrangement is permissible. For example one could have

block 1	block 2	block 3
A	C	E
B	D	
A		

We can spend a whole course discussing incomplete block designs. We do not have the time to go into further detail. A good reference for incomplete block designs is Statistical Design and Analysis of Experiments, by Peter W. M. John (1971, Chapter 11).

# Regression Analysis

## Regression Analysis

A regression model is a linear model

$$Y = X\beta + \epsilon ,$$

where  $X_{n \times p}$  and is usually of full rank  $p$ . The columns of  $X$  typically consist of continuous and categorical variables. If the columns of  $X$  consist of ALL categorical variables, then the model is NOT a regression model in the usual sense. Thus, a regression model needs at least one continuous variable. The variables that make up the columns of  $X$  are typically called covariates, predictors, regressors, explanatory variables, or independent variables. We will assume throughout our discussion of regression models that  $X$  has full rank  $p$  so that  $X'X$  is nonsingular.

When  $X$  has full rank  $p$ ,  $\beta$  is estimable, and every linear function of  $\beta$  is estimable. To see that  $\beta$  is estimable, note that we can write

$$\beta = P'X\beta$$

where

$$P' = (X'X)^{-1}X' .$$

# Regression Analysis

## Simple Linear Regression

The simple linear regression model (including an intercept) can be written as

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i , \quad i = 1, \dots, n ,$$

where the  $\epsilon_i$ 's are *i.i.d.* with  $E(\epsilon_i) = 0$  and  $\text{Var}(\epsilon_i) = \sigma^2$ , for  $i = 1, \dots, n$ . We can write the simple linear regression model in matrix form as

$$Y = X\beta + \epsilon ,$$

where

$$X = \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{pmatrix}_{n \times 2} .$$

We see here that  $r(X) = 2$  if the  $X_i$ 's are not all the same. We know from previous results that the BLUE of  $\beta$  is

$$\begin{aligned}\hat{\beta} &= \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} \\ &= (X'X)^{-1}X'Y .\end{aligned}$$

# Regression Analysis

We have

$$X'X = \begin{pmatrix} n & \sum_{i=1}^n X_i^2 \\ n\bar{X} & \end{pmatrix}$$

where  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ . It is easily seen that

$$(X'X)^{-1} = \frac{1}{n \sum_{i=1}^n (X_i - \bar{X})^2} \begin{pmatrix} \sum_{i=1}^n X_i^2 & -n\bar{X} \\ -n\bar{X} & n \end{pmatrix}$$

Note also that

$$X'Y = \begin{pmatrix} n\bar{Y} \\ \sum_{i=1}^n X_i Y_i \end{pmatrix}.$$

Thus

$$\begin{aligned}\hat{\beta} &= \frac{1}{n \sum_{i=1}^n (X_i - \bar{X})^2} \begin{pmatrix} \sum_{i=1}^n X_i^2 & -n\bar{X} \\ -n\bar{X} & n \end{pmatrix} \\ &\quad \times \begin{pmatrix} n\bar{Y} \\ \sum_{i=1}^n X_i Y_i \end{pmatrix} \\ &= \frac{1}{n \sum_{i=1}^n (X_i - \bar{X})^2} \begin{pmatrix} n\bar{Y} \sum_{i=1}^n X_i^2 - n\bar{X} \sum_{i=1}^n X_i Y_i \\ -n^2 \bar{X} \bar{Y} + n \sum_{i=1}^n X_i Y_i \end{pmatrix} \\ &= \begin{pmatrix} \bar{Y} - \bar{X} \left( \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} \right) \\ \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} \end{pmatrix}\end{aligned}$$

# Regression Analysis

Thus

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} \\ &= \frac{\hat{\text{Cov}}(X, Y)}{\hat{\text{Var}}(X)}\end{aligned}$$

where  $\hat{\text{Cov}}(X, Y)$  denotes the sample covariance of  $(X, Y)$  and  $\hat{\text{Var}}(X)$  denotes the sample variance of  $X$ . We also have

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}.$$

By doing the matrix multiplication, it is easily shown that the orthogonal projection operator onto  $C(X)$  has  $ij$ th element given by

$$m_{ij} = \frac{1}{n} + \frac{(X_i - \bar{X})(X_j - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2}.$$

# Regression Analysis

Also,

$$\begin{aligned}\text{Cov}(\hat{\beta}) &= \text{Cov}((X'X)^{-1}X'Y) \\ &= \left( (X'X)^{-1}X' \right) (\text{Cov}(Y)) \left( (X'X)^{-1}X' \right)' \\ &= (X'X)^{-1}X'(\sigma^2 I)(X(X'X)^{-1}) \\ &= \sigma^2 (X'X)^{-1}(X'X)(X'X)^{-1} \\ &= \sigma^2 (X'X)^{-1}.\end{aligned}$$

Confidence intervals and tests of hypotheses are carried out by assuming  $\epsilon \sim N_n(0, \sigma^2 I)$ .

# Regression Analysis

## Multiple Linear Regression

When there is more than one covariate, we call the regression model a multiple linear regression model. In scalar form, the multiple linear regression model with an intercept can be written as

$$Y_i = \beta_0 + \beta_1 X_{i1} + \dots + \beta_p X_{ip} + \epsilon_i ,$$

where the  $\epsilon_i$ 's are *i.i.d.* with  $E(\epsilon_i) = 0$  and  $\text{Var}(\epsilon_i) = \sigma^2$ . More generally, we can write the multiple linear regression model as

$$Y_i = \beta_1 X_{i1} + \dots + \beta_p X_{ip} + \epsilon_i .$$

In matrix form, we can write the multiple regression model as

$$Y = X\beta + \epsilon$$

where  $X_{n \times p}$ , and has the form

$$\begin{pmatrix} X_{11} & \dots & X_{1p} \\ \vdots & \dots & \vdots \\ X_{n1} & \dots & X_{np} \end{pmatrix} .$$

Thus, the BLUE of  $\beta$  is

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

and

$$\text{Cov}(\hat{\beta}) = \sigma^2(X'X)^{-1} .$$

# Regression Analysis

In multiple linear regression, we can decompose the total sums of squares as

$$\begin{aligned} Y'Y &= Y'MY + Y'(I - M)Y \\ &= SSR(X) + SSE \end{aligned}$$

where  $SSR(X) = Y'MY$  is called the regression sums of squares. The column space that makes up the regression sums of squares is  $C(X) = C(M)$ . Thus

$$\begin{aligned} R^n &= C(X) + C(X)^\perp \\ &= C(M) + C(I - M). \end{aligned}$$

Since  $r(M) = p$ ,  $C(M)$  can be decomposed into a sum of  $p$  orthogonal subspaces, each of dimension 1. Thus

$$C(M) = C(M_1) + \dots + C(M_p)$$

where  $r(M_i) = 1$ ,  $M_i M_j = 0$ , and  $M_i$  is an orthogonal projection operator. Let  $X_i$  denote the  $i$ th column of  $X$ ,  $i = 1, \dots, p$ , so that  $X = (X_1, \dots, X_p)$ . Define

$$Z_i = (X_1, \dots, X_i), \quad i = 1, \dots, p$$

and define  $Z_0 = 0$ . Let

$$P_{Z_i} = Z_i(Z_i' Z_i)^{-1} Z_i'$$

denote the orthogonal projection operator onto  $C(Z_i)$ ,  $i = 0, \dots, p$ . Now let

$$M_i = P_{Z_i} - P_{Z_{i-1}}, \quad i = 1, \dots, p.$$

# Regression Analysis

It is easily seen that each  $M_i$  is an orthogonal projection operator,  $r(M_i) = 1$ , and  $M_i M_j = 0$  for  $i \neq j$ . Thus

$$\begin{aligned}M &= M_1 + \dots + M_p \\&= P_{Z_1} + (P_{Z_2} - P_{Z_1}) + (P_{Z_3} - P_{Z_2}) \\&\quad + \dots + (P_{Z_p} - P_{Z_{p-1}}) \\&= P_{Z_p} \\&= M.\end{aligned}$$

Thus,

$$Y' M Y = Y' M_1 Y + \dots + Y' M_p Y.$$

Now denote

$$SSR(X_j | X_1, \dots, X_{j-1}) = Y' M_j Y, \quad j = 1, \dots, p.$$

Thus  $SSR(X_j | X_1, \dots, X_{j-1})$  can be interpreted as the sum of squares due to adding  $X_j$  given that  $X_1, \dots, X_{j-1}$  are already in the model. Thus, we can write

$$\begin{aligned}SSR(X) &= SSR(X_1) + SSR(X_2 | X_1) \\&\quad + SSR(X_3 | X_1, X_2) + \dots + SSR(X_p | X_1, \dots, X_{p-1}).\end{aligned}$$

# Regression Analysis

This orthogonal decomposition of  $C(X)$  depends on the order of the variables being fit into the model. Another ordering of the variables gives a different orthogonal breakdown. Thus the orthogonal decomposition of  $C(X)$  is not unique. This is not surprising since we already encountered this when we discussed unbalanced ANOVA. The orthogonal decomposition of  $C(X)$  however is unique if  $C(X_i) \perp C(X_j)$  for every  $i \neq j$ .

We can write the ANOVA table for the multiple linear regression model as

source	df	SS	MS
Regression	$p$	$Y' MY$	$Y' MY / p$
Error	$n - p$	$Y' (I - M) Y$	$\frac{Y' (I - M) Y}{n - p}$
Total	$n$	$Y' Y$	

# Regression Analysis

Breaking  $Y'Y$  into a sum of  $p$  independent quadratic forms yields

source	df	SS	MS
$X_1$	1	$Y'M_1Y$	$Y'M_1Y$
$X_2   X_1$	1	$Y'M_2Y$	$Y'M_2Y$
$X_3   X_1, X_2$	1	$Y'M_3Y$	$Y'M_3Y$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$X_p   X_1, \dots, X_{p-1}$	1	$Y'M_pY$	$Y'M_pY$
Error	$n - p$	$Y'(I - M)Y$	$\frac{Y'(I - M)Y}{n-p}$
Total	$n$	$Y'Y$	

If an intercept is included in the model, then we can let  $J_n = X_1$ , and  $M_1 = \frac{1}{n}J_nJ'_n$ .

To carry out tests of hypotheses and construct confidence regions, we assume  $\epsilon \sim N_n(0, \sigma^2 I)$ , and carry out the procedures as we did for the general linear model.

# Regression Analysis

## Best Linear Prediction

One of the main goals and uses of regression models is for prediction. We often want to construct models so that we can use the model to predict future observations. One can argue that the main goal in any statistical analysis is to predict and make inferences about observable quantities, and that parameters should not be the main focus in inference. That is, parameters should be viewed as a tool for making predictions and developing models, but they are not an end in themselves. There is a lot of philosophical discussion on estimation versus prediction in many text books and articles.

We can view the regression problem as a prediction problem. That is, regression can be considered as the problem of predicting  $Y$  on the basis of  $X_1, \dots, X_p$ . Let  $X = (X_1, \dots, X_p)'$  be a  $p \times 1$  vector, and  $Y$  is a scalar. Further, we assume  $X$  and  $Y$  are random. A reasonable criterion for choosing a predictor for  $Y$  is to pick a predictor  $f(X)$  that minimizes the mean square error. That is we, pick  $f(X)$  to minimize

$$E(Y - f(X))^2$$

where the expectation is taken with respect to the joint distribution  $(X, Y)$ . We are now led to the following theorem.

# Regression Analysis

## Theorem

Let  $m(X) = E(Y | X)$ . Then for any other predictor  $f(X)$ ,

$$E(Y - m(X))^2 \leq E(Y - f(X))^2,$$

where the expectation is taken with respect to the joint distribution of  $(X, Y)$ , and thus  $m(X) = E(Y | X)$  is the best predictor of  $Y$ ,

This theorem says that the best predictor of  $Y$  is the conditional expectation of  $Y$  given  $X$ .

## Proof

$$\begin{aligned} E(Y - f(X))^2 &= E(Y - m(X) + m(X) - f(X))^2 \\ &= E(Y - m(X))^2 + E(m(X) - f(X))^2 \\ &\quad + 2E((Y - m(X))(m(X) - f(X))) \end{aligned}$$

Since both  $E(Y - m(X))^2$  and  $E(m(X) - f(X))^2$  are nonnegative, it is enough to show that

$$E[(Y - m(X))(m(X) - f(X))] = 0.$$

# Regression Analysis

To this end, we note that we can write  $E_{(x,y)} = E_x E_{y|x}$  where  $E_{(x,y)}$  denotes the expectation with respect to the joint distribution of  $(X, Y)$ ,  $E_{y|x}$  denotes the expectation with respect to the conditional distribution of  $Y$  given  $X$ , and  $E_x$  denotes the expectation with respect to the marginal distribution of  $X$ . Thus

$$\begin{aligned} & E_{(x,y)} [(Y - m(X))(m(X) - f(X))] \\ &= E_x [E_{y|x} [(Y - m(X))(m(X) - f(X))]] \\ &= E_x [(m(X) - f(X))E_{y|x}(Y - m(X))] \\ &= E_x [(m(X) - f(X))(E(Y | X) - m(X))] \\ &= E_x [(m(X) - f(X)) 0] \\ &= 0 . \end{aligned}$$

# Regression Analysis

The last equation on the previous slide follows since  $m(X) = E(Y | X)$ . In order to use this general result, one needs to specify a joint distribution on  $(X, Y)$  so that  $E(Y | X)$  can be computed. For example, if

$$\begin{pmatrix} Y \\ X \end{pmatrix} \sim N_{p+1}(\mu, \Sigma),$$

then

$$E(Y | X) = (\mu_y - \Sigma_{12}\Sigma_{22}^{-1}\mu_x) + \Sigma_{12}\Sigma_{22}^{-1}X$$

where

$$\mu = \begin{pmatrix} \mu_y \\ \mu_x \end{pmatrix}$$

and

$$\Sigma = \begin{pmatrix} \sigma_y^2 & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}.$$

Here,  $\mu_y = E(Y)$ ,  $\Sigma_{12} = \text{Cov}(Y, X)$  and  $\Sigma_{22} = \text{Cov}(X)$ .

We note here that if we assume that the joint distribution between  $(X, Y)$  is multivariate normal, then the best predictor is a linear predictor.

# Regression Analysis

Now suppose the joint distribution of  $(X, Y)$  is not known, but means, variances, and covariances of  $(X, Y)$  (or their estimates) are all known. In this case, we can find the best linear predictor of  $Y$ . We seek a linear predictor of the form

$$\alpha + X'\beta$$

that minimizes

$$E(Y - \alpha - X'\beta)^2$$

for all scalars  $\alpha$  and  $p \times 1$  vectors  $\beta$ . Let  $E(Y) = \mu_y$ ,  $E(X) = \mu_x$ ,  $\text{Cov}(X) = V_{xx}$ , and

$$\text{Cov}(X, Y) = V_{xy} = E[(X - \mu_x)(Y - \mu_y)'] .$$

Also, denote  $V_{yx} = V'_{xy}$ . We are now led to the following theorem.

## Theorem

Let  $\beta_*$  be a solution to

$$V_{xx}\beta = V_{xy} .$$

Then

$$\mu_y + (X - \mu_x)'\beta_*$$

is the best linear predictor of  $Y$ .

# Regression Analysis

## Proof

Without loss of generality, we can write an arbitrary linear predictor as

$$\alpha + (X - \mu_x)' \beta .$$

Let us first consider the optimal choice of  $\alpha$ . We have

$$\begin{aligned} & E(Y - \alpha - (X - \mu_x)' \beta)^2 \\ = & E[Y - \mu_y - (X - \mu_x)' \beta + (\mu_y - \alpha)]^2 \\ = & E[Y - \mu_y - (X - \mu_x)' \beta]^2 + (\mu_y - \alpha)^2 \\ + & 2E[(\mu_y - \alpha)(Y - \mu_y - (X - \mu_x)' \beta)] \end{aligned}$$

But notice that

$$\begin{aligned} & E[(\mu_y - \alpha)(Y - \mu_y - (X - \mu_x)' \beta)] \\ = & (\mu_y - \alpha)E(Y - \mu_y - (X - \mu_x)' \beta) \\ = & (\mu_y - \alpha)(\mu_y - \mu_y - (\mu_x - \mu_x)' \beta) \\ = & (\mu_y - \alpha)0 \\ = & 0 . \end{aligned}$$

# Regression Analysis

Therefore,

$$\begin{aligned} & E [Y - \alpha - (X - \mu_x)' \beta]^2 \\ = & E [Y - \mu_y - (X - \mu_x)' \beta]^2 + (\mu_y - \alpha)^2 \\ \geq & E [Y - \mu_y - (X - \mu_x)' \beta]^2 . \end{aligned}$$

Thus  $\alpha = \mu_y$  is the optimal choice for  $\alpha$ .

It remains to show that the optimal choice of  $\beta$  is a solution to

$$V_{xx}\beta = V_{xy} .$$

Without loss of generality, we can assume  $\mu_x = \mu_y = 0$ , since we can subtract them off if they were not equal to zero. Let

$$V_{xx}\beta_* = V_{xy} .$$

Then

$$\begin{aligned} E(Y - X'\beta)^2 &= E(Y - X'\beta_* + X'\beta_* - X'\beta)^2 \\ &= E(Y - X'\beta_*)^2 + E(X'\beta_* - X'\beta)^2 \\ &\quad + 2 E [(Y - X'\beta_*)(X'\beta_* - X'\beta)] . \end{aligned}$$

# Regression Analysis

We want to show that the last term in the equation above is zero. Thus

$$\begin{aligned} & E[(Y - X'\beta_*)(X'\beta_* - X'\beta)] \\ &= E[YX'(\beta_* - \beta)] - E[\beta_*'XX'(\beta_* - \beta)] \\ &= V_{yx}(\beta_* - \beta) - \beta_*'V_{xx}(\beta_* - \beta) \\ &= V_{yx}(\beta_* - \beta) - V_{yx}(\beta_* - \beta) \\ &= 0. \end{aligned}$$

Note that we have made use of the assumption that  $V_{xx}\beta_* = V_{xy}$ . Thus we have,

$$\begin{aligned} E(Y - X'\beta)^2 &= E(Y - X'\beta_*)^2 + E(X'\beta_* - X'\beta)^2 \\ &\geq E(Y - X'\beta_*)^2, \end{aligned}$$

and therefore an optimal choice of  $\beta$  is  $\beta_*$  which satisfies

$$V_{xx}\beta_* = V_{xy}.$$

If  $V_{xx}$  is nonsingular, then  $\beta_*$  is unique so that the unique optimal choice of  $\beta$  is

$$\beta_* = V_{xx}^{-1}V_{xy}.$$

Thus we have shown that the function

$$\mu_y + (X - \mu_x)'\beta_*$$

is the best linear predictor of  $Y$  based on  $X$ .

# Regression Analysis

If  $(X, Y)$  have a multivariate normal distribution, the best predictor is linear, so that it is also the best linear predictor. Thus, if  $(X, Y)$  are multivariate normal, then the best predictor = best linear predictor = conditional expectation of  $Y | X$ .

Now we apply these results to the multiple linear regression model. Let  $X_{n \times p} = (X_1, \dots, X_p)$  be the matrix of covariates based on  $n$  observations, where each  $X_i$  is an  $n \times 1$  vector. Also let  $Y$  be the  $n \times 1$  vector of responses. Then

$$S_{xx} = \frac{X'(I - \frac{1}{n}J_n^n)X}{n-1}$$

and

$$S_{xy} = \frac{X'(I - \frac{1}{n}J_n^n)Y}{n-1}.$$

# Regression Analysis

Thus  $S_{xx}$  is the sample covariance matrix of the covariates, and  $S_{xy}$  is the vector of sample covariances between  $(X, Y)$ . Estimates of  $\mu_x$  and  $\mu_y$  are given by

$$\begin{aligned}\hat{\mu}_x &= \bar{X} \\ &= \frac{1}{n} J'_n X \\ &= \begin{pmatrix} \bar{X}_1 \\ \vdots \\ \bar{X}_p \end{pmatrix} \\ &= \left( \frac{1}{n} \sum_{i=1}^n X_{i1}, \dots, \frac{1}{n} \sum_{i=1}^n X_{ip} \right)'.\end{aligned}$$

Similarly,

$$\hat{\mu}_y = \bar{Y} = \frac{1}{n} J'_n Y = \frac{1}{n} \sum_{i=1}^n Y_i.$$

Thus, for multiple linear regression, the best linear predictor of  $Y_i$  is

$$\hat{Y}_i = \bar{Y} + (X_i - \bar{X})' \hat{\beta}_*$$

where  $\hat{\beta}_*$  is a solution to

$$X'(I - n^{-1} J_n^n) X \hat{\beta}_* = X'(I - n^{-1} J_n^n) Y$$

# Regression Analysis

If  $X'(I - n^{-1}J_n^n)X$  is nonsingular, then

$$\hat{\beta}_* = (X'(I - n^{-1}J_n^n)X)^{-1}X'(I - n^{-1}J_n^n)Y \quad (12)$$

We note that  $\hat{\beta}_*$  in (12) is the least squares estimate from the model

$$Y_i = \alpha + (X_i - \bar{X})'\beta + \epsilon_i \quad (13)$$

where  $E(\epsilon_i) = 0$  and  $\text{Var}(\epsilon_i) = \sigma^2$ .

To see this, note that we can write model (13) in matrix form as

$$\begin{aligned} Y &= \left( J_n, (I - n^{-1}J_n^n)X \right) \begin{pmatrix} \alpha \\ \beta \end{pmatrix} + \epsilon \\ &= J_n\alpha + (I - n^{-1}J_n^n)X\beta + \epsilon. \end{aligned}$$

We see that  $J_n$  and  $(I - n^{-1}J_n^n)X$  are orthogonal, since

$$\begin{aligned} J_n'(I - n^{-1}J_n^n)X &= J'X - n^{-1}J_n'J_n^nX \\ &= J'_nX - J'_nX \\ &= 0. \end{aligned}$$

# Regression Analysis

Since these columns are orthogonal, we can estimate  $\alpha$  and  $\beta$  separately and thus the least squares estimate of  $\beta$  is

$$\hat{\beta}_* = (X'(I - n^{-1}J_n^n)X)^{-1}X'(I - n^{-1}J_n^n)Y,$$

and the least squares estimate of  $\alpha$  is

$$\hat{\alpha} = (J_n'J_n)^{-1}J_n'Y = \frac{1}{n}J_n'Y = \bar{Y}.$$

# Regression Analysis

## The Coefficient of Determination

Consider the linear model

$$Y = J_n \beta_0 + X\beta + \epsilon \quad (1)$$

where  $X_{n \times p}$  is of rank  $p$  and  $J_n$  is the  $n \times 1$  vector of ones. Thus, we consider the linear regression model with an intercept. The intercept term is written separately from  $X$  for ease of exposition and notation. Thus we assume that  $X$  does not contain a column of ones. An alternative way to write the model in (1) is

$$Y = J_n \delta_0 + (I - n^{-1} J_n^n)X\delta + \epsilon. \quad (2)$$

Models (1) and (2) are equivalent because  $C(J_n, X) = C(J_n, (I - n^{-1} J_n^n)X)$ . The model in (2) is correcting all of the variables for their means. The parameters in the two models are related by  $\beta_0 = \delta_0 - n^{-1} J_n' X \delta$  and  $\beta = \delta$ . Also, we see that since  $C((I - n^{-1} J_n^n)X)$  is the orthogonal complement of  $C(J_n)$  with respect to  $C(J_n, X)$ , we have

$$\begin{aligned} & M^* - n^{-1} J_n^n \\ &= (I - n^{-1} J_n^n)X(X'(I - n^{-1} J_n^n)X)^{-1}X'(I - n^{-1} J_n^n) \end{aligned}$$

where  $M^*$  is the orthogonal projection operator onto  $C(J_n, X)$ .

# Regression Analysis

The Coefficient of Determination, written  $R^2$ , is defined from model (1) as

$$R^2 = \frac{SS_{reg}}{SSTOT - C} ,$$

where  $SSTOT = Y'Y$ ,  $C = \frac{1}{n}(J'_n Y)^2 = n\bar{Y}^2$ , and  $\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$ . Here,  $SS_{reg}$  is the regression sum of squares from model (1). Thus

$$SS_{reg} = Y'(M^* - n^{-1}J_n^n)Y ,$$

where  $M^*$  is the orthogonal projection operator onto  $C(J_n, X)$ . Also we have

$$\begin{aligned} SYY &= SSTOT - C \\ &= Y'(I - n^{-1}J_n^n)Y \\ &= \sum_{i=1}^n (Y_i - \bar{Y})^2 . \end{aligned}$$

Thus we can write

$$\begin{aligned} SS_{reg} &= Y'(M^* - n^{-1}J_n^n)Y \\ &= Y'M^*Y - n^{-1}Y'J_n^nY \\ &= Y'(I - n^{-1}J_n^n)Y - Y'(I - M^*)Y \\ &= SYY - SSE . \end{aligned}$$

# Regression Analysis

Therefore,

$$\begin{aligned} R^2 &= \frac{SS_{reg}}{SSTOT - C} \\ &= \frac{SYY - SSE}{SYY} \\ &= 1 - \frac{SSE}{SYY}. \end{aligned}$$

We see that  $0 \leq R^2 \leq 1$ .  $R^2$  is interpreted as the proportion of the total variability in  $Y$  explained by the independent variables  $(X_1, \dots, X_p)$ . The greater the proportion of the total variability of the data that is explained by the model, the better the fit of the model.

To motivate the use of  $R^2$  for assessing fit, we show that  $R^2$  is the natural estimate of the square of the multiple correlation coefficient. The correlation between two variables  $(X_1, X_2)$  is defined as

$$\text{Corr}(X_1, X_2) = \frac{\text{Cov}(X_1, X_2)}{[\text{Var}(X_1)\text{Var}(X_2)]^{1/2}}.$$

# Regression Analysis

The multiple correlation coefficient between a variable  $Y$  and a set of variables  $(X_1, \dots, X_p)$  is defined as the maximum correlation between  $Y$  and a linear function of  $X = (X_1, \dots, X_p)'$ . Thus, it is

$$\max_{(\alpha, \beta)} \{ \text{Corr}(Y, \alpha + X'\beta) \}$$

where  $\beta = (\beta_1, \dots, \beta_p)'$ . We note here that

$$\text{Corr}(Y, \alpha + X'\beta) = \frac{\text{Cov}(Y, X'\beta)}{[\text{Var}(Y)\text{Var}(X'\beta)]^{1/2}} .$$

The  $\alpha$  drops out since it is a constant. Defining  $\text{Cov}(Y, X) = V_{yx}$ , we have

$$\text{Cov}(Y, X'\beta) = V_{yx}\beta = \beta_*' V_{xx} \beta$$

where  $V_{xx}\beta_* = V_{xy}$  as defined earlier. Also, from the definition of variance, we have

$$\text{Var}(X'\beta) = \beta' V_{xx} \beta .$$

Using the above equations, we have

$$\begin{aligned} \text{Cov}(Y, \alpha + X'\beta_*) &= \beta_*' V_{xx} \beta_* \\ &= \text{Var}(\alpha + X'\beta_*) . \end{aligned}$$

# Regression Analysis

To establish an upper bound on the correlation between  $Y$  and  $\alpha + X'\beta$ , we need the Cauchy-Schwarz inequality. The Cauchy-Schwarz inequality states that

$$\left( \sum_{i=1}^t r_i s_i \right)^2 \leq \left( \sum_{i=1}^t r_i^2 \right) \left( \sum_{i=1}^t s_i^2 \right).$$

Since  $V_{xx}$  can be written as  $V_{xx} = RR'$  for some matrix  $R$ , we have

$$\begin{aligned} (\beta_*' V_{xx} \beta)^2 &= ((\beta_*' R)(R' \beta))^2 \\ &= \left( \sum_{i=1}^p r_i s_i \right)^2 \end{aligned}$$

where  $r_i$  is the  $i$ th component of the vector  $\beta_*' R$  and  $s_i$  is the  $i$ th component of the vector  $R' \beta$ . Thus, by Cauchy-Schwarz, we have

$$(\beta_*' RR' \beta)^2 \leq (\beta' R)^2 (\beta_*' R)^2,$$

and thus

$$(\beta_*' V_{xx} \beta)^2 \leq (\beta' V_{xx} \beta) (\beta_*' V_{xx} \beta_*).$$

# Regression Analysis

Let  $\sigma_y^2 = \text{Var}(Y)$ . Thus,

$$\begin{aligned} (\text{Corr}(Y, \alpha + X'\beta))^2 &= \frac{(\beta_*' V_{xx} \beta)^2}{(\beta' V_{xx} \beta) \sigma_y^2} \\ &\leq \frac{\beta_*' V_{xx} \beta_*}{\sigma_y^2} \\ &= \frac{(\beta_*' V_{xx} \beta_*)^2}{(\beta_*' V_{xx} \beta_*) \sigma_y^2} \\ &= (\text{Corr}(Y, \alpha + X'\beta_*))^2 . \end{aligned}$$

This establishes the result. From the above result, we have

$$(\text{Corr}(Y, \alpha + X'\beta_*))^2 = \frac{\beta_*' V_{xx} \beta_*}{\sigma_y^2} .$$

Now if we have observations  $(X_i, Y_i)$ ,  $i = 1, \dots, n$ , where  $X'_i = (X_{i1}, \dots, X_{ip})'$ , then the estimate of  $\sigma_y^2$  is

$$S_y^2 = \frac{SYY}{n - 1} .$$

# Regression Analysis

The natural estimate of the squared multiple correlation between  $Y$  and  $X$  is thus

$$\begin{aligned}\frac{\hat{\beta}'_* \hat{V}_{xx} \hat{\beta}_*}{S_y^2} &= \frac{\hat{\beta}'_* S_{xx} \hat{\beta}_*}{S_y^2} \\ &= \frac{\hat{\beta}'_* X'(I - n^{-1} J_n^n) X \hat{\beta}_*}{Y'(I - n^{-1} J_n^n) Y}\end{aligned}$$

where

$$\hat{\beta}_* = (X'(I - n^{-1} J_n^n) X)^{-1} X'(I - n^{-1} J_n^n) Y.$$

Let  $M^*$  denote the orthogonal projection operator onto  $C(J_n, X)$ . We note that since  $C((I - n^{-1} J_n^n) X)$  is the orthogonal complement of  $C(J_n)$  with respect to  $C(J_n, X)$ , we have

$$\begin{aligned}M^* - n^{-1} J_n^n \\ = (I - n^{-1} J_n^n) X (X'(I - n^{-1} J_n^n) X)^{-1} X'(I - n^{-1} J_n^n),\end{aligned}$$

so that

$$\hat{\beta}'_* \hat{V}_{xx} \hat{\beta}_* = Y'(M^* - n^{-1} J_n^n) Y.$$

Thus

$$\frac{\hat{\beta}'_* \hat{V}_{xx} \hat{\beta}_*}{S_y^2} = \frac{Y'(M^* - n^{-1} J_n^n) Y}{Y'(I - n^{-1} J_n^n) Y}.$$

# Regression Analysis

Thus,  $R^2$  is the estimate of the maximum squared correlation between  $Y$  and  $\alpha + X'\beta$ .

Since

$$R^2 = 1 - \frac{SSE}{SYY} ,$$

we have

$$\begin{aligned}\frac{SSE}{SYY} &= 1 - R^2 \\ \Leftrightarrow \frac{SSE}{SSE + SS_{reg}} &= 1 - R^2 \\ \Leftrightarrow SSE &= (1 - R^2)(SSE + SS_{reg}) \\ \Leftrightarrow (SSE)R^2 &= (1 - R^2)SS_{reg} \\ \Leftrightarrow \frac{SS_{reg}}{SSE} &= \frac{R^2}{1 - R^2} .\end{aligned}$$

# Regression Analysis

The  $F$  test for the hypothesis

$$H_0 : \beta_1 = \dots = \beta_p = 0$$

is sometimes called the overall  $F$  test for regression and is given by

$$\begin{aligned} F &= \frac{Y'(M^* - n^{-1}J_n^n)Y/p}{Y'(I - M^*)Y/(n - p - 1)} \\ &= \frac{SS_{reg}/p}{SSE/(n - p - 1)} \\ &= \left(\frac{n - p - 1}{p}\right)\left(\frac{R^2}{1 - R^2}\right). \end{aligned}$$

This  $F$  statistic above has an  $F(p, n - p - 1)$  distribution under  $H_0$ .

## Example

Consider Example 6.21. given on page 114 of Christensen. The model is given by

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon.$$

We have  $SS_{reg} = 1259.32$ ,  $SSTOT - C = 16657.6 - 15270.78 = 1305.82$ .

Therefore  $SSE = 1305.82 - 1259.32 = 46.50$ . Thus

$$R^2 = \frac{1259.32}{1305.82} = .964.$$

# Regression Analysis

## Partial Correlation Coefficients

In regression analysis, we are often interested in the correlation coefficient between two variables given (conditional) on a set of variables already in the model. This quantity is often of interest in variable selection. Suppose we are interested in the correlation between  $(Y_1, Y_2)$  given  $X = (X_1, \dots, X_p)'$  are in the model. Thus, we are interested in computing the conditional correlation coefficient

$$\rho_{y,x} = \text{Corr}(Y_1, Y_2 | X).$$

$\rho_{y,x}$  can be interpreted as a measure of the linear relationship between  $Y_1$  and  $Y_2$  after taking the effects of  $X = (X_1, \dots, X_p)'$  out of both variables. Let  $Y = (Y_1, Y_2)'$  be a  $2 \times 1$  vector. It can be shown that the conditional  $2 \times 2$  covariance matrix of  $Y | X$  is

$$\text{Cov}(Y | X) = \text{Cov}(Y - \mu_y - (X - \mu_x)' \beta_*)$$

where  $\beta_*$  is a solution to  $V_{xx}\beta_* = V_{xy}$  and  $V_{xy} = \text{Cov}(X, Y)$  is a  $p \times 2$  matrix. Notice that the above equation implies that the conditional covariance matrix can be obtained by computing the unconditional covariance of  $Y$  minus its best linear predictor. That is, we take the effects out of  $Y$  by looking at

$$Y - (\mu_y + (X - \mu_x)' \beta_*) . \quad (3)$$

# Regression Analysis

The partial correlation coefficient is now defined as the correlation between the two components of the vector in (3). To derive the covariance matrix of (3), note that

$$\begin{aligned} & \text{Cov}(Y - \mu_y - (X - \mu_x)'\beta_*) \\ = & \text{Cov}(Y - \mu_y) + \beta_*' \text{Cov}(X - \mu_x)\beta_* \\ - & \text{Cov}(Y - \mu_y, (X - \mu_x)')\beta_* - \beta_*' \text{Cov}(X - \mu_x, Y - \mu_y) \\ = & V_{yy} + \beta_*' V_{xx} \beta_* - V_{yx} \beta_* - \beta_*' V_{xy} \\ = & V_{yy} + \beta_*' V_{xx} \beta_* - \beta_*' V_{xx} \beta_* - \beta_*' V_{xx} \beta_* \\ = & V_{yy} - \beta_*' V_{xx} \beta_* \end{aligned}$$

Thus, for any generalized inverse of  $V_{xx}$ , i.e.,  $V_{xx} V_{xx}^- V_{xx} = V_{xx}$ , we have

$$\begin{aligned} \text{Cov}(Y - \mu_y - (X - \mu_x)'\beta_*) &= V_{yy} - \beta_*' V_{xx} V_{xx}^- V_{xx} \beta_* \\ &= V_{yy} - V_{yx} V_{xx}^- V_{xy} . \end{aligned}$$

# Regression Analysis

Suppose we have a sample

$$Y = (Y_1, Y_2) = \begin{pmatrix} Y_{11} & Y_{12} \\ \vdots & \vdots \\ Y_{n1} & Y_{n2} \end{pmatrix}_{n \times 2}$$

and

$$X = \begin{pmatrix} X_{11} & \dots & X_{1p} \\ \vdots & \vdots & \vdots \\ X_{n1} & \dots & X_{np} \end{pmatrix}.$$

The usual estimate of  $V_{yy} - V_{yx} V_{xx}^{-1} V_{xy}$  is

$$\begin{aligned} & (n-1)^{-1}(Y'(I - n^{-1}J_n^n)Y \\ & - Y'(I - n^{-1}J_n^n)X(X'(I - n^{-1}J_n^n)X)^{-1}X'(I - n^{-1}J_n^n)Y) \\ & = Y'(I - n^{-1}J_n^n)Y - Y'(M^* - n^{-1}J_n^n)Y \\ & = Y'(I - M^*)Y. \end{aligned}$$

The estimate of  $\rho_{y,x}$  is thus  $r_{y,x}$ , where

$$r_{y,x} = \frac{Y_2'(I - M^*)Y_1}{[(Y_1'(I - M^*)Y_1)(Y_2'(I - M^*)Y_2)]^{1/2}}.$$

# Regression Analysis

We call  $r_{y,x}$  the sample partial correlation coefficient between  $(Y_1, Y_2)$ . From the form of  $r_{y,x}$ , we see that  $r_{y,x}$  can be viewed as the correlation between the residuals of the regression models,

$$Y_1 = J_n \beta_0 + X\beta + \epsilon$$

and

$$Y_2 = J_n \beta_0 + X\beta + \epsilon .$$

The quantity  $r_{y,x}^2$  has a nice relationship to another linear model. Consider fitting

$$Y_1 = J_n \gamma_0 + X\gamma_1 + Y_2\gamma_2 + \epsilon .$$

Since  $C((I - M^*)Y_2)$  is the orthogonal complement of  $C(J_n, X)$  with respect to  $C(J_n, X, Y_2)$ , the sum of squares of testing whether  $Y_2$  adds to the model is

$$\begin{aligned} & SSR(Y_2 | J_n, X) \\ &= Y_1'(I - M^*)Y_2(Y_2'(I - M^*)Y_2)^{-1}Y_2'(I - M^*)Y_1 . \end{aligned}$$

# Regression Analysis

Since  $Y_2'(I - M^*)Y_2$  is a scalar, it is easily seen that

$$r_{y.x}^2 = \frac{SSR(Y_2 | J_n, X)}{SSE(J_n, X)} ,$$

where

$$SSE(J_n, X) = Y_1'(I - M^*)Y_1$$

is the error sum of squares for fitting the model

$$Y_1 = J_n\beta_0 + X\beta + \epsilon .$$

The hypothesis of

$$H_0 : \rho_{y.x} = 0$$

has the test statistic

$$F = \frac{r_{y.x}^2 / 1}{(1 - r_{y.x}^2) / (n - p - 1)} .$$

This statistic has an  $F(1, n - p - 1)$  distribution under  $H_0$ .  $r_{y.x}^2$  is often called the coefficient of partial determination.

# Regression Analysis

## Example

Consider Example 6.5.1 on page 126 of Christensen. We can compute the squared sample correlation coefficient between  $(Y, X_2)$  adjusting for  $X_1$ . We have  $SSR(X_2 | J_n, X_1) = 6.70$ ,

$$\begin{aligned} SSE(J_n, X_1) &= SSE(J_n, X_1, X_2) + SSR(X_2 | J_n, X_1) \\ &= 46.50 + 6.70 \\ &= 53.20 . \end{aligned}$$

Thus,

$$r_{y2.1} = \frac{6.70}{53.20} = .1259 .$$

# Regression Analysis

## Pure Error and Lack of Fit

The lack of fit problem can be described as follows. Suppose we have the linear model

$$Y = X\beta + \epsilon,$$

where  $\epsilon \sim N_n(0, \sigma^2 I)$ , and we suspect that the model is an inadequate explanation of the data. One way to correct this problem is to augment the model and include more predictors. That is, we fit a model

$$Y = Z\gamma_0 + \epsilon,$$

where  $C(X) \subset C(Z)$  and  $\epsilon \sim N_n(0, \sigma^2 I)$ . The two questions that arise here are:

- i) How does one choose  $Z$ ?
- ii) Is there really lack of fit?

Given a  $Z$ , question 2 can be addressed by carrying out the usual nested vs. full  $F$  test of

$$H_0 : E(Y) \in C(X)$$

$$H_a : E(Y) \in C(Z) \cap (C(X))^c.$$

In this setting, we call the error sum of squares based on fitting  $Z$ , the sum of squares for pure error (SSPE).

# Regression Analysis

Thus,

$$SSPE = SSE(Z) .$$

Also, the sum of squares for lack of fit (SSLF) is the difference in the sum of squares between  $X$  and  $Z$ . That is

$$SSLF = SSE(X) - SSE(Z) .$$

In general, there are few theoretical guidelines for choosing  $Z$ . The most common situation is the variable selection problem. We do not address this problem here. Here, we focus on obtaining  $Z$  from  $X$ . We discuss the following two methods.

- a) Some rows of  $X$  are replicated. That is, we take multiple observations at the same covariate values. This is the traditional method of doing lack of fit.
- b) Examination of different subsets of the data.

# Regression Analysis

We now discuss method a). We need notation for identifying which rows of  $X$  are identical. A regression model with replications can be written as

$$Y_{ij} = X'_i \beta + \epsilon_{ij},$$

where  $X'_i = (X_{i1}, \dots, X_{ip})$ ,  $i = 1, \dots, c$ ,  $j = 1, \dots, n_i$ . Thus, there are  $n_i$  observations for each  $X_i$ , and there are  $c$  distinct rows of  $X$ . Thus  $X$  is an  $n \times p$  matrix with  $n = \sum_{i=1}^c n_i$ . We can write  $X$  as

$$X = \begin{pmatrix} X'_1 \\ \vdots \\ X'_1 \\ \vdots \\ X'_c \\ \vdots \\ X'_c \end{pmatrix}_{n \times p}$$

# Regression Analysis

We assume that  $X'_i \neq X'_k$  for  $i \neq k$ . Using the notation established in Chapter 4, we write

$$Y = (Y_{ij})$$

and

$$X = (W'_{ij}) ,$$

where  $W'_{ij} = X'_i$ . The idea behind pure error is that when there are rows of  $X$  that are replicated, then several observations have the same mean value, and the variability about the mean value is in some sense pure error. The problem is to estimate the mean value. If we estimate the mean value in the  $i$ th group by  $X'_i \hat{\beta}$ , then estimate the variance for the group by looking at deviations about the estimated mean value, and then finally pool estimates from different groups, we get  $SSE(X)$ .

# Regression Analysis

We now describe the general method of finding  $Z$  when the rows of  $X$  are replicated. Consider the model

$$Y = Z\gamma_0 + \epsilon,$$

where  $C(X) \subset C(Z)$ ,  $\epsilon \sim N_n(0, \sigma^2 I)$ , and  $Z$  is chosen so that

$$Z = (Z'_{ij})$$

where  $Z'_{ij} = V'_i$  for some vector  $V'_i$ ,  $i = 1, \dots, c$ . Two rows of  $Z$  are the same if and only if the corresponding rows of  $X$  are the same. We will refer to the property that rows of  $Z$  are identical if and only if the corresponding rows of  $X$  are identical, as  $X$  and  $Z$  have the same row structure.

Since  $X$  has  $c$  distinct rows,  $r(X) \leq c$ . since  $Z$  has  $c$  distinct rows,  $r(Z) \leq c$ . The most general  $Z$  will have  $r(Z) = c$ . Thus, we need to find a  $Z$  such that  $C(X) \subset C(Z)$ ,  $r(Z) = c$ , and  $Z$  has the same row structure as  $X$ . We want to also show that the column space is the same for any such  $Z$ .

# Regression Analysis

We pick  $Z$  as follows.

- 1) Let  $Z$  be the design matrix for the one-way ANOVA model

$$Y_{ij} = \mu_i + \epsilon_{ij},$$

$i = 1, \dots, c$ , and  $j = 1, \dots, n_i$ . Thus  $Z$  corresponds to the design matrix for a one-way ANOVA model with  $c$  treatments and  $n_i$  observations per treatment. Let  $Z_{ij,k}$  denote the element in the  $ij$ th row and  $k$ th column of  $Z$ . Then

$$Z = (Z_{ij,k})$$

where  $Z_{ij,k} = \delta_{ik}$ , and  $\delta_{ik} = 1$  if  $i = k$  and 0 otherwise.  $Z$  is a matrix where the  $k$ th column is 0 everywhere except that it has ones in rows that correspond to the  $Y_{kj}$ 's. Since the values  $Z_{ij,k}$  do not depend on  $j$ , it is clear that  $Z$  has the same row structure as  $X$ . Since the  $c$  columns of  $Z$  are linearly independent (actually orthogonal), we have  $r(Z) = c$ . Also, we have

$$X = M_Z X,$$

where  $M_Z$  is the orthogonal projection operator onto  $C(Z)$ . Thus  $C(X) \subset C(Z)$ . Any matrix  $Z_1$  with the same row structure as  $X$  must have  $Z_1 = M_Z Z_1$  and  $C(Z_1) \subset C(Z)$ . If  $r(Z_1) = c$ , then  $C(Z_1) = C(Z)$ . Thus the column space of the most general model does not depend on  $Z$ .

# Regression Analysis

To test the hypothesis

$$\begin{aligned} H_0 &: E(Y) \in C(X) \\ H_a &: E(Y) \in C(Z) \cap (C(X))^c, \end{aligned}$$

where  $C(X) \subset C(Z)$ , we note that  $C(Z)$  is the column space for a one-way ANOVA. Let

$$M_z = Z(Z'Z)^{-1}Z'.$$

We know that  $Z$  has full rank  $c$  and  $M_z$  is block diagonal with  $i$ th block equal to  $n_i^{-1}J_{n_i}^{n_i}$ ,  $i = 1, \dots, c$ . Thus

$$\begin{aligned} SSPE &= SSE(Z) \\ &= Y'(I - M_z)Y \\ &= \sum_{i=1}^c \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.})^2. \end{aligned}$$

# Regression Analysis

We see that this is the SSE for the usual one-way ANOVA. Also, we have

$$\begin{aligned}SSLF &= Y'(M_z - M)Y \\&= \sum_{i=1}^c \sum_{j=1}^{n_i} (\bar{Y}_{i\cdot} - \hat{Y}_i)^2 \\&= \sum_{i=1}^c n_i (\bar{Y}_{i\cdot} - \hat{Y}_i)^2\end{aligned}$$

where  $M = X(X'X)^{-1}X'$ ,  $\hat{Y}_i = X_i'\hat{\beta}$ , and  $\hat{\beta}$  is a solution to  $X\hat{\beta} = MY$ . Also, we have

$$\bar{Y}_{i\cdot} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij} .$$

Notice here that we have

$$MY = (\hat{Y}_1, \dots, \hat{Y}_1, \dots, \hat{Y}_c, \dots, \hat{Y}_c)'$$

where  $\hat{Y}_i$  is repeated  $n_i$  times.

# Regression Analysis

The  $F$  test for lack of fit now takes the form

$$\begin{aligned} F &= \frac{Y'(M_z - M)Y/r(M_z - M)}{Y'(I - M_z)Y/r(I - M_z)} \\ &\sim F(r(M_z - M), r(I - M_z), \gamma) \end{aligned}$$

where

$$\gamma = \frac{\|(M_z - M)X\beta\|^2}{2\sigma^2}.$$

We note here that  $r(M_z - M) = c - r(M)$ , and if  $X$  is of rank  $p$ , then  $r(M) = p$ . Clearly, we need  $c > p$ . Also  $r(I - M_z) = n - c$ . Thus the  $F$  test can be written as

$$F = \frac{SSLF/(c - p)}{SSPE/(n - c)}.$$

The second method assumes no replication of the rows of  $X$ . In this case, we can still do a lack of fit test by partitioning the data. Write

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$$

and

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}$$

# Regression Analysis

The model  $Y = Z\gamma_0 + \epsilon$  can be chosen with

$$Z = \begin{pmatrix} X_1 & 0 \\ 0 & X_2 \end{pmatrix}_{n \times 2p}.$$

We clearly have  $C(X) \subset C(Z)$ . The idea behind the lack of fit test based on the partitioned data is the hope that  $X_1$  and  $X_2$  will be chosen so that the combined fit of  $Y_1 = X_1\beta + \epsilon$  and  $Y_2 = X_2\beta + \epsilon$  will be qualitatively better than the fit of  $Y = X\beta + \epsilon$ .

# Regression Analysis

## Example

Consider the model  $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ ,  $i = 1, \dots, 2r$  with  $X_1 \leq X_2 \dots \leq X_{2r}$ .

Suppose that the lack of fit is due to the true model being

$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \epsilon_i$  so that the true curve is a parabola. Clearly, one can approximate a parabola better with 2 lines than with 1 line. The combined fit of  $Y_i = \eta_0 + \eta_1 X_i + \epsilon_i$ ,  $i = 1, \dots, r$ , and  $Y_i = \tau_0 + \tau_1 X_i + \epsilon_i$ ,  $i = r + 1, \dots, 2r$  should be better than the unpartitioned fit.

The partitioning method can be extended to more than 2 partitions. For example, for 3 partitions, we have

$$X = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix},$$

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix},$$

# Regression Analysis

and

$$Z = \begin{pmatrix} X_1 & 0 & 0 \\ 0 & X_2 & 0 \\ 0 & 0 & X_3 \end{pmatrix} .$$

## Remark

In the traditional method, if one assumes that the true model is  $Y = W\delta + \epsilon$ , where  $W$  and  $X$  have the same row structure, then  $C(X) \subset C(W) \subset C(Z)$ . In this case, the lack of fit test based on  $X$  and  $Z$  has a noncentral  $F$  distribution under  $H_0$ , with noncentrality parameter

$$\gamma = \frac{\|(M_z - M)W\delta\|^2}{2\sigma^2} .$$

# Regression Analysis

## Polynomial Regression

The general polynomial regression model is given by

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \dots + \beta_{p-1} X_i^{p-1} + \epsilon_i , \quad (4)$$

where  $\epsilon_i$  are *i.i.d.*  $N(0, \sigma^2)$  random variables. We say that the model in (4) is a polynomial regression model of degree  $p - 1$ . These models should be considered when there is, say, one covariate and it is suspected that  $Y$  is a nonlinear function of the covariate.

In these models it is often of interest to determine the lowest degree polynomial which fits the data. Thus it is of interest to test hypotheses concerning the  $\beta_j$ 's. In fitting polynomial models, we follow the convention that if a particular power is included in the model, then all lower order powers are also included. Thus in doing tests of hypotheses, we test the higher order powers first.

Sometimes polynomial regression models are fit using orthogonal polynomials. This is a procedure that allows one to perform all of the appropriate tests on the  $\beta_j$ 's without having to fit more than one regression model. The technique uses the Gram-Schmidt algorithm to orthogonalize the columns of the design matrix and then fits a model to the orthogonalized columns.

# Regression Analysis

Since Gram-Schmidt orthogonalizes vectors sequentially, the matrix with orthogonal columns can be written

$$T = XP , \quad (5)$$

where  $X$  is the  $n \times p$  covariate matrix and  $P$  is a  $p \times p$  non-singular upper triangular matrix. Notice that (5) is the QR decomposition of  $X$  in disguise since

$$X = TP^{-1} .$$

Thus  $T$  plays the role of  $Q$  and  $P^{-1}$  plays the role of  $R$ .

The  $n \times p$  design matrix  $X$  from model (4) has the form

$$X = \begin{pmatrix} 1 & X_1 & X_1^2 & \dots & X_1^{p-1} \\ \vdots & X_2 & X_2^2 & \dots & X_2^{p-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_n & X_n^2 & \dots & X_n^{p-1} \end{pmatrix}$$

# Regression Analysis

and

$$P = \begin{pmatrix} P_{11} & P_{12} & \dots & P_{1p} \\ 0 & P_{22} & \dots & P_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & P_{pp} \end{pmatrix}.$$

Thus, we can write the  $i$ th row of  $T$  from (5) as

$$\begin{aligned} & (1, X_i, X_i^2, \dots, X_i^{p-1}) \begin{pmatrix} P_{11} & P_{12} & \dots & P_{1p} \\ 0 & P_{22} & \dots & P_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & P_{pp} \end{pmatrix} \\ &= (P_{11}, P_{12} + P_{22}X_i, P_{13} + P_{23}X_i + P_{33}X_i^2, \\ &\quad \dots, P_{1p} + P_{2p}X_i + \dots + P_{pp}X_i^{p-1}) \\ &= (\psi_0(X_i), \psi_1(X_i), \psi_2(X_i), \dots, \psi_{p-1}(X_i)). \end{aligned} \tag{6}$$

# Regression Analysis

Notice that (6) consists of polynomials. In particular,  $\psi_r(X_i)$  is a polynomial of degree  $r$ . Thus, we can write  $T$  as

$$T = \begin{pmatrix} P_{11} & \psi_1(X_1) & \dots & \psi_{p-1}(X_1) \\ P_{11} & \psi_1(X_2) & \dots & \psi_{p-1}(X_2) \\ \vdots & \vdots & \vdots & \vdots \\ P_{11} & \psi_1(X_n) & \dots & \psi_{p-1}(X_n) \end{pmatrix}.$$

Thus, the first column of  $T$  consists of polynomials of degree 0, the second column of  $T$  consists of polynomials of degree 1, the third column of  $T$  consists of polynomials of degree 2, and so on. Thus, the  $r$ th column of  $T$  consists of polynomials of degree  $r - 1$ ,  $r = 1, \dots, p$ .

Now the model

$$Y = X\beta + \epsilon, \quad \beta = (\beta_0, \beta_1, \dots, \beta_{p-1})', \quad (7)$$

is equivalent to

$$Y = T\gamma + \epsilon, \quad (8)$$

where

$$\gamma = P^{-1}\beta, \quad \gamma = (\gamma_0, \gamma_1 \dots \gamma_{p-1})'.$$

The columns of  $T$  are orthogonal, and form what are called orthogonal polynomials, that is,

$$\sum_{i=1}^n \psi_j(X_i)\psi_k(X_i) = 0, \quad j \neq k.$$

# Regression Analysis

Since  $P$  is upper triangular,  $P^{-1}$  is also upper triangular. Therefore,  $\gamma_j$  is a linear function of  $\beta_j, \beta_{j+1}, \dots, \beta_{p-1}$ . In particular,  $\gamma_j = \sum_{k=j}^{p-1} P^{(jk)} \beta_k$ ,  $j = 0, \dots, p-1$ , where  $P^{(jk)}$  is the  $(j, k)$ th element of the matrix  $P^{-1}$ . Thus, the test of  $H_0 : \gamma_{p-1} = 0$  is equivalent to the test of  $H_0 : \beta_{p-1} = 0$ . If  $\beta_{j+1} = \beta_{j+2} = \dots = \beta_{p-1} = 0$ , then the test of  $H_0 : \gamma_j = 0$  is equivalent to  $H_0 : \beta_j = 0$ . Thus the test for  $\gamma_j = 0$  is the same as the test for  $\beta_j = 0$  in the model

$$Y_i = \beta_0 + \beta_1 X_i + \dots + \beta_j X_i^j + \epsilon_i .$$

Since the columns of  $T$  are orthogonal, the sum of squares for testing  $H_0 : \gamma_j = 0$  depends only on the column of  $T$  associated with  $\gamma_j$ .

The  $p - 1$  different polynomials that are contained in each row of  $T$  are orthogonal only in that the coefficients of the polynomials were determined so that  $T$  has columns that are orthogonal.

Suppose that there are  $q$  distinct values of  $X_i$  in the design matrix. Then  $n - q$  of the rows of  $X$  will be identical and thus  $r(X) = q$ . In this case, the most general polynomial that can be fit must give a rank  $q$  design matrix, and thus the model must be

$$Y_{ij} = \beta_0 + \beta_1 X_i + \dots + \beta_{q-1} X_i^{q-1} + \epsilon_{ij} , \quad (9)$$

where  $j = 1, \dots, n_i$ ,  $i = 1, \dots, q$ , and  $n = \sum_{i=1}^q n_i$ .

# Regression Analysis

It also follows that the column space of the model in (9) is exactly the same as the column space for a one-way ANOVA model with  $q$  treatments. That is, the models

$$Y_{ij} = \beta_0 + \beta_i X_i + \dots + \beta_{q-1} X_i^{q-1} + \epsilon_{ij} \quad (10)$$

and

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij} \quad (11)$$

$i = 1, \dots, q$  and  $j = 1, \dots, n_i$  are equivalent.

Thus, the column space of the design matrix in (10) is the same as the column space of the model in (11). We are led to the following theorem.

## Theorem

The column spaces of the design matrices of models (10) and (11) are identical.

The proof is left as a straightforward exercise.

# Regression Analysis

Thus a test of

$$H_0 : \beta_1 = \dots = \beta_{q-1}$$

from (10) is identical to the test

$$H_0 : \alpha_1 = \dots = \alpha_q$$

from (11). Let  $C(X)$  denote the column space of (10) (or (11)). We can break up  $C(X)$  into  $q - 1$  orthogonal one-dimensional subspaces, each subspace corresponding to an orthogonal polynomial. Thus any vector in  $C(X)$  that is orthogonal to  $J_n$  corresponds to a contrast in the  $\alpha_i$ 's.

In particular each orthogonal polynomial corresponds to a contrast in the  $\alpha_i$ 's. We are led to the following definition.

## Defn

The orthogonal contrasts determined by the orthogonal polynomials are called the polynomial contrasts. The contrast corresponding to the first degree orthogonal polynomial is called the linear contrast. The contrasts for higher degree orthogonal polynomials are called the quadratic, cubic, quartic, etc. ... contrasts.

# Regression Analysis

Orthogonal contrasts are of interest in one-way ANOVA when we have quantitative levels of the treatment. Then, in this case, it is of interest to test for linear, quadratic, cubic, etc. ... trends.

To see how to construct a contrast from the orthogonal polynomials, consider the following example. Let us construct the linear contrast, that is, the contrast corresponding to the orthogonal polynomial of degree 1. The second column of the design matrix  $X$  is given by

$$\tilde{X}_1 = [t_{ij}]$$

where  $t_{ij} = X_i$  for all  $i$  and  $j$ . If we orthogonalize this with respect to  $J_n$ , we get the linear orthogonal polynomial.

Letting

$$\bar{X}_1 = \frac{\sum_{i=1}^q n_i X_i}{\sum_{i=1}^q n_i}$$

gives the linear orthogonal polynomial

$$T_1 = [w_{ij}]$$

where

$$w_{ij} = X_i - \bar{X}_1$$

# Regression Analysis

From Chapter 4, we know that this vector corresponds to a contrast of the form  $\sum \lambda_i \alpha_i$  where  $\frac{\lambda_i}{n_i} = X_i - \bar{X}_.$ . Solving for  $\lambda_i$  gives  $\lambda_i = n_i(X_i - \bar{X}_.)$ .

Thus, the sum of squares for testing  $H_0 : \gamma_1 = 0$  is

$$\frac{(\sum n_i(X_i - \bar{X}_.)\bar{Y}_{i.})^2}{\sum n_i(X_i - \bar{X}_.)^2}.$$

If  $n_i = N$  and the quantitative levels of the  $X_i$ 's are equally spaced, then the orthogonal polynomial contrasts depend only on  $q$ . In this case the contrasts can be tabled (as given in Handout).

The orthogonal polynomials can be written as

$$T = ZB$$

where the  $q \times q$  matrix  $B$  is

$$B = (b_0, b_1, \dots, b_{q-1}).$$

The  $i$ th degree polynomial is  $Zb_i$  and if  $\alpha = (\alpha_1, \dots, \alpha_q)'$ , then for  $i \geq 1$ ,  $Zb_i$  corresponds to a contrast  $b_i'Z'[J, Z]\beta = b_i'Z'Z\alpha = b_i'(NI)\alpha = Nb_i'\alpha$ . Since orthogonal polynomial contrasts are defined only up to constant multiples, the  $i$ th degree polynomial contrast is  $b_i'\alpha$ . Since  $T'T$  is diagonal,  $b_i'b_j = 0$  for  $i \neq j$ . With  $b_0 = J_q$  and tabled values for the other  $b_i$ 's, one can obtain  $T$ .

# Regression Analysis

## Polynomial Regression and the Balanced Two-way ANOVA

Let us first consider the balanced two-way ANOVA without interaction.

Suppose that the  $i$ th level of the  $\alpha$  treatment corresponds to some number  $W_i$  and that the  $j$ th level of the  $\eta$  treatments correspond to some number  $Z_j$ . We can write vectors taking powers of  $W_i$  and  $Z_j$ .

For  $r = 1, \dots, a - 1$ , and  $s = 1, \dots, b - 1$ , write

$$\begin{aligned} W^r &= [t_{ijk}] \quad , \text{ where } t_{ijk} = W_i^r \\ Z^s &= [t_{ijk}] \quad , \text{ where } t_{ijk} = Z_j^s. \end{aligned}$$

Note that  $W^0 = Z^0 = J_n$ .

# Regression Analysis

## Example

Consider the model

$$Y_{ijk} = \mu + \alpha_i + \eta_j + \epsilon_{ijk},$$

where  $i = 1, 2, 3$ ,  $j = 1, 2$ , and  $k = 1, 2$ .

Suppose that the  $\alpha$  treatments are 2, 4, 6 mg of drug A and that the  $\eta$  treatments are 5, 7 mg of drug B. Then we have  $Y = (y_{111}, y_{112}, \dots, y_{322})'$ ,

$$W^1 = (2, 2, 2, 4, 4, 4, 6, 6, 6)',$$

$$W^2 = (4, 4, 4, 16, 16, 16, 36, 36, 36)'$$

and

$$Z^1 = (5, 5, 7, 7, 5, 5, 7, 7, 5, 5, 7, 7)'.$$

In terms of column spaces, we have

$$C(J_n, W^1, \dots, W^{a-1}) = C(X_0, \dots, X_a)$$

and

$$C(J_n, Z^1, \dots, Z^{b-1}) = C(X_0, X_{a+1}, \dots, X_{a+b})$$

where the main effects model is given by

$$Y = X\beta + \epsilon,$$

and  $X = (J, X_1, \dots, X_a, X_{a+1}, \dots, X_{a+b})$ .

# Regression Analysis

Fitting the two-way ANOVA is the same as fitting a joint polynomial in  $W_i$  and  $Z_j$ . Algebraically, the model

$$Y_{ijk} = \mu + \alpha_i + \eta_j + \epsilon_{ijk} ,$$

$i = 1, \dots, a$ ,  $j = 1, \dots, b$ ,  $k = 1, \dots, N$  is equivalent to

$$\begin{aligned} Y_{ijk} &= \beta_{0,0} + \beta_{1,1} W_i \dots + \beta_{1,a-1} W_i^{a-1} + \beta_{2,1} Z_j \\ &\quad + \dots + \beta_{2,b-1} Z_j^{b-1} + \epsilon_{ijk} . \end{aligned}$$

The correspondence between contrasts and orthogonal polynomials remains valid.

If we consider the two-way balanced ANOVA with interaction, the equivalent polynomial regression model is given by

$$Y_{ijk} = \sum_{r=0}^{a-1} \sum_{s=0}^{b-1} \beta_{rs} W_i^r Z_j^s + \epsilon_{ijk} .$$

Thus the polynomial regression model now contains cross product terms between the  $W_i$ 's and the  $Z_j$ 's.

# Regression Analysis

## Example (continued)

The model

$$Y_{ijk} = \mu + \alpha_i + \eta_j + \gamma_{ij} + \epsilon_{ijk}$$

is equivalent to

$$Y_{ijk} = \beta_{0,0} + \beta_{10}W_i + \beta_{20}W_i^2 + \beta_{01}Z_j + \beta_{11}W_iZ_j + \beta_{21}W_i^2Z_j + \epsilon_{ijk}$$

where  $W_1 = 2, W_2 = 4, W_3 = 6, Z_1 = 5, Z_2 = 7.$

# Regression Analysis

The design matrix for the general polynomial model can be written as

$$S = (J_n, W^1, W^2, \dots, W^{a-1}, W^1 Z^1, \dots, W^1 Z^{b-1}, W^2 Z^1, \dots, W^{a-1} Z^{b-1}). \quad (12)$$

To establish the equivalence of the models, it is enough to notice that the row structure of  $X = (J_n, X_1, \dots, X_{a+b+ab})$  is the same as the row structure of  $S$  and that  $r(X) = ab = r(S)$ . It is also easily seen that  $C(X) \subset C(S)$ . Since  $r(X) = r(S)$ , it follows that  $C(X) = C(S)$ , and thus the models are equivalent. Now we would like to make a connection between tests in the interaction space determined by the polynomial representation of the model.

For example, we would like to know the test in the interaction space that is determined by, say, the quadratic contrast in the  $\alpha_i$ 's and the cubic contrast in the  $\eta_j$ 's. It turns out that it is a test for  $W^2 Z^3$ . That is, it is a test corresponding to the hypothesis  $H_0 : \beta_{23} = 0$  in the model

$$\begin{aligned} Y_{ijk} &= \beta_{0,0} + \beta_{10} W_i + \beta_{20} W_i^2 + \beta_{01} Z_j + \beta_{02} Z_j^2 + \beta_{03} Z_j^3 \\ &+ \beta_{11} W_i Z_j + \beta_{12} W_i Z_j^2 + \beta_{13} W_i Z_j^3 + \beta_{21} W_i^2 Z_j \\ &+ \beta_{22} W_i^2 Z_j^2 + \beta_{23} W_i^2 Z_j^3 + \epsilon_{ijk}. \end{aligned}$$

# Regression Analysis

In general, we want to identify columns  $W^r Z^s$ ,  $r \geq 1, s \geq 1$  with vectors in the interaction space. It turns out that the test of  $W^r Z^s$  based on the model  $C([W^i Z^j : i = 0, \dots, r, j = 0, \dots, s])$  is precisely the test of the vector in the interaction space defined by the  $r$ th degree polynomial contrast in the  $\alpha_i$ 's and the  $s$ th degree polynomial contrast in the  $\eta_j$ 's.

We note here that the test of the  $r$ th degree polynomial contrast in the  $\alpha_i$ 's is a test of whether the column  $W^r$  adds to the model based on  $C(J_n, W^1, \dots, W^r)$  and that the test of the  $s$ th degree polynomial contrast in the  $\eta_j$ 's is a test of whether the column  $Z^s$  adds to the model based on  $C(J_n, Z^1, \dots, Z^s)$ . We note here that the test for  $W^r Z^s$  adding to the model is not a test for  $W^r Z^s$  adding to the full model. It is a test for  $W^r Z^s$  adding to the model spanned by the vectors  $\{W^i Z^j, i = 0, \dots, r, j = 0, \dots, s\}$ , where  $W^0 = Z^0 = J_n$ .

Now we verify the connection between vectors in the interaction space and polynomial contrasts. The design matrix corresponding to the polynomial regression model is given by (12). We first apply Gram-Schmidt to  $S$  and obtain vectors  $R_{0,0}, R_{1,0} \dots, R_{a-1,0}, R_{0,1} \dots R_{0,b-1}, R_{1,1}, \dots R_{a-1,b-1}$ . The  $R_{ij}$ 's are obtained by orthogonalizing the columns of  $S$ .

# Regression Analysis

For notational convenience, for any vectors  $U = (U_1 \dots U_n)', V = (V_1 \dots V_n)',$  we define  $VU = (V_1 U_1, V_2 U_2, \dots, V_n U_n).$  The polynomial contrasts in the  $\alpha_i$ 's correspond to the vectors  $\{R_{i,0}, i = 1, \dots, a - 1\}$  with

$C(R_{1,0}, \dots, R_{a-1,0}) = C(M\alpha).$  Similarly, the polynomial contrast in the  $\eta_j$ 's corresponds to the vectors  $\{R_{0,j}, j = 1, \dots, b - 1\}$  with

$C(R_{0,1}, \dots, R_{0,b-1}) = C(M\eta).$

Now we need to examine the relationship between vectors in  $C(M_\alpha)$  and  $C(M_\eta)$  with vectors in the interaction space. Take a contrast in the  $\alpha_i$ 's with contrast coefficients  $(d_1, \dots, d_a)$  and a contrast in the  $\eta_j$ 's, with contrast coefficients  $(c_1, \dots, c_b).$

# Regression Analysis

Further, define

$$\rho_1 = [t_{ijk}] \quad , \quad t_{ijk} = \frac{d_i}{bN} \quad ,$$

and

$$\rho_2 = [t_{ijk}] \quad , \quad t_{ijk} = \frac{c_j}{aN} \quad .$$

Then  $\rho_1 \in C(M_\alpha)$  and  $\rho_2 \in C(M_\eta)$ . Also,

$$\rho'_1 X \beta = \sum_{i=1}^a d_i (\alpha_i + \bar{\gamma}_{i\cdot})$$

and

$$\rho'_2 X \beta = \sum_{j=1}^b c_j (\eta_j + \bar{\gamma}_{\cdot j}) \quad .$$

The vector

$$\rho_1 \otimes \rho_2 = [t_{ijk}] \quad ,$$

where  $t_{ijk} = \frac{d_i c_j}{N^2 ab}$ . This is a vector proportional to a vector in the interaction space corresponding to  $(d_1, \dots, d_a)$  and  $(c_1, \dots, c_b)$ .

# Regression Analysis

From this argument, it follows that since  $R_{r,0}$  is the vector in  $C(M\alpha)$  for testing the  $r$ th degree polynomial and  $R_{0,s}$  is the vector in  $C(M_\eta)$  for testing the  $s$ th degree polynomial, then  $R_{r,0}R_{0,s}$  is a vector in the interaction space.

Since the polynomial contrasts are defined to be orthogonal, and since  $R_{r,0}$  and  $R_{0,s}$  are defined by polynomial contrasts, it follows that the set  $\{R_{r,0}R_{0,s} : r = 1, \dots, a-1, s = 1, \dots, b-1\}$  is an orthogonal basis for the interaction space. Moreover with  $R_{r,0}$  and  $R_{0,s}$  orthonormal, we have

$$[R_{r,0}R_{0,s}]' [R_{r,0}R_{0,s}] = 1/abN .$$

# Regression Analysis

It remains to check whether the vector provides a test of the correct thing, that is, that  $W^r Z^s$  adds to a model containing all lower order terms. Since by Gram-Schmidt, for some  $a_i$ 's and  $b_j$ 's we have

$$R_{r,0} = a_0 W^r + a_1 W^{r-1} + \dots + a_{r-1} W^1 + a_r J_n$$

and

$$R_{0,s} = b_0 Z^s + b_1 Z^{s-1} + \dots + b_{s-1} Z^1 + b_s J_n,$$

we also have

$$\begin{aligned} R_{r,0} R_{0,s} &= a_0 b_0 W^r Z^s + \sum_{j=1}^s b_j Z^{s-1} W^r \\ &\quad + \sum_{i=1}^r a_i W^{r-i} Z^s + \sum_{j=1}^s \sum_{i=1}^r a_i b_j Z^{s-j} W^{r-i}. \end{aligned}$$

Letting  $R_{1,0} = R_{0,1} = J_n$ , it follows that

$$\begin{aligned} &C(R_{i,0}, R_{0,j} : i = 0, \dots, r, j = 0, \dots, s) \\ &\subset C(W^i Z^j, i = 0, \dots, r, j = 0, \dots, s). \end{aligned}$$

# Regression Analysis

The vectors listed in each of the sets are linearly independent and the number of vectors in each set is the same, so the ranks of the column spaces are the same, and thus

$$\begin{aligned} & C(R_{i,0}, R_{0,j} : i = 0, \dots, r, j = 0, \dots, s) \\ &= C(W^i Z^j, i = 0, \dots, r, j = 0, \dots, s). \end{aligned}$$

The vectors  $R_{i,0}R_{0,j}$  are orthogonal, so  $abN[R_{r,0}R_{0,s}][R_{r,0}R_{0,s}]'$  is the projection operator for testing if  $W^r Z^s$  adds to the model after fitting all terms of lower order. Since  $R_{r,0}R_{0,s}$  was found as the vector in the interaction space corresponding to the  $r$ th orthogonal polynomial contrast in the  $\alpha_i$ 's, and the  $s$ th polynomial contrast in the  $\eta_j$ 's, the technique for testing if  $W^r Z^s$  adds to the appropriate model is then a test of an interaction contrast.

# Regression Analysis

## Polynomial Contrasts for BIB Designs

Consider the BIB model given by

$$Y_{ij} = \mu + \beta_i + \tau_j + \epsilon_{ij}$$

where  $\epsilon_{ij}$  are *i.i.d.*  $N(0, \sigma^2)$ ,  $i = 1, \dots, b$ ,  $j \in D_i$ , where  $D_i$  is the set of indices of the treatments in block  $i$ .

We now want to derive polynomial contrasts for BIB models. It turns out that the orthogonal polynomial contrasts for the BIB design are the same as for a balanced one-way ANOVA.

Let  $Z$  be the design matrix of a balanced one-way ANOVA without the grand mean. That is,  $Z$  corresponds to the design matrix of

$$Y_{ij} = \mu_i + \epsilon_{ij},$$

$$i = 1, \dots, t, j = 1, \dots, N.$$

Define orthogonal polynomials by

$$T = ZB$$

by ignoring blocks, where  $B$  is matrix of contrast coefficients,

$B = (b_1, \dots, b_{t-1})$ . If the treatments are levels of a single quantitative factor, then the  $b_j$ 's are tabled orthogonal polynomial contrasts. If the treatments have a factorial structure, the  $b_j$ 's are obtained from tabled contrasts as in Example 9.4.1 on page 210 of Christensen.

# Regression Analysis

We note here that

$$J'_t b_j = 0, \quad j = 1, \dots, t-1,$$

and

$$b_i' b_j = 0, \quad i \neq j.$$

We can write the BIB model in regression form as

$$Y = X\beta + T\eta + \epsilon,$$

where

$$(\eta_1, \dots, \eta_{t-1})'.$$

For a simple treatment structure,  $\eta_j$  would be the coefficient for the  $j$ th degree polynomial. For a factorial treatment structure,  $\eta_j$  would be the coefficient for some main effect polynomial term or the coefficient of a cross-product term.

The model

$$Y = X\beta + T\eta + \epsilon \tag{13}$$

is equivalent to the model

$$Y = X\delta + (I - M)T\eta + \epsilon, \tag{14}$$

where  $\eta$  is identical in the two models. The test of  $H_0 : \eta_j = 0$  can be performed in model (14). It can be estimated independently from  $\delta$  in model (14).

# Regression Analysis

We are interested in obtaining the contrast in the  $\tau_j$ 's that corresponds to testing  $H_0 : \eta_j = 0$ . This contrast is

$$b'_j \tau ,$$

where  $\tau = (\tau_1, \dots, \tau_t)'$ . We see that the columns of  $(I - M)T$  are orthogonal since

$$\begin{aligned} T'(I - M)T &= B'Z'(I - M)ZB \\ &= \left(\frac{\lambda t}{k}\right) B'WB \\ &= \left(\frac{\lambda t}{k}\right) B'B \quad (\text{diagonal matrix}) \end{aligned}$$

The last equation follows from the fact that  $J'_t b_j = 0$  for all  $j$ .  $B'B$  is diagonal since  $b'_i b_j = 0$  for all  $i \neq j$ . Finally, the contrast that corresponds to testing  $H_0 : \eta_j = 0$  is

$$((I - M)Zb_j)'(I - M)Z\tau = b'_j Z'(I - M)Z\tau = \frac{\lambda t}{k} b'_j W\tau = \frac{\lambda t}{k} b'_j \tau .$$

Equivalently, the contrast is  $b'_j \tau$ . Examine Example 9.4.1, p.210.

# Bayesian Analysis of the Linear Model

## Bayesian Analysis of the Linear Model

The linear model is frequently used in many biostatistical applications, including

- ① dose response modeling
- ② polynomial regression
- ③ exposure assessment,
- ④ analysis of variance problems for comparing treatment groups

(See *Case Studies in Biometry* by Lange et al., John Wiley & Sons.)

The linear model can be written as

$$Y = X\beta + \epsilon , \quad (15)$$

where

$Y$  is  $n \times 1$ ,  $X$  is  $n \times p$ ,  $\beta$  is  $p \times 1$  and

$$\epsilon \sim N_n(0, \sigma^2 I) . \quad (16)$$

# Bayesian Analysis of the Linear Model

Let  $M = X(X'X)^{-1}X'$ , and  $\tau = \frac{1}{\sigma^2}$ , where the  $-$  denotes generalized inverse.  
Recall that the UMVUE of  $\mu = E(Y) = X\beta$  is  $MY$ .

We would like to derive the posterior distributions of  $\beta$  and  $\tau$  under noninformative priors.

## Theorem 1

Suppose  $\tau$  is known,  $X$  is of full rank  $p$ , and

$$\pi(\beta) \propto 1.$$

Then

$$\beta|y, \tau \sim N_p(\hat{\beta}, \tau^{-1}(X'X)^{-1}),$$

where

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

## Proof:

$$\begin{aligned} p(\beta|y, \tau) &\propto \exp\left\{-\frac{\tau}{2}(Y - X\beta)'(Y - X\beta)\right\} \\ &= \exp\left\{-\frac{\tau}{2}\left[Y'(I - M)Y + (\beta - \hat{\beta})'X'X(\beta - \hat{\beta})\right]\right\} \\ &\propto \exp\left\{-\frac{\tau}{2}\left[(\beta - \hat{\beta})'X'X(\beta - \hat{\beta})\right]\right\}. \end{aligned}$$

Note that

# Bayesian Analysis of the Linear Model

$$\begin{aligned} & Y'(I - M)Y + (\beta - \hat{\beta})'X'X(\beta - \hat{\beta}) \\ = & Y'(I - M)Y + \beta'X'X\beta - 2\hat{\beta}'X'X\beta + \hat{\beta}'X'X\hat{\beta} \\ = & Y'(I - M)Y + \beta'X'X\beta - 2Y'X(X'X)^{-1}(X'X)\beta + Y'MY \\ = & Y'Y + \beta'X'X\beta - 2Y'X\beta \\ = & (Y - X\beta)'(Y - X\beta). \end{aligned}$$

Thus

$$p(\beta|y, \tau) \propto \exp \left\{ -\frac{\tau}{2} [(\beta - \hat{\beta})'X'X(\beta - \hat{\beta})] \right\}.$$

We can recognize this as a normal kernel with mean  $\hat{\beta}$  and covariance matrix  $\tau^{-1}(X'X)^{-1}$ . Thus

$$\beta|y, \tau \sim N_p(\hat{\beta}, \tau^{-1}(X'X)^{-1}).$$

# Bayesian Analysis of the Linear Model

The posterior and predictive distributional results for the linear model are a generalization of the iid case discussed on previous pages.

To obtain the iid case for the linear model we set  $X = 1_{n \times 1}$  where  $1_{n \times 1}$  is a  $n \times 1$  vector of ones.

$$y_{n \times 1} = 1_{n \times 1}\beta_{1 \times 1} + \epsilon_{n \times 1}, \quad \epsilon \sim N(0, \sigma^2),$$

which implies that the  $y_i$ 's are iid  $N(\beta, \sigma^2)$ ,  $i = 1, \dots, n$ .

Substituting  $X = 1_{n \times 1}$ , for example, in the previous theorem, we get

$$\beta | y, \tau \sim N_1(\bar{y}, \tau^{-1} n^{-1}),$$

where

$$(X'X)^{-1}X'y = n^{-1}1'y = \bar{y}, \quad \text{and} \quad (X'X)^{-1} = (1'1)^{-1} = 1/n.$$

# Bayesian Analysis of the Linear Model

## Theorem

When  $\tau$  is known, Jeffreys's prior for  $\beta$  is a uniform prior, i.e.,

$$\pi(\beta) \propto 1.$$

## Proof:

$$\log[p(y|\beta, \tau)] = -\frac{n}{2} \log(2\pi) + \frac{n}{2} \log(\tau) - \frac{\tau}{2} (Y - X\beta)'(Y - X\beta).$$

$$\begin{aligned}\frac{\partial}{\partial \beta} \log[p(y|\beta, \tau)] &= \frac{\partial}{\partial \beta} \left[ -\frac{\tau}{2} (Y - X\beta)'(Y - X\beta) \right] \\ &= \frac{\partial}{\partial \beta} \left[ -\frac{\tau}{2} [Y'Y - 2\beta'X'Y + \beta'X'X\beta] \right] \\ &= \tau X'Y - \tau(X'X)\beta.\end{aligned}$$

Also,

$$\frac{\partial^2}{\partial \beta \partial \beta'} \log[p(y|\beta, \tau)] = -\tau(X'X),$$

and therefore,

$$I(\beta) = \tau(X'X).$$

Thus Jeffreys's prior for  $\beta$  is given by

$$\pi(\beta|\tau) \propto |\tau(X'X)|^{\frac{1}{2}} \propto \text{constant}.$$

Thus

$$\pi(\beta|\tau) \propto 1.$$

# Bayesian Analysis of the Linear Model

## Theorem

Consider the linear model in (15) and (16) where both  $\beta$  and  $\tau$  unknown. Then Jeffreys's joint prior for  $(\beta, \tau)$  is given by

$$\pi(\beta, \tau) \propto \tau^{\frac{p}{2}-1}.$$

## Proof:

$$\frac{\partial^2 \log[p(y|\beta, \tau)]}{\partial \beta \partial \beta'} = -\tau(X'X).$$

$$\frac{\partial^2 \log[p(y|\beta, \tau)]}{\partial \tau^2} = -\frac{n}{2\tau^2}.$$

$$\frac{\partial^2 \log[p(y|\beta, \tau)]}{\partial \beta \partial \tau} = X'Y - (X'X)\beta.$$

Thus,

$$I(\beta, \tau) = \begin{bmatrix} \tau(X'X) & 0 \\ 0 & \frac{n}{2\tau^2} \end{bmatrix}.$$

# Bayesian Analysis of the Linear Model

Note that

$$\begin{aligned}-E\left(\frac{\partial^2 \log[\rho(y|\beta, \tau)]}{\partial \beta \partial \tau}\right) &= -E(X'Y) + (X'X)\beta \\ &= -(X'X)\beta + (X'X)\beta \\ &= 0.\end{aligned}$$

Now

$$\begin{aligned}|I(\beta, \tau)| &= |\tau(X'X)|^{\frac{n}{2}} \tau^{-2} \\ &= \tau^p |X'X|^{\frac{n}{2}} \tau^{-2} \\ &= \tau^{p-2} \frac{n}{2} |X'X| \\ &\propto \tau^{p-2}.\end{aligned}$$

Thus

$$\begin{aligned}\pi(\beta, \tau) &\propto |I(\beta, \tau)|^{\frac{1}{2}} \\ &= \tau^{\frac{p-2}{2}} = \tau^{\frac{p}{2}-1}.\end{aligned}$$

Notice that with  $p = 1$  (i.i.d. case),

$$\pi(\beta, \tau) \propto \tau^{-\frac{1}{2}}.$$

# Bayesian Analysis of the Linear Model

## Theorem

Consider the linear model in (15) and (16) with  $(\beta, \tau)$  unknown, and suppose

$$\pi(\beta, \tau) \propto \tau^{-1}.$$

Then

$$\beta|y \sim S_p(n-p, \hat{\beta}, s^2(X'X)^{-1}),$$

where  $s^2 = \frac{Y'(I-M)Y}{n-p}$  and

$$\tau|y \sim \text{gamma}\left(\frac{n-p}{2}, \frac{(n-p)s^2}{2}\right).$$

## Proof:

We have

$$\begin{aligned} p(\beta, \tau|y) &\propto \tau^{\frac{n}{2}-1} \exp\left\{-\frac{\tau}{2} [(Y - X\beta)'(Y - X\beta)]\right\} \\ &= \tau^{\frac{n}{2}-1} \exp\left\{-\frac{\tau}{2} [Y'(I-M)Y + (\beta - \hat{\beta})'X'X(\beta - \hat{\beta})]\right\}. \end{aligned}$$

# Bayesian Analysis of the Linear Model

Thus

$$\begin{aligned} p(\beta|y) & \propto \int_0^{\infty} \tau^{\frac{n}{2}-1} \exp\left\{-\frac{\tau}{2} [Y'(I-M)Y + (\beta - \hat{\beta})'X'X(\beta - \hat{\beta})]\right\} d\tau \\ & \propto [Y'(I-M)Y + (\beta - \hat{\beta})'X'X(\beta - \hat{\beta})]^{-\frac{n}{2}} \end{aligned}$$

Let  $s^2 = \frac{Y'(I-M)Y}{n-p}$ . Then the above integral is

$$\begin{aligned} &= [(n-p)s^2 + (\beta - \hat{\beta})'X'X(\beta - \hat{\beta})]^{-\frac{(n-p+p)}{2}} \\ &\propto \left[1 + \frac{1}{s^2(n-p)}(\beta - \hat{\beta})'X'X(\beta - \hat{\beta})\right]^{-\frac{(n-p+p)}{2}}. \end{aligned}$$

Thus

$$\beta|y \sim S_p(n-p, \hat{\beta}, s^2(X'X)^{-1}).$$

# Bayesian Analysis of the Linear Model

Now

$$p(\tau|y)$$

$$\propto \int_{-\infty}^{\infty} \tau^{\frac{n}{2}-1} \exp \left\{ -\frac{\tau}{2} [Y'(I-M)Y + (\beta - \hat{\beta})'(X'X)(\beta - \hat{\beta})] \right\} d\beta$$

$$= \tau^{\frac{n}{2}-1} \exp \left\{ -\frac{\tau}{2} [Y'(I-M)Y] \right\} \int_{-\infty}^{\infty} \exp \left\{ -\frac{\tau}{2} (\beta - \hat{\beta})'(X'X)(\beta - \hat{\beta}) \right\} d\beta$$

$$\propto \tau^{\frac{n}{2}-1} \exp \left\{ -\frac{\tau}{2} [Y'(I-M)Y] \right\} \tau^{-\frac{p}{2}}$$

$$= \tau^{\frac{n-p}{2}-1} \exp \left\{ -\frac{\tau}{2} [(n-p)s^2] \right\} .$$

Thus

$$\tau|y \sim \text{gamma} \left( \frac{n-p}{2}, \frac{(n-p)s^2}{2} \right) .$$

# Bayesian Analysis of the Linear Model

## Theorem

Consider the linear model in (15) and (16) with  $(\beta, \tau)$  unknown, and suppose

$$\begin{aligned}\beta | \tau &\sim N_p(\mu_0, \tau^{-1} \Sigma_0), \\ \tau &\sim \text{gamma } \left( \frac{\delta_0}{2}, \frac{\gamma_0}{2} \right).\end{aligned}$$

Then

$$\beta | y \sim S_p \left( n + \delta_0, \tilde{\beta}, \tilde{s}^2 (X'X + \Sigma_0^{-1})^{-1} \right),$$

where

$$\tilde{\beta} = \Lambda \mu_0 + (I - \Lambda) \hat{\beta},$$

$$\Lambda = (X'X + \Sigma_0^{-1})^{-1} \Sigma_0^{-1},$$

$$\hat{\beta} = (X'X)^{-1} X' Y,$$

$$\tilde{s}^2 = (n + \delta_0)^{-1} [Y'(I - M)Y + (\hat{\beta} - \mu_0)'(\Lambda' X' X)(\hat{\beta} - \mu_0) + \gamma_0],$$

and

# Bayesian Analysis of the Linear Model

$$\tau|y \sim \text{gamma} \left( \frac{n + \delta_0}{2}, \frac{(n + \delta_0)\tilde{s}^2}{2} \right) .$$

**Proof:**

$$p(\beta|y) \propto \int_0^\infty \tau^{\frac{n+p+\delta_0}{2}-1} e^{-\frac{\tau}{2} [\gamma_0 + Y'(I-M)Y + (\beta - \hat{\beta})'(X'X)(\beta - \hat{\beta}) + (\beta - \mu_0)' \Sigma_0^{-1}(\beta - \mu_0)]} d\tau .$$

Now

$$\begin{aligned} & (\beta - \hat{\beta})'(X'X)(\beta - \hat{\beta}) + (\beta - \mu_0)' \Sigma_0^{-1}(\beta - \mu_0) \\ &= (\beta - \tilde{\beta})'(X'X + \Sigma_0^{-1})(\beta - \tilde{\beta}) - \tilde{\beta}'(X'X + \Sigma_0^{-1})\tilde{\beta} + \hat{\beta}'X'X\hat{\beta} + \mu_0'\Sigma_0^{-1}\mu_0 \end{aligned}$$

Note that

$$\begin{aligned} \Lambda &= (X'X + \Sigma_0^{-1})^{-1}\Sigma_0^{-1}, \\ I - \Lambda &= (X'X + \Sigma_0^{-1})^{-1}X'X, \end{aligned}$$

since  $I - \Lambda + \Lambda$

# Bayesian Analysis of the Linear Model

$$\begin{aligned} &= (X'X + \Sigma_0^{-1})^{-1} X'X + (X'X + \Sigma_0^{-1})^{-1} \Sigma_0^{-1} \\ &= (X'X + \Sigma_0^{-1})^{-1} (X'X + \Sigma_0^{-1}) = I . \end{aligned}$$

Now observe that

$$\begin{aligned} &- \tilde{\beta}'(X'X + \Sigma_0^{-1})\tilde{\beta} + \hat{\beta}'X'X\hat{\beta} + \mu_0'\Sigma_0^{-1}\mu_0 \\ &= -(\Lambda\mu_0 + (I - \Lambda)\hat{\beta})'(X'X + \Sigma_0^{-1})(\Lambda\mu_0 + (I - \Lambda)\hat{\beta}) + Y'MY + \mu_0'\Sigma_0^{-1}\mu_0 \\ &= -\mu_0'\Lambda' (X'X + \Sigma_0^{-1}) \Lambda\mu_0 - \mu_0'\Lambda' (X'X + \Sigma_0^{-1}) (I - \Lambda)\hat{\beta} \\ &\quad - \hat{\beta}'(I - \Lambda)' (X'X + \Sigma_0^{-1}) \Lambda\mu_0 \\ &\quad - \hat{\beta}'(I - \Lambda)' (X'X + \Sigma_0^{-1}) (I - \Lambda)\hat{\beta} + Y'MY + \mu_0'\Sigma_0^{-1}\mu_0 . \end{aligned}$$

Now

$$\begin{aligned} &- \mu_0'\Lambda' (X'X + \Sigma_0^{-1}) \Lambda\mu_0 + \mu_0'\Sigma_0^{-1}\mu_0 \\ &= -\mu_0' (\Sigma_0^{-1}(X'X + \Sigma_0^{-1})^{-1}) (X'X + \Sigma_0^{-1})(X'X + \Sigma_0^{-1})^{-1} \Sigma_0^{-1}\mu_0 + \mu_0'\Sigma_0^{-1}\mu_0 \\ &= -\mu_0'\Sigma_0^{-1}(X'X + \Sigma_0^{-1})^{-1} \Sigma_0^{-1}\mu_0 + \mu_0'\Sigma_0^{-1}\mu_0 \\ &= \mu_0'\Sigma_0^{-1} (I - (X'X + \Sigma_0^{-1})^{-1} \Sigma_0^{-1}) \mu_0 \\ &= \mu_0'\Sigma_0^{-1}(X'X + \Sigma_0^{-1})^{-1} X'X\mu_0 \\ &= \mu_0'\Lambda' X'X\mu_0 . \end{aligned}$$

# Bayesian Analysis of the Linear Model

Now

$$\begin{aligned} & -2\mu_0' \Lambda' \left( X'X + \Sigma_0^{-1} \right) (I - \Lambda) \hat{\beta} \\ &= -2\mu_0' \Sigma_0^{-1} (X'X + \Sigma_0^{-1})^{-1} (X'X + \Sigma_0^{-1}) (X'X + \Sigma_0^{-1})^{-1} X'X \hat{\beta} \\ &= -2\mu_0' \Sigma_0^{-1} (X'X + \Sigma_0^{-1})^{-1} X'X \hat{\beta} \\ &= -2\mu_0' \Lambda' X'X \hat{\beta}. \end{aligned}$$

Finally

$$\begin{aligned} & -\hat{\beta}' (I - \Lambda)' (X'X + \Sigma_0^{-1}) (I - \Lambda) \hat{\beta} \\ &= -\hat{\beta}' \left( X'X (X'X + \Sigma_0^{-1})^{-1} (X'X + \Sigma_0^{-1}) (X'X + \Sigma_0^{-1})^{-1} X'X \right) \hat{\beta} \\ &= -\hat{\beta}' X'X (X'X + \Sigma_0^{-1})^{-1} X'X \hat{\beta} \\ &= -\hat{\beta}' (I - \Lambda') X'X \hat{\beta} \\ &= -\hat{\beta}' X'X \hat{\beta} + \hat{\beta}' \Lambda' X'X \hat{\beta} \\ &= -Y' M Y + \hat{\beta}' \Lambda' X'X \hat{\beta}. \end{aligned}$$

# Bayesian Analysis of the Linear Model

Thus

$$\begin{aligned} & -\tilde{\beta}'(X'X + \Sigma_0^{-1})\tilde{\beta} + \hat{\beta}'X'X\hat{\beta} + \mu_0'\Sigma_0^{-1}\mu_0 \\ &= \mu_0'\Lambda'X'X\mu_0 - 2\mu_0'\Lambda'X'X\hat{\beta} - Y'MY + \hat{\beta}'\Lambda'X'X\hat{\beta} + Y'MY \\ &= \mu_0'\Lambda'X'X\mu_0 - 2\mu_0'\Lambda'X'X\hat{\beta} + \hat{\beta}'\Lambda'X'X\hat{\beta} \\ &= (\hat{\beta} - \mu_0)'(\Lambda'X'X)(\hat{\beta} - \mu_0). \end{aligned}$$

Thus

$$\begin{aligned} & p(\beta|y) \\ & \propto \int_0^{\infty} \tau^{\frac{n+p+\delta_0}{2}-1} \times \\ & e^{-\frac{\tau}{2} [\gamma_0 + Y'(I-M)Y + (\beta - \tilde{\beta})'(X'X + \Sigma_0^{-1})(\beta - \tilde{\beta}) + (\hat{\beta} - \mu_0)'(\Lambda'X'X)(\hat{\beta} - \mu_0)]} d\tau \\ &= \int_0^{\infty} \tau^{\frac{n+p+\delta_0}{2}-1} e^{-\frac{\tau}{2} [(n+\delta_0)\tilde{s}^2 + (\beta - \tilde{\beta})'(X'X + \Sigma_0^{-1})(\beta - \tilde{\beta})]} d\tau \\ &\propto [(n + \delta_0)\tilde{s}^2 + (\beta - \tilde{\beta})'(X'X + \Sigma_0^{-1})(\beta - \tilde{\beta})]^{-\frac{(n+\delta_0+p)}{2}} \\ &\propto \left[ 1 + \frac{1}{(n + \delta_0)\tilde{s}^2} (\beta - \tilde{\beta})'(X'X + \Sigma_0^{-1})(\beta - \tilde{\beta}) \right]^{-\frac{(n+\delta_0+p)}{2}}. \end{aligned}$$

# Bayesian Analysis of the Linear Model

Thus

$$\beta|y \sim S_p \left( n + \delta_0, \tilde{\beta}, \tilde{s}^2 (X'X + \Sigma_0^{-1})^{-1} \right) .$$

Now

$$\begin{aligned} p(\tau|y) &\propto \int_{-\infty}^{\infty} \tau^{\frac{n+p+\delta_0}{2}-1} \\ &\quad \times e^{-\frac{\tau}{2}[(n+\delta_0)\tilde{s}^2 + (\beta - \tilde{\beta})'(X'X + \Sigma_0^{-1})(\beta - \tilde{\beta})]} d\beta \\ &= \tau^{\frac{n+p+\delta_0}{2}-1} e^{-\frac{\tau}{2}[(n+\delta_0)\tilde{s}^2]} \\ &\quad \times \int_{-\infty}^{\infty} e^{-\frac{\tau}{2}[(\beta - \tilde{\beta})'(X'X + \Sigma_0^{-1})(\beta - \tilde{\beta})]} d\beta \\ &\propto \tau^{\frac{n+p+\delta_0}{2}-1} e^{-\frac{\tau}{2}[(n+\delta_0)\tilde{s}^2]} \tau^{-\frac{p}{2}} \\ &= \tau^{\frac{n+\delta_0}{2}-1} e^{-\frac{\tau}{2}[(n+\delta_0)\tilde{s}^2]}. \end{aligned}$$

Thus

$$\tau|y \sim \text{gamma} \left( \frac{n + \delta_0}{2}, \frac{(n + \delta_0)\tilde{s}^2}{2} \right)$$

# Bayesian Analysis of the Linear Model

## Theorem

Consider the linear model in (15) and (16). Let  $Z$  be a  $q \times 1$  vector of future observations taken at  $X_f$ , where  $X_f$  is  $q \times p$ . That is

$$Z = X_f\beta + \epsilon,$$

where

$$\epsilon \sim N_q(0, \sigma^2 I).$$

Suppose

$$\pi(\beta, \tau) \propto \tau^{-1}.$$

Then

$$Z|X_f, Y \sim S_q \left( n - p, X_f \hat{\beta}, s^2(I + X_f(X'X)^{-1}X_f') \right),$$

where

$$s^2 = \frac{Y'(I - M)Y}{n - p}.$$

# Bayesian Analysis of the Linear Model

## Proof

$$\begin{aligned} p(z|X_f, y) &= \int_0^\infty \int_{-\infty}^\infty p(z|\beta, \tau) p(\beta, \tau|y) d\beta d\tau \\ &\propto \int_0^\infty \int_{-\infty}^\infty \tau^{\frac{q}{2}} e^{-\frac{\tau}{2}[(z - X_f\beta)'(z - X_f\beta)]} \tau^{\frac{n}{2}-1} \\ &\quad \times e^{-\frac{\tau}{2}[(n-p)s^2 + (\beta - \hat{\beta})'X'X(\beta - \hat{\beta})]} d\beta d\tau . \end{aligned}$$

Now

$$\begin{aligned} &(z - X_f\beta)'(z - X_f\beta) \\ &= (\beta - \hat{\beta}_z)'(X_f'X_f)(\beta - \hat{\beta}_z) + z'(I - M_{X_f})z \end{aligned}$$

Where

$$\hat{\beta}_z = (X_f'X_f)^{-1}X_f'z ,$$

$$M_{X_f} = X_f(X_f'X_f)^{-1}X_f' .$$

# Bayesian Analysis of the Linear Model

Now

$$\begin{aligned} & (\beta - \hat{\beta}_z)'(X_f'X_f)(\beta - \hat{\beta}_z) + (\beta - \hat{\beta})'X'X(\beta - \hat{\beta}) \\ &= (\beta - \tilde{\beta}_f)'(X_f'X_f + X'X)(\beta - \tilde{\beta}_f) \\ &\quad - \tilde{\beta}_f'(X_f'X_f + X'X)\tilde{\beta}_f + \hat{\beta}_z'X_f'X_f\hat{\beta}_z + \hat{\beta}'X'X\hat{\beta}, \end{aligned}$$

where

$$\tilde{\beta}_f = (X_f'X_f + X'X)^{-1} \left[ (X_f'X_f)\hat{\beta}_z + (X'X)\hat{\beta} \right]$$

$$= \Lambda_f \hat{\beta}_z + (I - \Lambda_f) \hat{\beta},$$

$$\Lambda_f = (X_f'X_f + X'X)^{-1}X_f'X_f.$$

Note that  $I - \Lambda_f = (X_f'X_f + X'X)^{-1}X'X$

# Bayesian Analysis of the Linear Model

since  $I = \Lambda_f + \Lambda_f$

$$\begin{aligned} &= (X_f' X_f + X' X)^{-1} X' X + (X_f' X_f + X' X)^{-1} X_f' X_f \\ &= (X_f' X_f + X' X)^{-1} (X' X + X_f' X_f) = I . \end{aligned}$$

Now

$$\begin{aligned} &- \tilde{\beta}_f' (X_f' X_f + X' X) \tilde{\beta}_f + \hat{\beta}_z' X_f' X_f \hat{\beta}_z + \hat{\beta}' X' X \hat{\beta} \\ &= (\hat{\beta} - \hat{\beta}_z)' (\Lambda_f' X' X) (\hat{\beta} - \hat{\beta}_z) , \end{aligned}$$

as in the previous proof.

Thus

$$\begin{aligned} &p(z | X_f, y) \\ &\propto \int_0^\infty \int_{-\infty}^\infty \tau^{\frac{n+q}{2}-1} e^{-\frac{\tau}{2} [(n-p)s^2 + (\beta - \tilde{\beta}_f)' (X_f' X_f + X' X)^{-1} (\beta - \tilde{\beta}_f)]} \\ &\quad \times e^{-\frac{\tau}{2} [(\hat{\beta} - \hat{\beta}_z)' (\Lambda_f' X' X) (\hat{\beta} - \hat{\beta}_z) + z' (I - M_{X_f}) z]} d\beta d\tau . \\ &= \int_0^\infty \tau^{\frac{n+q}{2}-1} \tau^{-\frac{p}{2}} e^{-\frac{\tau}{2} [(n-p)s^2 + (\hat{\beta} - \hat{\beta}_z)' (\Lambda_f' X' X) (\hat{\beta} - \hat{\beta}_z) + z' (I - M_{X_f}) z]} d\tau . \end{aligned}$$

# Bayesian Analysis of the Linear Model

Now write

$$\begin{aligned} & (\hat{\beta} - \hat{\beta}_z)'(\Lambda_f' X' X)(\hat{\beta} - \hat{\beta}_z) + z'(I - M_{X_f})z \\ &= \hat{\beta}' \Lambda_f' X' X \hat{\beta} - 2\hat{\beta}_z' \Lambda_f' X' X \hat{\beta} + \hat{\beta}_z' \Lambda_f' X' X \hat{\beta}_z + z'(I - M_{X_f})z \\ &= \hat{\beta}' \Lambda_f' X' X \hat{\beta} - 2z' X_f (X_f' X_f)^{-1} (X_f' X_f) (X_f' X_f + X' X)^{-1} (X' X) \hat{\beta} \\ &+ z' X_f (X_f' X_f)^{-1} (X_f' X_f) (X_f' X_f + X' X)^{-1} (X' X) (X_f' X_f)^{-1} X_f' z \\ &+ z'(I - M_{X_f})z \\ \\ &= \hat{\beta}' \Lambda_f' X' X \hat{\beta} - 2z' X_f (X_f' X_f + X' X)^{-1} X' X \hat{\beta} \\ &+ z' X_f (I - \Lambda_f) (X_f' X_f)^{-1} X_f' z + z'(I - M_{X_f})z \\ \\ &= \hat{\beta}' \Lambda_f' X' X \hat{\beta} - 2z' X_f (X_f' X_f + X' X)^{-1} X' X \hat{\beta} \\ &+ z' X_f (X_f' X_f)^{-1} X_f' z - z' X_f \Lambda_f (X_f' X_f)^{-1} X_f' z \\ &+ z' z - z' X_f (X_f' X_f)^{-1} X_f' z \\ \\ &= \hat{\beta}' \Lambda_f' X' X \hat{\beta} - 2z' X_f (X_f' X_f + X' X)^{-1} X' X \hat{\beta} \\ &+ z'(I - X_f (X_f' X_f + X' X)^{-1} X_f')z . \end{aligned}$$

Note that

$$\begin{aligned} \Lambda_f (X_f' X_f)^{-1} &= (X_f' X_f + X' X)^{-1} (X_f' X_f) (X_f' X_f)^{-1} \\ &= (X_f' X_f + X' X)^{-1} . \end{aligned}$$

# Bayesian Analysis of the Linear Model

Now we need to establish 2 non-trivial identities in order to finish completing the square.

## Identity 1:

$$(I - X_f(X'_f X_f + X' X)^{-1} X'_f)^{-1} = I + X_f(X' X)^{-1} X'_f .$$

## Proof:

It suffices to show that

$$\left[ I - X_f(X' X + X'_f X_f)^{-1} X'_f \right] \left[ I + X_f(X' X)^{-1} X'_f \right] = I .$$

Let's multiply the left and show the right. Doing this, we get

$$\begin{aligned} & I - X_f(X' X + X'_f X_f)^{-1} X'_f - X_f(X' X + X'_f X_f)^{-1} X'_f X_f(X' X)^{-1} X'_f \\ & + X_f(X' X)^{-1} X'_f . \end{aligned}$$

# Bayesian Analysis of the Linear Model

Now we need to show that the sum of the latter three terms is 0.

$$\begin{aligned} & X_f(X'X)^{-1}X'_f \\ & - \left[ X_f(X'X + X'_fX_f)^{-1}X'_f + X_f(X'X + X'_fX_f)^{-1}X'_fX_f(X'X)^{-1}X'_f \right] \\ & = X_f(X'X)^{-1}X'_f \\ & - X_f \left[ (X'X + X'_fX_f)^{-1} + (X'X + X'_fX_f)^{-1}X'_fX_f(X'X)^{-1} \right] X'_f . \end{aligned}$$

Now it suffices to show that

$$(X'X + X'_fX_f)^{-1} + (X'X + X'_fX_f)^{-1}X'_fX_f(X'X)^{-1} = (X'X)^{-1} .$$

Multiplying both sides by  $X'X + X'_fX_f$ , we get

$$\begin{aligned} I + X'_fX_f(X'X)^{-1} &= (X'X + X'_fX_f)(X'X)^{-1} \\ &= I + (X'_fX_f)(X'X)^{-1} . \end{aligned}$$

This proves Identity 1.

# Bayesian Analysis of the Linear Model

## Identity 2:

$$(I + X_f(X'X)^{-1}X'_f)X_f(X'_fX_f + X'X)^{-1}X'X = X_f .$$

### Proof:

Taking  $[(X'_fX_f + X'X)^{-1}X'X]^{-1}$  of both sides, we get

$$\begin{aligned}(I + X_f(X'X)^{-1}X'_f)X_f &= X_f(X'X)^{-1}(X'_fX_f + X'X) \\ &= X_f(X'X)^{-1}X'_fX_f + X_f .\end{aligned}$$

The left hand side is

$$(I + X_f(X'X)^{-1}X'_f)X_f = X_f + X_f(X'X)^{-1}X'_fX_f .$$

Thus identity 2 is established.

# Bayesian Analysis of the Linear Model

Now we have

$$\begin{aligned} & \hat{\beta}' \Lambda_f' X' X \hat{\beta} - 2z' X_f (X_f' X_f + X' X)^{-1} X' X \hat{\beta} \\ & + z' (I - X_f (X_f' X_f + X' X)^{-1} X_f') z \\ & = \hat{\beta}' \Lambda_f' X' X \hat{\beta} + (z - \mu_z)' (I - X_f (X_f' X_f + X' X)^{-1} X_f') (z - \mu_z) \\ & - \mu_z' (I - X_f (X_f' X_f + X' X)^{-1} X_f') \mu_z , \end{aligned}$$

where

$$\begin{aligned} \mu_z &= (I - X_f (X_f' X_f + X' X)^{-1} X_f')^{-1} (X_f (X_f' X_f + X' X)^{-1} X' X \hat{\beta}) \\ &= (I + X_f (X' X)^{-1} X_f') (X_f (X_f' X_f + X' X)^{-1} X' X \hat{\beta}) \quad (\text{Identity 1}) \\ &= X_f \hat{\beta} \quad (\text{Identity 2}) \end{aligned}$$

Thus  $\mu_z = X_f \hat{\beta}$ , and we have

$$\begin{aligned} & \hat{\beta}' \Lambda_f' X' X \hat{\beta} + (z - X_f \hat{\beta})' (I - X_f (X_f' X_f + X' X)^{-1} X_f') (z - X_f \hat{\beta}) \\ & - \hat{\beta}' X_f' (I - X_f (X_f' X_f + X' X)^{-1} X_f') X_f \hat{\beta} . \end{aligned}$$

# Bayesian Analysis of the Linear Model

Now

$$\begin{aligned}& \hat{\beta}' \Lambda_f' X' X \hat{\beta} - \hat{\beta}' X_f' (I - X_f (X_f' X_f + X' X)^{-1} X_f') X_f \hat{\beta} \\&= \hat{\beta}' \Lambda_f' X' X \hat{\beta} - \hat{\beta}' X_f' X_f \hat{\beta} + \hat{\beta}' X_f' X_f \Lambda_f \hat{\beta} \\&= \hat{\beta}' [\Lambda_f' X' X - X_f' X_f + X_f' X_f \Lambda_f] \hat{\beta}.\end{aligned}$$

Now observe that

$$\begin{aligned}& \Lambda_f' X' X - X_f' X_f + X_f' X_f \Lambda_f \\&= X_f' X_f (X_f' X_f + X' X)^{-1} X' X - X_f' X_f \\&\quad + X_f' X_f (X_f' X_f + X' X)^{-1} X_f' X_f.\end{aligned}$$

# Bayesian Analysis of the Linear Model

**Claim:**

$$X_f' X_f (X_f' X_f + X' X)^{-1} X_f' X_f + X_f' X_f (X_f' X_f + X' X)^{-1} X' X = X_f' X_f .$$

To see this, left multiply both sides by  $(X_f' X_f)^{-1}$ , which yields

$$(X_f' X_f + X' X)^{-1} X' X + (X_f' X_f + X' X)^{-1} X_f' X_f = I .$$

Now multiply both sides by  $X_f' X_f + X' X$ , which yields

$$X' X + X_f' X_f = X_f' X_f + X' X .$$

Thus

$$\Lambda_f' X' X - X_f' X_f + X_f' X_f \Lambda_f = 0 .$$

# Bayesian Analysis of the Linear Model

Finally, we have

$$\begin{aligned} p(z|X_f, y) &\propto \int_0^{\infty} \tau^{\frac{n+q-p}{2}-1} \\ &\times e^{-\frac{\tau}{2} \left[ (n-p)s^2 + (z - X_f \hat{\beta})' \left( I - X_f (X_f' X_f + X' X)^{-1} X_f' \right) (z - X_f \hat{\beta}) \right]} d\tau \\ &\propto \left[ (n-p)s^2 + (z - X_f \hat{\beta})' \left( I - X_f (X_f' X_f + X' X)^{-1} X_f' \right) (z - X_f \hat{\beta}) \right]^{-\frac{n-p+q}{2}} \\ &\propto \left[ 1 + \frac{1}{(n-p)s^2} (z - X_f \hat{\beta})' \left( I + X_f (X' X)^{-1} X_f' \right)^{-1} (z - X_f \hat{\beta}) \right]^{-\frac{n-p+q}{2}}. \end{aligned}$$

Thus

$$Z|X_f, y \sim S_q \left( n-p, X_f \hat{\beta}, s^2 (I + X_f (X' X)^{-1} X_f') \right).$$

# Bayesian Analysis of the Linear Model

## Theorem

Consider the linear model in (15) and (16) and suppose

$$\beta | \tau \sim N_p(\mu_0, \tau^{-1} \Sigma_0)$$

$$\tau \sim \text{gamma}\left(\frac{\delta_0}{2}, \frac{\gamma_0}{2}\right).$$

Let  $Z$  be a  $q \times 1$  future vector of observations taken at  $X_f$ , with

$$Z = X_f \beta + \epsilon_f, \quad \epsilon_f \sim N_q(0, \sigma^2 I).$$

Then

$$Z | X_f, y \sim S_q \left( n + \delta_0, X_f \tilde{\beta}, \tilde{s}^2 (I + X_f (\Sigma_0^{-1} + X' X)^{-1} X_f') \right),$$

where

$$\tilde{s}^2 = (n + \delta_0)^{-1} \left[ Y'(I - M)Y + (\hat{\beta} - \mu_0)' (\Lambda' X' X) (\hat{\beta} - \mu_0) + \gamma_0 \right],$$

$$\tilde{\beta} = \Lambda \mu_0 + (I - \Lambda) \hat{\beta},$$

$$\Lambda = (X' X + \Sigma_0^{-1})^{-1} \Sigma_0^{-1}.$$

# Bayesian Analysis of the Linear Model

Proof: We have

$$p(z|X_f, y)$$

$$\begin{aligned} &\propto \int_0^\infty \int_{-\infty}^\infty \tau^{\frac{n+p+q+\delta_0}{2}-1} e^{-\frac{\tau}{2}[\gamma_0+(n-p)s^2]} \\ &\quad \times e^{-\frac{\tau}{2}[(\beta-\hat{\beta})'(X'X)(\beta-\hat{\beta})+(\beta-\mu_0)'\Sigma_0^{-1}(\beta-\mu_0)]} \\ &\quad \times e^{-\frac{\tau}{2}[z-X_f\beta]'[z-X_f\beta]} d\beta d\tau \end{aligned}$$

$$\begin{aligned} &\propto \int_0^\infty \int_{-\infty}^\infty \tau^{\frac{n+p+q+\delta_0}{2}-1} e^{-\frac{\tau}{2}[\gamma_0+(n-p)s^2]} \\ &\quad \times e^{-\frac{\tau}{2}[(\beta-\tilde{\beta})'(X'X+\Sigma_0^{-1})(\beta-\tilde{\beta})+(\hat{\beta}-\mu_0)'(\Lambda'X'X)(\hat{\beta}-\mu_0)]} \\ &\quad \times e^{-\frac{\tau}{2}[z-X_f\beta]'[z-X_f\beta]} d\beta d\tau \end{aligned}$$

$$\begin{aligned} &= \int_0^\infty \int_{-\infty}^\infty \tau^{\frac{n+p+q+\delta_0}{2}-1} e^{-\frac{\tau}{2}[(n+\delta_0)\tilde{s}^2]} \\ &\quad \times e^{-\frac{\tau}{2}[(\beta-\tilde{\beta})'(X'X+\Sigma_0^{-1})(\beta-\tilde{\beta})]} \\ &\quad \times e^{-\frac{\tau}{2}[z-X_f\beta]'[z-X_f\beta]} d\beta d\tau \end{aligned}$$

$$\begin{aligned} &= \int_0^\infty \int_{-\infty}^\infty \tau^{\frac{n+p+q+\delta_0}{2}-1} e^{-\frac{\tau}{2}[(n+\delta_0)\tilde{s}^2]} \\ &\quad \times e^{-\frac{\tau}{2}[(\beta-\tilde{\beta})'(X'X+\Sigma_0^{-1})(\beta-\tilde{\beta})]} \\ &\quad \times e^{-\frac{\tau}{2}[(\beta-\hat{\beta}_z)'(X_f'X_f)(\beta-\hat{\beta}_z)+z'(I-M_{X_f})z]} d\beta d\tau . \end{aligned}$$

# Bayesian Analysis of the Linear Model

Now

$$\begin{aligned} & (\beta - \tilde{\beta})' (X'X + \Sigma_0^{-1})(\beta - \tilde{\beta}) \\ & + (\beta - \hat{\beta}_z)' (X_f'X_f)(\beta - \hat{\beta}_z) + z'(I - M_{x_f})z \\ = & (\tilde{\beta} - \hat{\beta}_z)' (\tilde{\Lambda}'X'X)(\tilde{\beta} - \hat{\beta}_z) + z'(I - M_{x_f})z \\ & + (\beta - \tilde{\beta})' \left( X_f'X_f + X'X + \Sigma_0^{-1} \right) (\beta - \tilde{\beta}), \end{aligned}$$

where

$$\tilde{\Lambda} = \left( X_f'X_f + X'X + \Sigma_0^{-1} \right) X_f'X_f.$$

This is clear from the previous proof.

Thus, we have

$$\begin{aligned} & \int_0^\infty \int_{-\infty}^\infty \tau^{\frac{n+p+q+\delta_0}{2}-1} e^{-\frac{\tau}{2}[(n+\delta_0)\tilde{s}^2]} \\ & \times e^{-\frac{\tau}{2}[(\beta - \tilde{\beta})' (X_f'X_f + X'X + \Sigma_0^{-1})(\beta - \tilde{\beta})]} \\ & \times e^{-\frac{\tau}{2}[(\tilde{\beta} - \hat{\beta}_z)' (\tilde{\Lambda}'X'X)(\tilde{\beta} - \hat{\beta}_z) + z'(I - M_{x_f})z]} d\beta d\tau \\ = & \int_0^\infty \tau^{\frac{n+q+\delta_0}{2}-1} e^{-\frac{\tau}{2}[(n+\delta_0)\tilde{s}^2]} \\ & \times e^{-\frac{\tau}{2}[(\tilde{\beta} - \hat{\beta}_z)' (\tilde{\Lambda}'X'X)(\tilde{\beta} - \hat{\beta}_z) + z'(I - M_{x_f})z]} d\tau. \end{aligned}$$

# Bayesian Analysis of the Linear Model

Following the earlier proof, we have

$$\begin{aligned} & (\tilde{\beta} - \hat{\beta}_z)' (\tilde{\lambda}' X' X) (\tilde{\beta} - \hat{\beta}_z) + z' (I - M_{X_f}) z \\ = & (z - X_f \tilde{\beta})' \left( I - X_f (X_f' X_f + X' X + \Sigma_0^{-1})^{-1} X_f' \right) (z - X_f \tilde{\beta}) \\ = & (z - X_f \tilde{\beta})' \left( I + X_f (X' X + \Sigma_0^{-1})^{-1} X_f' \right)^{-1} (z - X_f \tilde{\beta}). \end{aligned}$$

Thus

$$\begin{aligned} & p(z | X_f, y) \\ \propto & \int_0^\infty \tau^{\frac{n+q+\delta_0}{2}-1} e^{-\frac{\tau}{2} [(n+\delta_0)\tilde{s}^2]} \\ & \times e^{-\frac{\tau}{2} \left[ (z - X_f \tilde{\beta})' \left( I + X_f (X' X + \Sigma_0^{-1})^{-1} X_f' \right)^{-1} (z - X_f \tilde{\beta}) \right]} d\tau \\ \propto & \left[ (n + \delta_0)\tilde{s}^2 + (z - X_f \tilde{\beta})' \left( I + X_f (X' X + \Sigma_0^{-1})^{-1} X_f' \right)^{-1} (z - X_f \tilde{\beta}) \right]^{-\frac{m}{2}} \\ \propto & \left[ 1 + \frac{1}{(n + \delta_0)\tilde{s}^2} (z - X_f \tilde{\beta})' \left( I + X_f (X' X + \Sigma_0^{-1})^{-1} X_f' \right)^{-1} (z - X_f \tilde{\beta}) \right]^{-\frac{m}{2}}, \end{aligned}$$

where  $m = n + \delta_0 + q$ . Thus

$$z | X_f, y \sim S_q \left( n + \delta_0, X_f \tilde{\beta}, \tilde{s}^2 \left( I + X_f (X' X + \Sigma_0^{-1})^{-1} X_f' \right) \right).$$

# Bayesian Analysis of the Linear Model

Note that the noninformative case can be obtained from the informative case by formally setting

$$\delta_0 = -p, \quad \Sigma_0^{-1} = 0, \quad \gamma_0 = 0.$$

## Square completion in $n$ dimensions

Suppose  $x = (x_1, \dots, x_n)'$  is a  $n \times 1$  vector,  $A$  is a  $n \times n$  nonsingular matrix, and  $b$  is an  $n \times 1$  vector. Then

$$x'Ax + b'x = \left(x + \frac{A^{-1}b}{2}\right)'A\left(x + \frac{A^{-1}b}{2}\right) - \frac{b'A^{-1}b}{4}$$

This is the multivariate analog of the one dimensional square completion given earlier. This result is very useful in computing multivariate normal integrals.

## Combining quadratic forms

Suppose  $x$  is an  $n \times 1$  vector,  $\mu_1$ , and  $\mu_2$  are  $n \times 1$  vectors,  $A_1, A_2$ , are  $n \times n$  matrices such that  $A_1 + A_2$  is nonsingular. Then

$$\begin{aligned} & (x - \mu_1)'A_1(x - \mu_1) + (x - \mu_2)'A_2(x - \mu_2) \\ &= (x - \mu^*)'(A_1 + A_2)(x - \mu^*) + \mu_1'A_1\mu_1 + \mu_2'A_2\mu_2 - \mu^{*'}(A_1 + A_2)\mu^*, \end{aligned}$$

# Bayesian Analysis of the Linear Model

where

$$\mu^* = (A_1 + A_2)^{-1}(A_1\mu_1 + A_2\mu_2).$$

We can generalize this result to combining  $m$  quadratic forms, We have

$$\begin{aligned} & (x - \mu_1)'A_1(x - \mu_1) + \cdots + (x - \mu_m)'A_m(x - \mu_m) \\ &= (x - \mu^*)'B(x - \mu^*) + \sum_{i=1}^m \mu_i' A_i \mu_i - \mu^{*\prime} B \mu^*, \end{aligned}$$

where  $B = \sum_{i=1}^m A_i$  and  $\mu^* = B^{-1}(\sum_{i=1}^m A_i \mu_i)$ .

## ① Exponential family

- Definition
- Examples
- Properties

## ② Likelihood theory

- Bartlett identities
- Maximum likelihood estimation
- Central limit theorems
- Estimation theory: consistency and asymptotic normality
- Hypothesis testing
- The delta method

## Definition 2.1

If the density of a random variable  $Y$  with respect to a  $\sigma$ -finite measure  $\Lambda(y)$  has the form

$$p(y|\xi) = \exp \left\{ \phi[y\theta - b(\theta) - c(y)] - \frac{1}{2}s(y, \phi) \right\}, \quad (2.1)$$

where  $\xi = (\theta, \phi)$ ,  $c(\cdot)$ ,  $b(\cdot)$ , and  $s(\cdot, \cdot)$  are some known functions, then  $Y$  or its distribution belongs to an **exponential family**, denoted by  $Y \sim D(\theta, \phi)$ . In addition,  $\theta$  is the **natural parameter** of the exponential family and  $\phi$  is the dispersion parameter. Moreover, if  $\phi$  is known, then (2.1) is a **linear exponential family**; however, if  $\phi$  is unknown, then (2.1) is called **exponential dispersion model** to emphasize the role of  $\phi$ .

## Exponential family: examples

Many commonly used univariate distributions belong to the exponential family. For instance, the **normal, gamma, chi-square, beta, Dirichlet, Bernoulli, binomial, multinomial, Poisson, negative binomial, and geometric distributions** are all exponential families. However, the **Weibull, Cauchy, and uniform** distributions are not.

# Exponential family: examples

## Example 2.1

Let  $Y$  be a normally distributed random variable with mean  $\mu$  and variance  $\sigma^2$ . Then the density of  $y$  is given by

$$\begin{aligned} p(y|\mu, \sigma) &= \exp\{-0.5(y - \mu)^2/\sigma^2 - \log(\sqrt{2\pi}\sigma)\} \\ &= \exp\left\{(y\mu - \mu^2/2 - y^2/2)/\sigma^2 - \frac{1}{2}\log(2\pi\sigma^2)\right\}. \end{aligned}$$

Thus,  $\theta = \mu$ ,  $\phi = 1/\sigma^2$ ,  $b(\theta) = \mu^2/2$ ,  $c(y) = y^2/2$ , and  $s(y, \phi) = \log(2\pi\sigma^2)$ .

## Example 2.2

Let  $Y$  be a Poisson random variable with mean  $\lambda$ , denoted by  $y \sim \text{Poisson}(\lambda)$ , whose distribution can be written as

$$P(y) = \frac{\lambda^y e^{-\lambda}}{y!} = \exp(y \log \lambda - \lambda)/y!.$$

Thus,  $\phi = 1$ ,  $\theta = \log \lambda$ ,  $b(\theta) = \lambda = \exp(\theta)$ ,  $c(y) = \log(y!)$ , and  $s(y, \phi) = 0$ .

# Exponential family: examples

## Example 2.3

Let  $Y$  be a Binomial  $(M, p)$  random variable with

$$P(y) = \exp \left\{ M \log(1 - p) + y \log(p/(1 - p)) + \log \binom{M}{y} \right\}$$

where  $\theta = \log(p/(1 - p))$ ,  $\binom{M}{y} = M!/[y!(M - y)!]$ ,

$$b(\theta) = M \log(1 + \exp(\theta)), \quad s(y, \phi) = 0, \quad \text{and} \quad c(y) = -\log \binom{M}{y}.$$

## Exponential family: properties

Because  $\int p(y|\xi)d\Lambda(y) = 1$ , for a given  $\phi$ , we define

$$\Theta = \left\{ \theta : 0 < \int \exp\{\phi y\theta - d(y, \phi)\}d\Lambda(y) = \exp(\phi b(\theta)) < \infty \right\}, \quad (2.2)$$

which is a convex set, where  $d(y, \phi) = \phi c(y) + 0.5s(y, \phi)$ .

Moreover, the *potential* function  $b(\theta)$  is a convex function in  $\theta$  and, in the interior of  $\Theta$ , all derivatives of  $b(\theta)$  and all moments of  $Y$  exist.

## Theorem 2.1

*The potential function  $b(\theta)$  is a convex function in  $\theta$ .*

Proof of Theorem 2.1. For any  $\theta \in \Theta$ , we have

$b(\theta) = c \log(\int \exp\{\phi y \theta - d(y, \phi)\} d\Lambda(y))$ . Let  $\alpha_1 \in [0, 1]$ , we get

$$\begin{aligned} & b(\alpha_1 \theta_1 + (1 - \alpha_1) \theta_2) \\ = & c \log\left(\int \exp\{\phi y [\alpha_1 \theta_1 + (1 - \alpha_1) \theta_2] - d(y, \phi)\} d\Lambda(y)\right) \\ \leq & \alpha_1 c \log\left(\int \exp\{\phi y \theta_1 - d(y, \phi)\} d\Lambda(y)\right) \\ & + (1 - \alpha_1) c \log\left(\int \exp\{\phi y \theta_2 - d(y, \phi)\} d\Lambda(y)\right) \\ = & \alpha_1 b(\theta_1) + (1 - \alpha_1) b(\theta_2). \end{aligned}$$

# Exponential family: properties

## Definition 2.2

*The moment generating function of  $Y$  is defined as follows:*

$$M_Y(t) = E[\exp(tY)] = \int \exp(tx) dF_Y(x),$$

*where  $F_Y(\cdot)$  is the distribution function of  $Y$ . Moreover, the cumulant generating function  $K_Y(t)$  is given by*

$$K_Y(t) = \log\{M_Y(t)\}.$$

## Exponential family: properties

$M_Y(t)$  is associated with the moments of  $Y$ . If  $M_Y(t)$  exists in an interval around  $t = 0$ , then the  $j$ -th moment is given by

$$m_j(Y) = m_j(Y) = E(Y^j) = \frac{d^j}{dt^j} M_Y(t) \Big|_{t=0}.$$

Moreover, because  $\exp(tx) = 1 + tx + \frac{(tx)^2}{2!} + \dots$ , we have

$$M_Y(t) = 1 + tm_1 + \frac{t^2 m_2}{2!} + \dots$$

## Exponential family: properties

The *cumulants*  $k_j(Y) = k_j$  are defined by

$$K_Y(t) = \log M_Y(t) = \sum_{j=1}^{\infty} \frac{k_j t^j}{j!} = \mu t + \frac{\sigma^2 t^2}{2} + \dots, \quad (2.3)$$

where  $\mu = E(Y)$  and  $\sigma^2 = \text{var}(Y)$ . Equivalently, we have

$$k_j(Y) = \left. \frac{d^j}{dt^j} K_Y(t) \right|_{t=0}. \quad (2.4)$$

Furthermore, the relationship between  $m_j$  and  $k_j$  is determined by

$$1 + \sum_{j=1}^{\infty} \frac{m_j t^j}{j!} = \exp \left( \sum_{j=1}^{\infty} \frac{k_j t^j}{j!} \right).$$

## Exponential family: properties

Based on the above equation, we can get a recursion formula as follows:

$$k_j = m_j - \sum_{i=1}^{j-1} \binom{j-1}{i-1} k_i m_{j-i}, \quad (2.5)$$

where  $\binom{j}{i} = j!/(i!(j-i)!)$ . In particular, we have

$$\begin{aligned} m_1 &= k_1, \quad m_2 = k_2 + k_1^2, \quad m_3 = k_3 + 3k_2k_1 + k_1^3, \\ \text{and } m_4 &= k_4 + 4k_3k_1 + 3k_2^2 + 6k_2k_1^2 + k_1^4. \end{aligned}$$

## Exponential family: properties

An important question is why we need to consider  $K_Y(t)$  and the cumulants  $k_j$ . An important reason is due to the fact that cumulants satisfy homogeneity and additivity as follows:  
 $k_j(cY) = c^j k_j(Y)$  and  $k_j(X + Y) = k_j(X) + k_j(Y)$  when  $X$  and  $Y$  are independent.

$$K_{cY}(t) = \log M_{cY}(t) = \log E[\exp(ctY)] = K_Y(ct)$$

and

$$K_{X+Y}(t) = \log M_{X+Y}(t) = \log\{M_X(t)M_Y(t)\} = K_X(t) + K_Y(t).$$

## Exponential family: properties

Let us consider  $n^{-1/2} \sum_{i=1}^n Y_i$  and  $Y_1, \dots, Y_n$  are independently and identically distributed with mean zero. Then, we have

$$k_j(n^{-1/2} \sum_{i=1}^n Y_i) = n^{-j/2} k_j(\sum_{i=1}^n Y_i) = n^{-j/2+1} k_j(Y_1), \text{ for } j = 1, \dots . \quad (2.6)$$

Thus,  $k_1(n^{-1/2} \sum_{i=1}^n Y_i) = n^{1/2} k_1(Y_1)$  and  $k_2(n^{-1/2} \sum_{i=1}^n Y_i) = k_2(Y_1)$ , whereas  $k_j(n^{-1/2} \sum_{i=1}^n Y_i)$  converges to zero for  $j > 2$ . Then the cumulants of  $n^{-1/2} \sum_{i=1}^n Y_i$  converge to those of a normal random variable with mean zero and variance  $\text{var}(Y_1)$ .

# Exponential family: properties

## Theorem 2.2

For the exponential family defined in Definition (2.1),

$$M_Y(t) = \exp\{\phi[b(\theta + t/\phi) - b(\theta)]\} \text{ and } K_Y(t) = \phi[b(\theta + t/\phi) - b(\theta)]. \quad (2.7)$$

Proof of Theorem 2.2.

$$\begin{aligned} M_Y(t) &= \int \exp(tx) \exp\{\phi[x\theta - b(\theta) - c(x)] - \frac{1}{2}s(x, \phi)\} d\Lambda(x) \\ &= \int \exp\{\phi[x(\theta + t/\phi) - b(\theta) - c(x)] - \frac{1}{2}s(x, \phi)\} d\Lambda(x) \\ &= \int \exp\{\phi[x(\theta + t/\phi) - b(\theta + t/\phi) - c(x)] - \frac{1}{2}s(x, \phi)\} \\ &\quad \exp\{\phi[b(\theta + t/\phi) - b(\theta)]\} d\Lambda(x) \\ &= \exp\{\phi[b(\theta + t/\phi) - b(\theta)]\} \times 1. \end{aligned}$$

## Exponential family: properties

Based on  $M_Y(t)$ , we can calculate all moments of  $Y$ . Moreover, differentiating  $K_Y(t)$  with respect to  $t$ , we can easily obtain

$$k_j(Y) = \phi^{1-j} \partial_\theta^j b(\theta) \quad \text{for } j = 1, \dots, \quad (2.8)$$

where  $\partial_\theta^j$  denotes the  $j$ th derivative with respect to  $\theta$ . In particular,

$$E(Y) = \mu = \partial_\theta b(\theta) \quad \text{and} \quad \text{Var}(Y) = \phi^{-1} \partial_\theta^2 b(\theta). \quad (2.9)$$

## Exponential family: properties

We can also derive the moment generating function for the “random error”  $e = Y - \mu$  to get the central moments of  $y$ . Specifically, the moment generating function of  $e = Y - \mu$  is given by

$$E[\exp(te)] = \exp[\phi\{b(\theta + t/\phi) - b(\theta)\} - t\mu].$$

After some algebraic calculations, we get

$$E(e) = 0, E(e^2) = \phi^{-1}\partial_\theta^2 b(\theta), E(e^3) = \phi^{-2}\partial_\theta^3 b(\theta), \quad (2.10)$$

and  $E(e^4) = 3\phi^{-2}[\partial_\theta^2 b(\theta)]^2 + \phi^{-3}\partial_\theta^4 b(\theta)$ .

## Exponential family: properties

If  $\partial_\theta^2 b(\theta) = \phi \text{Var}(Y)$  is positive, then there is a one-to-one transformation between  $\mu = \mu(\theta)$  and  $\theta = \theta(\mu)$ . Thus, we can regard  $\mu$  in (2.9) as a new parameterization (coordinate system) of the exponential family in (2.1), and then the exponential family in (2.1) can be represented as

$$p(y|\mu, \phi) = \exp\{\phi[y\theta(\mu) - b(\theta(\mu)) - c(y)] - 0.5s(y, \phi)\},$$

in which  $\mu$  is the *expectation parameter*. It follows from equation (2.9) that

$$\partial_\theta \mu = \partial_\theta^2 b(\theta) \quad \text{and} \quad \partial_\mu \theta = \{\partial_\theta^2 b(\theta)\}^{-1}. \quad (2.11)$$

Geometrically, these two parameterizations (coordinate systems) are conjugate.

## Example 2.4

We consider a continuation of Example 2.1. By using (2.9), we get

$$\mu = \partial_\theta b(\theta) = \theta \quad \text{and} \quad \text{var}(Y) = \phi^{-1}.$$

## Example 2.5

We consider a continuation of Example 2.2. We can obtain  $k_j(Y) = \exp(\theta)$  for all  $j \geq 1$ . In particular,  $\mu = \exp(\theta)$  and  $\theta = \log(\mu)$ .

# Exponential family: examples

- Suppose the density of a random variable  $Y$  can be written as

$$p(y|\xi) = \exp\{\phi[q(y)t(\theta) - b(\theta) - c(y)] - \frac{1}{2}s(y, \phi)\}, \quad (2.12)$$

where  $q(y)$  is a known function of  $y$  and  $t(\theta)$  is a known function of  $\theta$ . Does  $p(y|\xi)$  belong to the exponential family with a dispersion parameter?

- Consider

$$p(y|\xi) = \exp \left\{ \phi[y\theta - b(\theta) - c(y)] - \frac{1}{2}s(y, \phi) \right\}. \quad (2.13)$$

Are the expressions of  $\phi$ ,  $\theta$ ,  $b(\theta)$ ,  $c(y)$  and  $s(y, \phi)$  unique?

## Exponential family: examples

Many multivariate distributions belong to exponential families, because, in general, exponential families can be defined for multi-dimensional  $\theta$  and  $\mathbf{y}$ . A multivariate exponential family (MEF) is defined as

$$p(\mathbf{y}|\xi) = \exp\{Q(\mathbf{y})^T T(\theta) - b(\theta) - c(\mathbf{y})\},$$

where  $\xi = \theta$ ,  $T(\theta) = (t_1(\theta), \dots, t_k(\theta))^T$  and  $Q(\mathbf{y}) = (q_1(\mathbf{y}), \dots, q_k(\mathbf{y}))^T$  contains some specific functions of  $\mathbf{y}$ . Moreover, if all components of  $T(\theta)$  and all components of  $Q(\mathbf{y})$  are **linearly independent**, then the exponential family is said to be of **full rank**.

# Exponential family: properties

## Example 2.6

Consider a multinomial random variable  $Y \sim M(1, (\pi_1, \dots, \pi_{I-1}))$ , where  $y$  takes values  $1, \dots, I$ ,  $\pi_i \geq 0$  and  $\sum_{i=1}^{I-1} \pi_i \leq 1$ . Then,

$$p(y|\pi_1, \dots, \pi_{I-1}) = \pi_1^{\mathbf{1}(y=1)} \cdots \pi_{I-1}^{\mathbf{1}(y=I-1)} [1 - \pi_1 - \cdots - \pi_{I-1}]^{\mathbf{1}(y=I)}.$$

Because  $\mathbf{1}(y = I) + \mathbf{1}(y = I - 1) + \cdots + \mathbf{1}(y = 1) = 1$ , we have

$$\begin{aligned} & p(y|\pi_1, \dots, \pi_{I-1}) \\ &= \exp \left\{ \sum_{k=1}^{I-1} \mathbf{1}(y = k) \log(\pi_k / \pi_I) + \log(1 - \pi_1 - \cdots - \pi_{I-1}) \right\} \end{aligned}$$

This distribution belongs to the MEF with  $\theta = (\log(\pi_1/\pi_I), \dots, \log(\pi_{I-1}/\pi_I))$ ,  $Q(\mathbf{y}) = (\mathbf{1}(y = 1), \dots, \mathbf{1}(y = I-1))^T$  and  $b(\theta) = -\log(1 - \pi_1 - \cdots - \pi_{I-1})$ .

# Exponential family: properties

## Example 2.7

Consider an Extended Bernoulli (EB) random variable, denoted  $Y \sim EB(\pi, \theta)$ , where  $y$  takes values 0, 1, 2, where  $\pi \in (0, 1)$  and  $\theta > 0$ . Specifically,

$$Pr(y) = \binom{2}{y} \pi^y (1 - \pi)^{2-y} \theta^{-y(2-y)} / f(\pi, \theta),$$

where

$$f(\pi, \theta) = \sum_{k=0}^2 \binom{2}{k} \pi^k (1 - \pi)^{2-k} \theta^{-k(2-k)}.$$

It can be shown that this distribution is in the MEF.

## Elementary Likelihood Theory: Bartlett identities

Let  $p(\mathbf{u}, \xi)$  be the joint probability density function of  $\mathbf{U} = (U_1, \dots, U_n)$ , and therefore

$$\int p(\mathbf{u}, \xi) d\Lambda(\mathbf{u}) \equiv 1, \quad (2.14)$$

where  $\xi$  contains all the unknown parameters. By differentiating (2.14) with respect to  $\xi \in \Xi$ , a subset of  $R^q$ , we are led to the following theorem.

# Elementary Likelihood Theory: Bartlett identities

## Theorem 2.3

Suppose that differentiation and integration are exchangeable and all the necessary expectations are finite. We have the following results:

$$E_\xi(\partial_j \ell_n) = 0, \quad (2.15)$$

$$E_\xi(\partial_{j,k}^2 \ell_n) + E_\xi(\partial_j \ell_n \partial_k \ell_n) = 0, \quad (2.16)$$

$$\begin{aligned} E_\xi(\partial_{j,k,l}^3 \ell_n) + \text{cov}_\xi(\partial_{j,k}^2 \ell_n, \partial_l \ell_n) + \text{cov}_\xi(\partial_{j,l}^2 \ell_n, \partial_k \ell_n) \\ + \text{cov}_\xi(\partial_{l,k}^2 \ell, \partial_j \ell_n) + E_\xi(\partial_i \ell_n \partial_j \ell_n \partial_k \ell_n) = 0, \end{aligned} \quad (2.17)$$

where  $j, k, l = 1, \dots, q$ ,  $\partial$  denotes partial differentiation (e.g.,  $\partial_j = \partial_{\xi_j}$ ,  $\partial_{j,k,l}^3 = \partial_{\xi_j \xi_k \xi_l}^3$ ), and  $\ell_n = \ell_n(\xi) = \log p(\mathbf{u}, \xi)$  is the log-likelihood function.

Each of equations (2.15)-(2.17) plays an important role in statistics. Equation (2.15), called the **score equation**, is the basic equation for defining the maximum likelihood (ML) estimate  $\hat{\xi}$ . Equation (2.16) gives

$$I_n(\xi) = E_\xi(\partial_\xi \ell_n^{\otimes 2}) = -E_\xi(\partial_\xi^2 \ell_n),$$

where  $\mathbf{a}^\otimes = \mathbf{a}\mathbf{a}^T$  for any vector  $\mathbf{a}$  and  $I_n(\xi)$  is the Fisher information matrix of  $\mathbf{U}$  in the  $\xi$  parametrization. The Fisher information matrix is associated with the asymptotic covariance of the ML estimate. Equation (2.17) is called the Bartlett identity of order 2.

# Elementary Likelihood Theory: Bartlett identities

**Proof.** Differentiating equation (2.14) with respect to  $\xi_j$ , we get

$$0 = \int \partial_j p(\mathbf{u}, \xi) d\Lambda(\mathbf{u}) = \int \{\partial_j \log p(\mathbf{u}, \xi)\} p(\mathbf{u}, \xi) d\Lambda(\mathbf{u}), \quad (2.18)$$

which leads to equation (2.15). Differentiating equation (2.18) with respect to  $\xi_k$  leads to

$$\begin{aligned} 0 &= \partial_k \int \{\partial_j \log p(\mathbf{u}, \xi)\} p(\mathbf{u}, \xi) d\Lambda(\mathbf{u}) \\ &= \int \{\partial_k \partial_j \log p(\mathbf{u}, \xi)\} p(\mathbf{u}, \xi) d\Lambda(\mathbf{u}) + \int \{\partial_j \log p(\mathbf{u}, \xi)\} \partial_k p(\mathbf{u}, \xi) d\Lambda(\mathbf{u}) \\ &= E_\xi(\partial_{j,k}^2 \ell_n) + E_\xi(\partial_j \ell_n \partial_k \ell_n), \end{aligned}$$

which yields equation (2.16). Similarly, we can prove equation (2.17).

# Elementary Likelihood Theory: Bartlett identities

## Corollary 2.1

Suppose that  $Y_1, \dots, Y_n$  are independently and identically distributed according to the exponential family in (2.1). We have the following results:

$$\sum_{i=1}^n E(Y_i) = n\partial_\theta b(\theta),$$

$$\sum_{i=1}^n \text{var}(Y_i) = n\phi^{-1}\partial_\theta^2 b(\theta),$$

and

$$\sum_{i=1}^n E\{[Y_i - \partial_\theta b(\theta)]^3\} = n\phi^{-2}\partial_\theta^3 b(\theta). \quad (2.19)$$

# Elementary Likelihood Theory: Bartlett identities

**Proof of Corollary 2.1.** The joint probability density function is given by

$$\exp \left\{ \phi \left[ \sum_{i=1}^n y_i \theta - nb(\theta) - \sum_{i=1}^n c(y_i) \right] - \frac{1}{2} \sum_{i=1}^n s(y_i, \phi) \right\}.$$

Thus,  $\ell_n(\xi) = \phi[\sum_{i=1}^n y_i \theta - nb(\theta) - \sum_{i=1}^n c(y_i)] - 0.5 \sum_{i=1}^n s(y_i, \phi)$ .

Differentiating  $\ell_n(\xi)$  with respect to  $\theta$  gives

$$\partial_\theta \ell_n(\xi) = \phi \left[ \sum_{i=1}^n y_i - n \partial_\theta b(\theta) \right], \quad \partial_\theta^2 \ell_n(\xi) = -n \phi \partial_\theta^2 b(\theta),$$

and  $\partial_\theta^3 \ell_n(\xi) = -n \phi \partial_\theta^3 b(\theta)$ . Applying Theorem 2.3 completes the proof of Corollary 2.1.

# Elementary Likelihood Theory: Maximum likelihood estimation

For simplicity, we assume in this subsection that  $U_1, \dots, U_n$  are independent random variables each with probability density function  $p(u; \xi)$ , such as the exponential family in (2.1). The joint probability density function of  $\mathbf{U} = (U_1, \dots, U_n)$  is given by  $p(\mathbf{u}; \xi) = \prod_{i=1}^n p(u_i; \xi)$ , and thus its log-likelihood function is a sum of  $n$  terms

$$\ell_n(\xi) = \sum_{i=1}^n \log p(u_i; \xi).$$

# Elementary Likelihood Theory: Maximum likelihood estimation

## Definition 2.3

*The maximum likelihood estimate of  $\xi$  is defined as*

$$\hat{\xi} = \operatorname{argmax}_{\xi \in \Xi} \ell_n(\xi), \quad (2.20)$$

*that is,  $\ell_n(\hat{\xi}) \geq \ell_n(\xi)$  for all  $\xi \in \Xi$ .*

# Elementary Likelihood Theory: Maximum likelihood estimation

For a smooth  $\ell_n(\xi)$ , a necessary condition for the existence of  $\hat{\xi}$  is that

$$\partial_{\xi} \ell_n(\hat{\xi}) = \mathbf{0} \text{ and } -\partial_{\xi}^2 \ell_n(\hat{\xi}) \text{ is positive definite.} \quad (2.21)$$

Although the score equation  $\partial_{\xi} \ell_n(\hat{\xi}) = 0$  may not have an explicit closed-form, the Newton-Raphson algorithm can be used to numerically calculate  $\hat{\xi}$  by using the iterative scheme

$$\xi^{t+1} = \xi^t + \{-\partial_{\xi}^2 \ell_n(\xi^t)\}^{-1} \partial_{\xi} \ell_n(\xi^t). \quad (2.22)$$

This algorithm stops until the absolute difference between consecutive  $\xi^t$ 's is smaller than a small number, say  $10^{-6}$ .

## Elementary Likelihood Theory: Maximum likelihood estimation

The key idea of the Newton-Raphson algorithm is based on the first-order Taylor's series expansion of the score function at a trial value  $\xi^t$  given by

$$\partial_{\xi} \ell_n(\xi^{t+1}) \approx \partial_{\xi} \ell_n(\xi^t) + \partial_{\xi}^2 \ell_n(\xi^t)(\xi^{t+1} - \xi^t). \quad (2.23)$$

By setting  $\partial_{\xi} \ell(\xi^{t+1})$  to zero and solving  $\xi^{t+1}$ , we obtain the iterative equation (2.22).

## Elementary Likelihood Theory: Maximum likelihood estimation

In particular, for the exponential family in (2.1),  $\xi = (\theta, \phi)$ . Then, it follows from Corollary 2.1 that  $\hat{\xi} = (\hat{\theta}, \hat{\phi})$  satisfies

$$\bar{y} = n^{-1} \sum_{i=1}^n y_i = \partial_\theta b(\hat{\theta}) = \mu(\hat{\theta}) = \hat{\mu}. \quad (2.24)$$

Because  $\mu(\cdot)$  is invertible, the maximum likelihood (ML) estimator of  $\theta$  is  $\hat{\theta} = \mu^{-1}(n^{-1} \sum_{i=1}^n y_i)$ .

# Elementary Likelihood Theory: Maximum likelihood estimation

When  $\phi$  is unknown, we can calculate the maximum likelihood estimate of  $\phi$ , denoted by  $\hat{\phi}$ , by maximizing

$$\phi \left[ \sum_{i=1}^n y_i \hat{\theta} - nb(\hat{\theta}) - \sum_{i=1}^n c(y_i) \right] - \frac{1}{2} \sum_{i=1}^n s(y_i, \phi).$$

Equivalently, we solve the score function of  $\phi$  as follows:

$$0.5 \sum_{i=1}^n \partial_\phi s(y_i, \phi) = \sum_{i=1}^n y_i \hat{\theta} - nb(\hat{\theta}) - \sum_{i=1}^n c(y_i).$$

# Elementary Likelihood Theory: Maximum likelihood estimation

## Example 2.8

From Example 2.1, we have  $\hat{\mu} = \hat{\theta} = n^{-1} \sum_{i=1}^n y_i = \bar{y}$  and  
 $\hat{\phi}^{-1} = n^{-1} \sum_{i=1}^n (y_i - \bar{y})^2$ .

## Example 2.9

From Example 2.2, we have  $\hat{\mu} = \exp(\hat{\theta}) = n^{-1} \sum_{i=1}^n y_i$ , and then  
 $\hat{\theta} = \log(n^{-1} \sum_{i=1}^n y_i)$ .

# Elementary Likelihood Theory: Maximum likelihood estimation

For the exponential family, if  $\mu^{-1}(\cdot)$  does not have an explicit form, then we can resort to the iterative equation of the Newton-Raphson algorithm, taking the form

$$\theta^{t+1} = \theta^t + \{-\partial_\theta^2 \ell_n(\theta^t, \phi)\}^{-1} \partial_\theta \ell_n(\theta^t, \phi) \quad (2.25)$$

$$= \theta^t + \{\partial_\theta^2 b(\theta^t)\}^{-1} (n\phi)[\bar{y} - \partial_\theta b(\theta^t)]. \quad (2.26)$$

# Elementary Likelihood Theory: Maximum likelihood estimation

## Example 2.10

Let  $Y_1, \dots, Y_n$  be independent random variables from a  $\text{Binomial}(M, p)$  distribution. In this case,  $\xi = \theta$  and  $\phi$  is a constant. Thus,  $\ell_n(\xi) = \ell_n(\theta)$  can be written as

$$\ell_n(\theta) = \theta \sum_{i=1}^n y_i - nM \log(1 + \exp(\theta)),$$

where  $\theta = \log(p) - \log(1 - p)$ . Differentiating  $\ell_n(\theta)$  with respect to  $\theta$  twice, we can obtain

$$\partial_\theta \ell_n(\theta) = \sum_{i=1}^n y_i - nM \frac{\exp(\theta)}{1 + \exp(\theta)} \quad \text{and} \quad -\partial_\theta^2 \ell_n(\theta) = nM \frac{\exp(\theta)}{[1 + \exp(\theta)]^2}.$$

# Elementary Likelihood Theory: Maximum likelihood estimation

Therefore, the iterative equation of the Newton-Raphson algorithm can be written as

$$\theta^{t+1} = \theta^t + \left\{ nM \frac{\exp(\theta^t)}{[1 + \exp(\theta^t)]^2} \right\}^{-1} \left[ \sum_{i=1}^n y_i - nM \frac{\exp(\theta^t)}{1 + \exp(\theta^t)} \right].$$

As an example, we set  $\theta^0 = 0$ ,  $M = 1$ ,  $n = 30$ , and  $\sum_{i=1}^n y_i = 5$ . The Newton-Raphson algorithm converges in 4 iterations (Table 1.1 and Figure 1.1).

Let's consider another way of calculating  $\hat{\theta}$ . Because  $\hat{\mu} = \hat{p} = \bar{y} = 1/6$  and  $\hat{\theta} = \log(\hat{p}/(1 - \hat{p}))$ , we have  $\hat{\theta} = -\log(5) = -1.609$ .

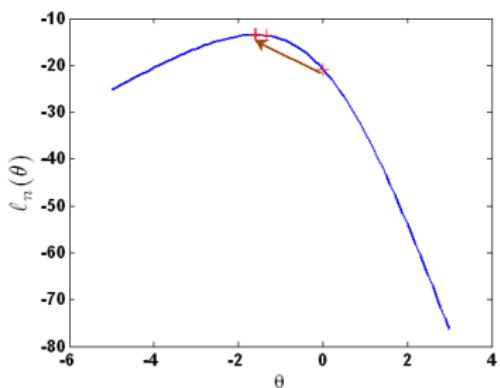
# Elementary Likelihood Theory: Maximum likelihood estimation

Table: Table 1.1: Iterations of the Newton-Raphson algorithm for the Binomial example in which  $\xi = \theta$

Iteration	$\theta$	$\ell_n(\theta)$
0	0	-20.79
1	-1.33	-13.69
2	-1.59	-13.52
3	-1.61	-13.51
4	-1.609	-13.5168

# Elementary Likelihood Theory: Maximum likelihood estimation

Figure: Figure 1.1: The function  $\ell_n(\theta)$  and the trajectory of  $\{\theta^t : t = 0, \dots\}$  in the Newton-Raphson algorithm



## Elementary Likelihood Theory: Central limit theorems

The symbol  $o_p(1)$  denotes that a sequence of random vectors converges to zero in probability. The symbol  $O_p(1)$  denotes that a sequence of random vectors is bounded in probability.

The equation  $W_n = o_p(r_n)$  means  $W_n = s_n r_n$  and  $s_n = o_p(1)$ .

The equation  $W_n = O_p(r_n)$  means  $W_n = S_n r_n$  and  $S_n = O_p(1)$ .

Also, the following facts are useful:

$$o_p(1) + o_p(1) = o_p(1), \quad o_p(1) + O_p(1) = O_p(1),$$

$$O_p(1)o_p(1) = o_p(1), \quad [1 + o_p(1)]^{-1} = O_p(1),$$

$$o_p(r_n) = r_n o_p(1), \quad O_p(r_n) = r_n O_p(1), \quad \text{and} \quad o_p(O_p(1)) = o_p(1).$$

Suppose that  $U_1, \dots, U_n$  are independent and identically distributed with  $E(U) = \mu$  and  $\text{Var}(U) = \sigma^2$ . Then, using the law of large numbers, we have

$$n^{-1} \sum_{i=1}^n U_i \rightarrow E(U) = \mu = O(1)$$

in probability or almost surely, as  $n \rightarrow \infty$ . In this case,

$$n^{-1} \sum_{i=1}^n U_i = O_p(1).$$

Using the central limit theorem, we have

$$n^{-1/2} \sum_{i=1}^n (U_i - \mu) \rightarrow N(0, \sigma^2) = O_p(1)$$

in distribution, as  $n \rightarrow \infty$ . Then, it follows that

$$n^{-1/2} \sum_{i=1}^n (U_i - \mu) = O_p(1).$$

Suppose that  $U_1, \dots, U_n$  are independently and identically distributed with density  $p(u, \xi_*)$ . If  $\partial_\xi \log p(u, \xi_*)$  has a finite covariance matrix, then the central limit theorem ensures that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \partial_\xi \log p(U_i, \xi_*) \xrightarrow{L} N(0, E\{[\partial_\xi \log p(U, \xi_*)] \otimes 2\}) = O_p(1),$$

as  $n \rightarrow \infty$ . Thus, it follows that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \partial_\xi \log p(U_i, \xi_*) = O_p(1). \quad (2.27)$$

Furthermore, if  $E\{-\partial_{\xi}^2 \log p(U, \xi_*)\}$  exists, then the law of large numbers ensures that

$$-\frac{1}{n} \sum_{i=1}^n \partial_{\xi}^2 \log p(U_i, \xi_*) \rightarrow E\{-\partial_{\xi}^2 \log p(U, \xi_*)\} = O(1)$$

in probability or almost surely, as  $n \rightarrow \infty$ . Thus, it follows that

$$-\frac{1}{n} \sum_{i=1}^n \partial_{\xi}^2 \log p(U_i, \xi_*) = O_p(1). \quad (2.28)$$

If a function  $|h(U, \xi)| \leq f(U)$  for all  $\xi$  in a neighborhood of  $\xi_*$ , denoted by  $N(\xi_*, \delta)$ , and  $E[f(U)] < \infty$ , then

$$\left| \frac{1}{n} \sum_{i=1}^n h(U_i, \xi) \right| \leq \frac{1}{n} \sum_{i=1}^n f(U_i) \rightarrow E[f(U)] = O(1)$$

holds for all  $\xi \in N(\xi_*, \delta)$ . Then, it follows that

$$\frac{1}{n} \sum_{i=1}^n h(U_i, \xi) = O_p(1). \quad (2.29)$$

## Elementary Likelihood Theory: Central limit theorems

Suppose that  $U_1, \dots, U_n$  are independently distributed random variables. Then, using the law of large numbers, we have

$$n^{-1} \sum_{i=1}^n [U_i - E(U_i)] = o_p(1)$$

in probability as  $n \rightarrow \infty$ . In this case,

$$n^{-1} \sum_{i=1}^n U_i = n^{-1} \sum_{i=1}^n E(U_i) + o_p(1).$$

# Elementary Likelihood Theory: Central limit theorems

Using the central limit theorem, we have

$$n^{-1/2} \sum_{i=1}^n (U_i - E(U_i)) \rightarrow N(0, \sigma^2) = O_p(1)$$

in distribution, as  $n \rightarrow \infty$ , where  $\sigma^2 = \lim_{n \rightarrow \infty} n^{-1} \sum_{i=1}^n \text{var}(U_i)$ .  
Then, it follows that

$$n^{-1/2} \sum_{i=1}^n (U_i - E(U_i)) = O_p(1).$$

## Example 2.11

Suppose that  $Y_i$ ,  $i = 1, \dots, n$ , are independently generated from a  $N(\mu_i, 1)$  distribution and the  $\mu_i$  are generated from a  $U[-1, 1]$  distribution for  $i = 1, \dots, n$ . Design a simulation study to verify that

$$n^{-1} \sum_{i=1}^n (Y_i - \mu_i) = o_p(1)$$

and

$$n^{-1/2} \sum_{i=1}^n (Y_i - \mu_i) \rightarrow N(0, \sigma^2)$$

in distribution, as  $n \rightarrow \infty$ . What is the true value of  $\sigma^2$ ?

## Example 2.12

Suppose  $\mu_i \equiv \mu$ ,  $i = 1, \dots, n$ . Is it true that

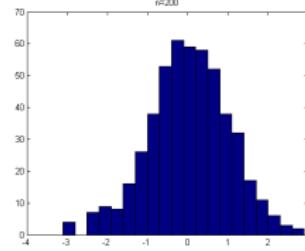
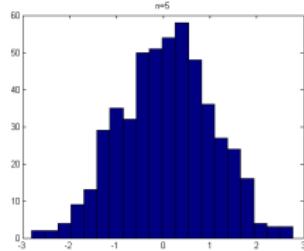
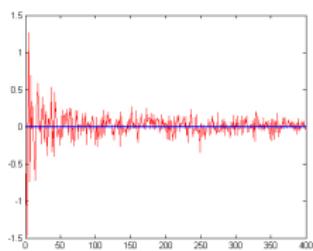
$$n^{-1} \sum_{i=1}^n (Y_i - \mu) = o_p(1)$$

and

$$n^{-1/2} \sum_{i=1}^n (Y_i - \mu) \rightarrow N(0, \sigma^2)?$$

# Elementary Likelihood Theory: Central limit theorems

**Figure:** Figure 1.2: (a)  $n^{-1} \sum_{i=1}^n Y_i$ ; (b)  $n^{-1/2} \sum_{i=1}^n Y_i : n = 5$ ; (c)  $n^{-1/2} \sum_{i=1}^n Y_i : n = 200$ .



# Elementary Likelihood Theory: Central limit theorems

Suppose that  $U_1, \dots, U_n$  are independently distributed random variables. Then, by the law of large numbers, we have

$$n^{-1} \sum_{i=1}^n [f(U_i) - E(f(U_i))] = o_p(1)$$

in probability as  $n \rightarrow \infty$ . It follows that

$$n^{-1} \sum_{i=1}^n f(U_i) = n^{-1} \sum_{i=1}^n E[f(U_i)] + o_p(1).$$

# Elementary Likelihood Theory: Central limit theorems

Using the central limit theorem, we have

$$O_p(1) = n^{-1/2} \sum_{i=1}^n [f(U_i) - E(f(U_i))] \rightarrow N(0, \sigma_f^2) = O_p(1)$$

in distribution, as  $n \rightarrow \infty$ , where  $\sigma_f^2 = \lim_{n \rightarrow \infty} n^{-1} \sum_{i=1}^n \text{var}[f(U_i)]$ .

## Example 2.13

Suppose  $Y_1, \dots, Y_n$  are independent and  $Y_i \sim N(\mu_i, \sigma_i^2)$ , where  $\mu_i$  are fixed numbers and  $\sigma_i \in [a, b]$ , in which  $0 < a \leq b < \infty$ . If we set  $f(U_i) = (Y_i - \mu_i)^2$ , then we can show that

$$n^{-1} \sum_{i=1}^n \{f(U_i) - E[f(U_i)]\} = n^{-1} \sum_{i=1}^n [(Y_i - \mu_i)^2 - \sigma_i^2] = o_p(1)$$

$$n^{-1/2} \sum_{i=1}^n [(Y_i - \mu_i)^2 - \sigma_i^2] \rightarrow N(0, \sigma_f^2),$$

where  $\sigma_f^2 = \lim_{n \rightarrow \infty} n^{-1} \sum_{i=1}^n \text{var}[(Y_i - \mu_i)^2] = 2 \lim_{n \rightarrow \infty} n^{-1} \sum_{i=1}^n \sigma_i^4$ .

**Consistency:**  $\hat{\xi}$  converges to  $\xi_*$  (the ‘true value’) in probability (or almost surely);

**Asymptotic Normality:**  $\sqrt{n}(\hat{\xi} - \xi_*)$  converges a normal random vector  $N(\mathbf{0}, I(\xi_*)^{-1})$ , where  $I(\xi_*) = \lim_{n \rightarrow \infty} n^{-1} I_n(\xi_*)$  denotes the average Fisher information matrix of  $U_1, \dots, U_n$  for  $\xi$ , and  $I_n(\xi_*)$  denotes the Fisher information based on  $n$  observations.

# Elementary Likelihood Theory: Estimation theory

Let  $U_1, \dots, U_n$  be independently and identically distributed as  $p(U_i, \xi_*)$ .

According to Definition 2.3, the ML estimator  $\hat{\xi}$  maximizes the function  $M_n(\xi)$  given by

$$M_n(\xi) = n^{-1} \sum_{i=1}^n \log \frac{p(U_i, \xi)}{p(U_i, \xi_*)}.$$

Under some conditions, we have

$$\sup_{\xi \in \Xi} |M_n(\xi) - M(\xi)| \xrightarrow{P} 0, \quad (2.30)$$

where  $\xrightarrow{P}$  denotes convergence in probability and

$$M(\xi) = E_{\xi_*} \log \frac{p(U_1, \xi)}{p(U_1, \xi_*)}.$$

The number  $-M(\xi)$  is called the *Kullback-Leibler divergence* between  $p_\xi$  and  $p_{\xi_*}$ , which measures the discrepancy between  $p_\xi$  and  $p_{\xi_*}$ .

# Elementary Likelihood Theory: Estimation theory

## Example 2.14

Suppose we generate a dataset with 100 observations  $y_i$ ,  $i = 1, \dots, 100$ , independently from a Poisson(4) distribution. Assume that we fit a Poisson model  $Y_i \sim \text{Poisson}(\lambda)$ . In this case,  $\xi = \lambda$ ,  $\xi_* = \lambda_* = 4$  and  $u_i = y_i$  for  $i = 1, \dots, n = 100$ . For each observation,  $\log p(u, \xi) = u \log(\lambda) - \lambda - \log(u!)$  and  $\log p(u, \xi_*) = u \log(\lambda_*) - \lambda_* - \log(u!)$ . Given the data  $\{y_1, \dots, y_{100}\}$ , we have

$$\begin{aligned}\ell_n(\xi) &= \sum_{i=1}^{100} \{y_i \log(\lambda) - \lambda - \log(y_i!)\} \\ \ell_n(\xi_*) &= \sum_{i=1}^{100} \{y_i \log(4) - 4 - \log(y_i!)\}.\end{aligned}$$

# Elementary Likelihood Theory: Estimation theory

Thus,  $M_n(\xi)$  is given by

$$M_n(\xi) = 100^{-1} \sum_{i=1}^{100} \{y_i \log(\lambda/4) - (\lambda - 4)\}.$$

$\hat{\xi} = \sum_{i=1}^{100} y_i / 100$  maximizes  $M_n(\xi)$ . Using the law of large numbers, we can show that  $M_n(\xi)$  converges to

$$M(\xi) = \int [u \log(\lambda/4) - (\lambda - 4)] p(u, 4) d\Lambda(u) = 4 \log(\lambda/4) - (\lambda - 4).$$

It can be shown that  $\partial_\xi M(\xi) = 4/\lambda - 1$  and thus  $M(\xi)$  attains its maximum uniquely at  $\xi_* = \lambda_* = 4$  by solving  $\partial_\xi M(\xi) = 4/\lambda - 1 = 0$ .

## Lemma 2.1

Suppose that  $p(U, \xi_*)$  is identifiable, that is,  $p(U, \xi_*) \neq p(U, \xi)$  for every  $\xi \neq \xi_*$ . Then,  $M(\xi)$  attains its maximum uniquely at  $\xi_*$ .

**Proof of Lemma 2.1.** Because  $\log x \leq 2(\sqrt{x} - 1)$  for every  $x \geq 0$ , we have

$$\begin{aligned} E_{\xi_*} \log \left( \frac{p(U, \xi)}{p(U, \xi_*)} \right) &\leq 2E_{\xi_*} \left( \sqrt{\frac{p(U, \xi)}{p(U, \xi_*)}} - 1 \right) \\ = 2 \int \sqrt{p(u, \xi)p(u, \xi_*)} d\Lambda(u) - 2 &= - \int (\sqrt{p(u, \xi)} - \sqrt{p(u, \xi_*)})^2 d\Lambda(u). \end{aligned}$$

Thus, the right hand side equals zero if and only if  $p(u, \xi) = p(u, \xi_*)$ .

## Theorem 2.4

If for every  $\epsilon > 0$ ,

$$\sup_{\xi \in \Xi} |M_n(\xi) - M(\xi)| \rightarrow^P 0, \quad \sup_{\xi: ||\xi - \xi_*|| \geq \epsilon} M(\xi) < M(\xi_*),$$

then  $\hat{\xi}$  converges to  $\xi_*$  in probability.

**Proof of Theorem 2.4.** Since  $M_n(\hat{\xi}) \geq M_n(\xi_*)$ , we have  $M_n(\hat{\xi}) \geq M(\xi_*) - o_P(1)$ , and therefore

$$\begin{aligned} 0 &\leq M(\xi_*) - M(\hat{\xi}) \leq M_n(\hat{\xi}) - M(\hat{\xi}) + o_p(1) \\ &\leq \sup_{\xi} |M_n(\xi) - M(\xi)| + o_p(1) \xrightarrow{P} 0. \end{aligned}$$

Thus, for every  $\epsilon > 0$ , there exists a  $\eta > 0$  such that  $M(\xi) < M(\xi_*) - \eta$  for every  $\xi$  with  $\|\xi - \xi_*\| \geq \epsilon$ , where  $\|\cdot\|$  denotes the Euclidean norm of a vector or a matrix. Then, the event  $\{\|\hat{\xi} - \xi_*\| \geq \epsilon\}$  is contained in the event  $\{M(\hat{\xi}) < M(\xi_*) - \eta\}$ , whose probability converges to 0.

## Theorem 2.5

Suppose that the following conditions hold:

- (a)  $\sum_{i=1}^n \partial_\xi \log p(U_i, \hat{\xi}) = \mathbf{0}$ ;
- (b)  $\hat{\xi}$  is a consistent estimate of  $\xi_*$ ;
- (c) for each  $\xi$  in an open subset of Euclidean space,  $\partial_\xi \log p(U, \xi)$  is twice continuously differentiable for every  $x$  and  
 $||\partial_\xi^3 \log p(U, \xi)|| \leq f(U)$  holds for every  $\xi$  in a neighborhood of  $\xi_*$ ,  
where  $f(U)$  is a fixed integrable function;
- (d)  $E[\partial_\xi \log p(U, \xi_*)] = 0$ ,  $E||\partial_\xi \log p(U, \xi_*)||^2 < \infty$ , and the  
matrix  $E[\partial_\xi^2 \log p(U_i, \xi_*)]$  exists and is nonsingular.

# Elementary Likelihood Theory: Estimation theory

**Theorem 2.5 (continued)** Under these conditions, we have

$$\sqrt{n}(\hat{\xi} - \xi_*) = [-E(\partial_{\xi}^2 \log p(U, \xi_*))]^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n \partial_{\xi} \log p(U_i, \xi_*) + o_p(1). \quad (2.31)$$

In particular,  $\sqrt{n}(\hat{\xi} - \xi_*)$  converges to a  $N(0, [-E(\partial_{\xi}^2 \log p(U, \xi_*))]^{-1})$  distribution as  $n \rightarrow \infty$ .

## Proof of Theorem 2.5.

The proof consists of four steps as follows.

In Step 1, Taylor's theorem yields

$$\mathbf{0} = \partial_{\xi} \ell_n(\hat{\xi}) = \partial_{\xi} \ell_n(\xi_*) + \partial_{\xi}^2 \ell_n(\xi_*) \Delta \xi + 0.5 \Delta \xi^T [\partial_{\xi}^3 \ell_n(\tilde{\xi}_n) \Delta \xi],$$

where  $\Delta \xi = \hat{\xi} - \xi_*$  and  $\tilde{\xi}_n = t \xi_* + (1-t) \hat{\xi}$ .

In Step 2, using the central limit theorem, we can show that  $\partial_{\xi} \ell_n(\xi_*) / \sqrt{n}$  converges to a  $N(0, E[\partial_{\xi} \log p(U, \xi_*)^{\otimes 2}])$  distribution as  $n \rightarrow \infty$ , where  $\mathbf{a}^{\otimes 2} = \mathbf{a} \mathbf{a}^T$  for any vector  $\mathbf{a}$ .

# Elementary Likelihood Theory: Estimation theory

In Step 3, using the law of large numbers, we can show that

$$\frac{1}{n} \partial_{\xi}^2 \ell(\xi_*) \xrightarrow{P} E[\partial_{\xi}^2 \log p(U, \xi_*)] \text{ and } \frac{1}{n} \|\partial_{\xi}^3 \ell_n(\tilde{\xi}_n)\| \leq \frac{1}{n} \sum_{i=1}^n f(U_i) = O_p(1).$$

In Step 4, we have

$$-\frac{1}{\sqrt{n}} \partial_{\xi} \ell_n(\xi_*) = (E\{[\partial_{\xi} \log p(U, \xi_*)]^{\otimes 2}\} + o_p(1)) \sqrt{n} (\hat{\xi} - \xi_*).$$

This completes the proof of Theorem 2.5.

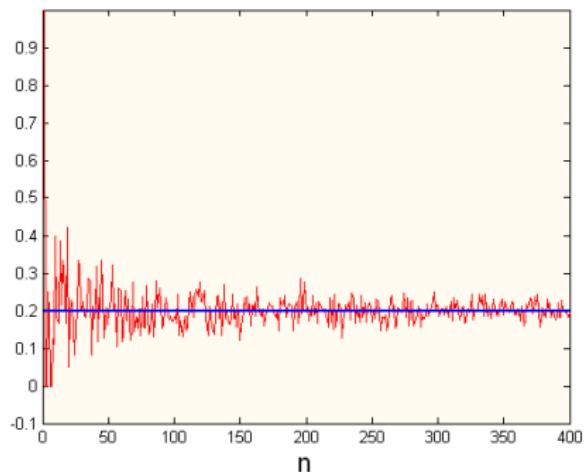
## Example 2.15

Let  $Y_1, \dots, Y_n$  be independent Bernoulli( $1, p_*$ ) random variables. In this case,  $\xi = p$ . It can be shown that the maximum likelihood estimate is  $\hat{\xi} = n^{-1} \sum_{i=1}^n y_i = \bar{y}_{(n)}$ . An application of Lemma 2.1 and Theorem 2.5 yields that  $\hat{\xi} - p_* = o_p(1)$  and  $\sqrt{n}(\hat{\xi} - p_*)$  converges to a  $N(0, p_*(1 - p_*))$  distribution. We leave it to the reader to check all the assumptions in Lemma 2.1 and Theorem 2.5 for this simple example.

We can numerically verify the consistency of  $\bar{y}_{(n)}$  as an estimator of  $p_*$ . First, we generate  $y_1, \dots, y_n$  from a Bernoulli(0.2) distribution and then calculate  $\bar{y}_{(n)}$  for each  $n$ . Second, we compute  $\bar{y}_{(n)}$  for each  $n$ ,  $n = 1, \dots, 100$ . Third, we plot the index of  $\bar{y}_{(n)}$  against  $n$ . If  $\bar{y}_{(n)} - 0.2 = o_p(1)$  holds, we should observe that  $|\bar{y}_{(n)} - 0.2|$  gets smaller as  $n$  increases (Figure 1.3).

# Elementary Likelihood Theory: Estimation theory

Figure: Figure 1.3: The index plot of  $\bar{y}_{(n)}$

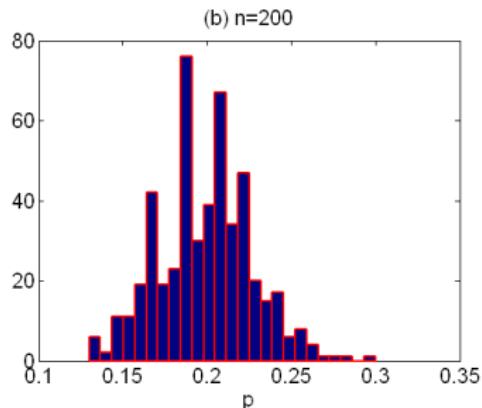
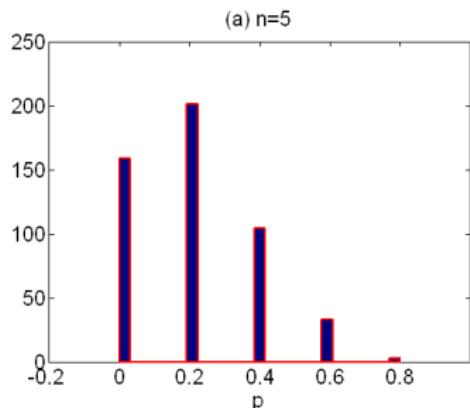


## Elementary Likelihood Theory: Estimation theory

Also, numerically, we can verify the asymptotic normality of  $\bar{y}_n$ . We choose two different values of  $n$ ,  $n = 5$  and  $n = 200$ . For each  $n$ , we generate 500 datasets containing  $\{y_1, \dots, y_n\}$  independently generated from a Bernoulli(0.2) distribution and then we calculate  $\bar{y}_{(n)}$  for each of the 500 datasets. We plot the histograms of these 500  $\bar{y}_{(n)}$ 's for each  $n$ . As expected, the histogram of  $\bar{y}_{(n)}$  gets closer to the normal distribution as  $n$  increases (Figure 1.4).

# Elementary Likelihood Theory: Estimation theory

Figure: Figure 1.4: The histograms of the  $\bar{y}_{(n)}$ 's for  $n = 5$  and  $n = 200$



# Elementary Likelihood Theory: Hypothesis Testing

- **Hypothesis Testing**
- Suppose that in Example 2.15, we are interested in whether  $p = 0.5$  or not. We can set up the hypotheses as

$$H_0 : p = 0.5 \quad \text{vs.} \quad H_1 : p \neq 0.5.$$

The null hypothesis can be written as  $p - 0.5 = 0$ .

- Suppose that we are interested in whether a new treatment is more effective than the old (or standard) treatment. We can set up the hypotheses as

$$H_0 : \mu_{\text{new}} = \mu_{\text{old}} \quad \text{vs.} \quad H_1 : \mu_{\text{new}} > \mu_{\text{old}}.$$

The null hypothesis can be written as  $\mu_{\text{new}} - \mu_{\text{old}} = 0$ .

## Steps in Hypothesis Testing

- Set up Null and Alternative Hypotheses
- Design Considerations: Determine Sample Size and Power under a Specific Test Statistic
- Calculate the Test Statistic - the test statistic has small values under the null hypothesis, but has large values under the alternative hypothesis
- Derive the Asymptotic (or exact) Distribution of the Test Statistic under the Null and Alternative Hypotheses
- Compute the  $p$ -value

# Elementary Likelihood Theory: Hypothesis Testing

The two-sample t-test is used to determine if two population means are equal. A common application of this is to test if a new process or treatment is superior to a current process or treatment. Suppose that we have  $n_1$  observations  $y_{11}, \dots, y_{1n_1}$  from Population 1 and  $n_2$  observations  $y_{21}, \dots, y_{2n_2}$  from Population 2.

- Set up Null and Alternative Hypotheses:  $H_0 : \mu_1 = \mu_2$  vs.  $H_1 : \mu_1 \neq \mu_2$ .
- Sample Size and Power Determination: For the two-sample t-test, the power formula is given by

$$\text{Power} = P \left( \left| \frac{\bar{Y}_1 - \bar{Y}_2}{\sigma \sqrt{1/n_1 + 1/n_2}} \right| > t_{n_1+n_2-2, \alpha/2} \right).$$

- Test Statistic:  $T_n = \frac{\bar{Y}_1 - \bar{Y}_2}{s_p \sqrt{1/n_1 + 1/n_2}}$ , where  $s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$ .

# Elementary Likelihood Theory: Hypothesis Testing

- The Test Statistic  $T_n$  under the Null (and Alternative Hypothesis) follows a  $t_{n_1+n_2-2}$  distribution.
- $p\text{-value} = P(|t_{n_1+n_2-2}| \geq T_n)$ .

## Example 2.16

*The first column is miles per gallon for U.S. cars and the second column is miles per gallon for Japanese cars.*

*SAMPLE 1:  $n_1 = 249$ ,  $\bar{Y}_1 = 20.14458$ ,  $s_1 = 6.414700$ ;*

*SAMPLE 2:  $n_2 = 79$ ,  $\bar{Y}_2 = 30.48101$ ,  $s_2 = 6.107710$ ;*

*If we assume  $\sigma_1 = \sigma_2$ , then we have  $s_p = 6.342600$ ,*

*$\bar{Y}_1 - \bar{Y}_2 = -10.33643$ ,  $T_n = -12.62059$  and  $df = 326$ . The p-value is smaller than 0.0001.*

# Elementary Likelihood Theory: Hypothesis Testing

- **Data:**  $U_1, \dots, U_n$ .
- **Model:**  $p(U_1, \xi), \dots, p(U_n, \xi)$ .
- **Maximum likelihood estimate:**  $\hat{\xi} = \operatorname{argmax} \sum_{i=1}^n \log p(U_i, \xi)$ .
- We consider testing linear or non-linear hypotheses of the form

$$H_0 : h_0(\xi) = b_0 \quad \text{vs.} \quad H_1 : h_0(\xi) \neq b_0, \quad (2.32)$$

where  $h_0(\cdot)$  is an  $r \times 1$  vector function of the  $q$ -vector  $\xi$  with  $q \geq r$  and  $b_0$  is an  $r \times 1$  specified vector. For example, we may have  $h_0(\xi) = \xi_1^2 - 4$ .

## General Test Procedures

- Likelihood ratio test
- Wald test
- Rao's Score test
- All these tests are equivalent to each other in first-order asymptotics
- They differ from each other in second-order asymptotics.

# Elementary Likelihood Theory: Hypothesis Testing

The Wald test is defined by

$$W_n = [h_0(\hat{\xi}) - b_0]^T \{H(\hat{\xi})E[-\partial_{\xi}^2 \ell_n(\hat{\xi})]^{-1}H(\hat{\xi})^T\}^{-1}[h_0(\hat{\xi}) - b_0], \quad (2.33)$$

where  $H(\xi) = \partial h_0(\xi)/\partial \xi$  is an  $r \times q$  matrix. For independent data,  
 $E[\partial_{\xi}^2 \ell_n(\xi)] = \sum_{i=1}^n E[\partial_{\xi}^2 \log p(U_i, \xi)]$ . In practice, we may replace  $E[\partial_{\xi}^2 \ell_n(\hat{\xi})]$  by  
 $\partial_{\xi}^2 \ell_n(\hat{\xi})$ , since  $n^{-1}\{\partial_{\xi}^2 \ell_n(\hat{\xi}) - E[\partial_{\xi}^2 \ell_n(\hat{\xi})]\}$  converges to zero in probability.  
Moreover, using the delta method as developed later, we can show that  
 $H(\hat{\xi})E[-\partial_{\xi}^2 \ell_n(\hat{\xi})]^{-1}H(\hat{\xi})^T$  is an approximation of the covariance matrix of  
 $h_0(\hat{\xi})$ .

# Elementary Likelihood Theory: Hypothesis Testing

Rao's score test is defined as

$$SC_n = \partial_{\xi} \ell_n(\tilde{\xi})^T \left\{ E[\partial_{\xi} \ell_n(\xi)^{\otimes 2}] \right\}^{-1} \Big|_{\xi=\tilde{\xi}} \partial_{\xi} \ell_n(\tilde{\xi}), \quad (2.34)$$

for which we can replace  $E[\partial_{\xi} \ell_n(\tilde{\xi})^{\otimes 2}]$  by  $E[-\partial_{\xi}^2 \ell_n(\tilde{\xi})]$  or  $-\partial_{\xi}^2 \ell_n(\tilde{\xi})$ .

# Elementary Likelihood Theory: Hypothesis Testing

The likelihood ratio test is defined as

$$LRT_n = 2[\ell_n(\hat{\xi}) - \ell_n(\tilde{\xi})], \quad (2.35)$$

where  $\tilde{\xi}$  is the ML estimate of  $\xi$  under the restriction  
 $H_0 : h(\xi) = b_0$ .

# Elementary Likelihood Theory: Hypothesis Testing

We consider testing the linear hypotheses:

$$H_0 : R\xi = b_0 \quad \text{vs.} \quad H_1 : R\xi \neq b_0, \quad (2.36)$$

where  $R$  is an  $r \times q$  matrix of full row rank and  $b_0$  is an  $r \times 1$  specified vector.

# Elementary Likelihood Theory: Hypothesis Testing

Without loss of generality, we assume that  $R\xi = \xi(1)$ , where  $R = (\mathbf{I}_r, \mathbf{0})$ ,  $\xi = (\xi(1)^T, \xi(2)^T)^T$  and  $\xi(1)$  and  $\xi(2)$  are, respectively,  $r \times 1$  and  $(q - r) \times 1$  subvectors of  $\xi$ . Thus, the linear hypotheses can be rewritten as

$$H_0 : \xi(1) = b_0 \quad \text{vs.} \quad H_1 : \xi(1) \neq b_0. \quad (2.37)$$

# Elementary Likelihood Theory: Hypothesis Testing

The ML estimate of  $\xi$  can be written as  $\hat{\xi} = (\hat{\xi}(1)^T, \hat{\xi}(2)^T)^T$ . Under  $H_0$ , the **constrained ML estimate** of  $\xi$  is denoted by  $\tilde{\xi} = (b_0^T, \tilde{\xi}(2)^T)^T$ . We define  $\hat{\xi}(2)(\xi(1))$  as a function of  $\xi(1)$ , which maximizes  $\ell_n(\xi(1), \xi(2))$  for each  $\xi(1)$ . Thus,  
 $\tilde{\xi}(2) = \hat{\xi}(2)(b_0)$ .

# Elementary Likelihood Theory: Hypothesis Testing

For a fixed  $\xi(1)$ , we define

$$(\xi(1), \tilde{\xi}(2)(\xi(1))) = \operatorname{argmax}_{\xi(2)} \ell_n(\xi(1), \xi(2)).$$

The quantity  $\ell_n(\xi(1), \tilde{\xi}(2)(\xi(1)))$  is called the **profile likelihood**.  
Let  $\hat{\xi}(1)^* = \operatorname{argmax}_{\xi(1)} \ell_n(\xi(1), \tilde{\xi}(2)(\xi(1)))$ . If there is a unique  
 $\hat{\xi} = (\hat{\xi}(1), \hat{\xi}(2))$ , then  $\hat{\xi} = (\hat{\xi}(1)^*, \tilde{\xi}(2)(\hat{\xi}(1)^*))$ .

# Elementary Likelihood Theory: Hypothesis Testing

According to the partition  $(\xi(1)^T, \xi(2)^T)$  of  $\xi^T$ , we define

$$\begin{aligned}\partial_\xi \ell_n(\xi)^T &= (\dot{L}_1^T, \dot{L}_2^T), \\ \ddot{L}(\xi) = \partial_\xi^2 \ell_n &= \begin{pmatrix} L_{11} & L_{12} \\ L_{21} & L_{22} \end{pmatrix}, \\ \text{and } \ddot{L}(\xi)^{-1} &= (\partial_\xi^2 \ell_n)^{-1} = \begin{pmatrix} L^{11} & L^{12} \\ L^{21} & L^{22} \end{pmatrix}.\end{aligned}$$

# Elementary Likelihood Theory: Hypothesis Testing

- Wald test statistic:

$$W_n = (\hat{\xi}(1) - b_0)^T \left[ \text{Cov}(\hat{\xi}(1)) \right]^{-1} (\hat{\xi}(1) - b_0). \quad (2.38)$$

- Score test statistic:

$$SC_n = -\dot{L}_1(\tilde{\xi})^T L^{11}(\tilde{\xi}) \dot{L}_1(\tilde{\xi}). \quad (2.39)$$

- Likelihood ratio test statistic:

$$LRT_n = 2[\ell_n(\hat{\xi}) - \ell_n(\tilde{\xi})]. \quad (2.40)$$

# Elementary Likelihood Theory: Hypothesis Testing

**Score test statistic:** We need to calculate  $\partial_{\xi}\ell_n(\tilde{\xi})$  and  $\partial_{\xi}^2\ell_n(\tilde{\xi})$  (or  $E[\partial_{\xi}\ell_n(\tilde{\xi})]^{\otimes 2}]$ ). It can be shown that

$$\partial_{\xi}\ell_n(\tilde{\xi}) = (\dot{L}_1(\tilde{\xi})^T, \dot{L}_2(\tilde{\xi})^T)^T = (\dot{L}_1(\tilde{\xi})^T, \mathbf{0}^T)^T.$$

Thus, using equation (2.34), we can show that

$$SC_n = -(\dot{L}_1(\tilde{\xi})^T, \mathbf{0}^T) \begin{pmatrix} L^{11} & L^{12} \\ L^{21} & L^{22} \end{pmatrix} \begin{pmatrix} \dot{L}_1(\tilde{\xi}) \\ \mathbf{0} \end{pmatrix} = -\dot{L}_1(\tilde{\xi})^T L^{11}(\tilde{\xi}) \dot{L}_1(\tilde{\xi}).$$

We can establish a connection between the ML estimates of  $\xi$  under  $H_0$  and the estimates under the unrestricted space  $\Xi$ .

## Lemma 2.2

*Under  $H_0$  and assumptions of Theorem 2.5,  $\hat{\xi}$  and  $\tilde{\xi}$  satisfy*

$$\hat{\xi}(1) - b_0 = -L^{11}(\tilde{\xi})\dot{L}_1(\tilde{\xi}) + O_p(n^{-1}); \quad (2.41)$$

$$\hat{\xi}(2) - \tilde{\xi}(2)(b_0) = -L_{22}^{-1}L_{21}(\hat{\xi}(1) - b_0) + O_p(n^{-1}). \quad (2.42)$$

# Elementary Likelihood Theory: Hypothesis Testing

**Proof of Lemma 2.2.** Using a Taylor's series expansion, we get

$$\mathbf{0} = \partial_{\xi} \ell_n(\hat{\xi}) = \partial_{\xi} \ell_n(\tilde{\xi}) + \ddot{L}(\tilde{\xi})[\hat{\xi} - \tilde{\xi}] \left[ 1 + O_p\left(\frac{1}{\sqrt{n}}\right) \right].$$

Thus, because  $\dot{L}_2(\tilde{\xi}) = \mathbf{0}$ , we have

$$\begin{pmatrix} \dot{L}_1(\tilde{\xi}) \\ \mathbf{0} \end{pmatrix} = - \begin{pmatrix} L_{11} & L_{12} \\ L_{21} & L_{22} \end{pmatrix} \begin{pmatrix} \hat{\xi}(1) - b_0 \\ \hat{\xi}(2) - \tilde{\xi}(2) \end{pmatrix} \left[ 1 + O_p\left(\frac{1}{\sqrt{n}}\right) \right].$$

Thus,  $\hat{\xi} - \tilde{\xi} = -[\ddot{L}(\tilde{\xi})]^{-1}\dot{L}(\tilde{\xi}) + O_p(n^{-1})$  and

$L_{21}(\hat{\xi}(1) - b_0) + L_{22}(\hat{\xi}(2) - \tilde{\xi}(2)) = \mathbf{0}$ . This completes the proof of Lemma 2.2.

# Elementary Likelihood Theory: Hypothesis Testing

The **Wald test statistic** is given by

$$W_n = (\hat{\xi}(1) - b_0)^T \left[ \text{Cov}(\hat{\xi}(1)) \right]^{-1} (\hat{\xi}(1) - b_0). \quad (2.43)$$

We can estimate  $\text{Cov}(\hat{\xi}(1))$  using  $R\text{Cov}(\hat{\xi})R^T = -L^{11}$ , where  $\mathbf{I}_r$  is an  $r \times r$  identity matrix. In particular, if  $R = \mathbf{e}_m$  is a  $q \times 1$  vector with  $m$ th element equal to 1 and 0 otherwise, then

$$W_n = (\hat{\xi}_m - b_0)^2 / \text{Var}(\hat{\xi}_m).$$

## Elementary Likelihood Theory: Hypothesis Testing

The Wald test statistic is asymptotically distributed as  $\chi^2(r)$ , a chi-square distribution with  $r$  degrees of freedom under the null hypothesis  $H_0$ .

In Step 1, using Theorem 2.5, we can show that  $\sqrt{n}(\hat{\xi} - \xi_*)$  converges to a  $N(0, [-E\partial_{\xi}^2 \log p(U, \xi_*)]^{-1})$  distribution.

In Step 2, if the null hypothesis (2.37) is true, that is,  $\xi(1)_* = b_0$ , then

$$\sqrt{n}R(\hat{\xi} - \xi_*) = \sqrt{n}(\hat{\xi}(1) - b_0) \xrightarrow{L} N(0, R[-E\partial_{\xi}^2 \log p(U, \xi_*)]^{-1}R^T),$$

where  $R = (\mathbf{I}_r, \mathbf{0})$  and  $\xrightarrow{L}$  denotes the convergence in distribution. Thus, we can show that  $W_n \xrightarrow{L} \chi^2(r)$  as  $n \rightarrow \infty$ .

The score test statistic is given by

$$SC_n = -\dot{L}_1(\tilde{\xi})^T L^{11}(\tilde{\xi}) \dot{L}_1(\tilde{\xi}). \quad (2.44)$$

An advantage of using the score test statistic is that it avoids the calculation of an estimator under the alternative hypothesis.

# Elementary Likelihood Theory: Hypothesis Testing

As  $n \rightarrow \infty$ ,  $SC_n \xrightarrow{L} \chi^2(r)$  under the null hypothesis  $H_0$ .

In Step 1, similar to Theorem 2.5, we can show that

$$\tilde{\xi}(2) - \xi(2)_* = -L_{22}(\xi_*)^{-1}\dot{L}_2(\xi_*) + O_p(n^{-1}).$$

In Step 2, we expand  $\dot{L}_1(\tilde{\xi})$  about  $\xi_*$  to get

$$\begin{aligned}\dot{L}_1(\tilde{\xi}) &= \dot{L}_1(\xi_*) + L_{12}(\xi_*)(\tilde{\xi}(2) - \xi(2)_*)[1 + o_p(1)] \\ &= (\mathbf{I}_r, -L_{12}L_{22}^{-1})\dot{L}(\xi_*)[1 + o_p(1)].\end{aligned}$$

Since  $[L^{11}]^{1/2}(\mathbf{I}_r, -L_{12}L_{22}^{-1})\dot{L}(\xi_*) \xrightarrow{L} N(\mathbf{0}, \mathbf{I}_r)$ , this completes the proof.

# Elementary Likelihood Theory: Hypothesis Testing

The likelihood ratio test statistic is given by

$$LRT_n = 2[\ell_n(\hat{\xi}) - \ell_n(\tilde{\xi})]. \quad (2.45)$$

The statistic  $LRT_n$  is asymptotically distributed as  $\chi^2(r)$  under the null hypothesis  $H_0$ .

# Elementary Likelihood Theory: Hypothesis Testing

In Step 1, using  $\partial_\xi \ell_n(\hat{\xi}) = \mathbf{0}$ , we use a Taylor's series expansion to obtain

$$LRT_n = -(\hat{\xi} - \tilde{\xi})^T \partial_\xi^2 \ell_n(\hat{\xi})(\hat{\xi} - \tilde{\xi})[1 + o_p(1)]. \quad (2.46)$$

In Step 2, we expand  $\partial_\xi \ell_n(\hat{\xi}) = 0$  at  $\tilde{\xi}$  to get

$$\mathbf{0} = \partial_\xi \ell_n(\tilde{\xi}) + \partial_\xi^2 \ell_n(\tilde{\xi})(\hat{\xi} - \tilde{\xi})[1 + o_p(1)].$$

Thus, we get

$$\hat{\xi} - \tilde{\xi} = [-\partial_\xi^2 \ell_n(\tilde{\xi})]^{-1} \partial_\xi \ell_n(\tilde{\xi}) + O_p(n^{-1}). \quad (2.47)$$

# Elementary Likelihood Theory: Hypothesis Testing

In Step 3, substituting (2.47) into (2.46), we get that

$$LRT_n = \partial_{\xi} \ell_n(\tilde{\xi})^T [-\partial_{\xi}^2 \ell_n(\tilde{\xi})]^{-1} [-\partial_{\xi}^2 \ell_n(\hat{\xi})] [-\partial_{\xi}^2 \ell_n(\tilde{\xi})]^{-1} \partial_{\xi} \ell_n(\tilde{\xi}) + o_p(1). \quad (2.48)$$

Because  $\partial_{\xi}^2 \ell_n(\tilde{\xi}) \approx \partial_{\xi}^2 \ell_n(\hat{\xi})$  under  $H_0$ ,

$$LRT_n = \partial_{\xi} \ell_n(\tilde{\xi})^T [-\partial_{\xi}^2 \ell_n(\tilde{\xi})]^{-1} \partial_{\xi} \ell_n(\tilde{\xi}) + o_p(1) = -\dot{L}_1(\tilde{\xi})^T L^{11}(\tilde{\xi}) \dot{L}_1(\tilde{\xi}),$$

which converges to a  $\chi^2(r)$  distribution.

An **asymptotically valid test** can be obtained by comparing the sample values of the test statistic  $W_n$  with the critical value of the right-hand tail of a  $\chi^2(r)$  distribution at a pre-specified significance level  $\alpha$ .

We reject  $H_0$  if  $W_n = w_n \geq \chi_\alpha^2(r)$ , and do not reject  $H_0$  otherwise, where  $\chi_\alpha^2(r)$  is the upper  $\alpha$ -percentile of the  $\chi^2(r)$  distribution.

## Example 2.17

Let  $y_1, \dots, y_{100}$  be independent  $Bernoulli(1, 0.2)$  random variables. We are interested in testing  $H_0 : p = 0.3$  vs  $H_1 : p \neq 0.3$ . Let  $\xi = p$ . The log-likelihood function  $\ell_n(\xi)$  is given by

$$\ell_n(\xi) = n\bar{y} \log \xi + n(1 - \bar{y}) \log(1 - \xi).$$

# Elementary Likelihood Theory: Hypothesis Testing

## Example 2.17 (continued)

Differentiating  $\ell_n(\xi)$  with respect to  $\xi$  twice, we have

$$\partial_\xi \ell_n(\xi) = \frac{n(\bar{y} - \xi)}{\xi(1 - \xi)} \quad \text{and} \quad -\partial_\xi^2 \ell_n(\xi) = n \frac{(1 - 2\xi)\bar{y} + \xi^2}{\xi^2(1 - \xi)^2}.$$

Thus,

$$W_n = \frac{n(\hat{p} - 0.3)^2}{\hat{p}(1 - \hat{p})}, \quad SC_n = \frac{n(\bar{y} - 0.3)^2}{(1 - 0.6)\bar{y} + 0.3^2},$$

and

$$LRT_n = 2n\bar{y} \log \left( \frac{\hat{p}}{0.3} \right) + 2n(1 - \bar{y}) \log \left( \frac{1 - \hat{p}}{0.7} \right).$$

## Example 2.18

For instance, if  $\hat{p} = \bar{y} = 0.24$ , then  $W_n = 1.97$ ,  $SC_n = 1.9355$ , and  $LRT_n = 1.7894$  and their corresponding p-values are, respectively, given by 0.16, 0.164, and 0.181. Note that we substituted  $\partial_\xi \ell_n(0.3)$  into  $SC_n$ , whereas we can also use  $E[\partial_\xi \ell_n(0.3)]$  in  $SC_n$ .

# Elementary Likelihood Theory: Hypothesis Testing

We can apply the three test statistics (score, Wald, and likelihood ratio) to test linear hypotheses of  $\xi$  as follows:

$$H_0 : \xi(1) = b_0 \quad \text{vs.} \quad H_1 : \xi(1) \neq b_0, \quad (2.49)$$

where  $\xi(1)$  is the first  $r \times 1$  subvector of  $\xi$ . First, the Wald test statistic is given by

$$W_n = (\hat{\xi}(1) - b_0)^T \left[ R I_n(\hat{\xi})^{-1} R^T \right]^{-1} (\hat{\xi}(1) - b_0),$$

where  $R = [\mathbf{I}_r, \mathbf{0}]$  and  $I_n(\hat{\xi}) = E[-\partial_{\xi}^2 \ell_n(\hat{\xi})] \approx \{Cov(\hat{\xi})\}^{-1}$ .

# Elementary Likelihood Theory: Hypothesis Testing

To calculate the score test statistic, we proceed as follows:

- (a) We calculate  $\tilde{\xi} = (b_0, \tilde{\xi}(2))$ , where  $\xi(2)$  is the last  $p - r$  subvector of  $\xi$ . Specifically,  $\tilde{\xi}(2)$  maximizes  $\ell_n(b_0, \xi(2))$  when  $\xi(1) = b_0$ .
- (b) We calculate  $\partial_{\xi} \ell_n(\tilde{\xi})$ .
- (c) We calculate  $I_n(\tilde{\xi})$  given by

$$I_n(\tilde{\xi}) = E[-\partial_{\xi}^2 \ell_n(\tilde{\xi})].$$

- (d) We calculate  $SC_n = \partial_{\xi} \ell_n(\tilde{\xi})^T I_n(\tilde{\xi})^{-1} \partial_{\xi} \ell_n(\tilde{\xi})$ .

# Elementary Likelihood Theory: Hypothesis Testing

The likelihood ratio test statistic is given by

$$LRT_n = 2[\ell_n(\hat{\xi}) - \ell_n(\tilde{\xi})].$$

## Example 2.19

Let  $y_1, \dots, y_{100}$  be iid  $N(\mu, \sigma^2)$  random variables. We are interested in testing  $H_0 : \mu = 1$  vs.  $H_1 : \mu \neq 1$ . In this case,  $\xi = (\mu, \sigma^2)$ . The log-likelihood function  $\ell_n(\xi)$  is given by

$$\ell_n(\xi) = -\sum_{i=1}^{100} (y_i - \mu)^2 / (2\sigma^2) - 0.5n \log(2\pi\sigma^2).$$

## Example 2.19 (continued)

Differentiating  $\ell_n(\xi)$  with respect to  $\xi$  twice, we have

$$\partial_\xi \ell_n(\xi) = (\dot{L}_1, \dot{L}_2)^T = \left( \sum_{i=1}^{100} (y_i - \mu)/\sigma^2, \sum_{i=1}^{100} (y_i - \mu)^2/(2\sigma^4) - 0.5n/\sigma^2 \right),$$

and

$$\partial_\xi^2 \ell_n(\xi) = \begin{pmatrix} -100/\sigma^2 & -\sum_{i=1}^{100} (y_i - \mu)/\sigma^4 \\ -\sum_{i=1}^{100} (y_i - \mu)/\sigma^4 & -\sum_{i=1}^{100} (y_i - \mu)^2/(\sigma^6) + 0.5n/\sigma^4 \end{pmatrix}.$$

Thus, the Fisher information matrix of  $\xi$  is given by

$$I_n(\hat{\xi}) = E[-\partial_\xi^2 \ell_n(\xi)] = \begin{pmatrix} 100/\sigma^2 & 0 \\ 0 & 0.5n/\sigma^4 \end{pmatrix}.$$

## Example 2.19 (continued)

Furthermore, we calculate

$$\hat{\xi} = (\hat{\mu}, \hat{\sigma}^2)^T = \left( \sum_{i=1}^{100} y_i / 100, \sum_{i=1}^{100} (y_i - \bar{y})^2 / 100 \right)^T$$

$$\text{and } \tilde{\xi} = (\mu_0, \tilde{\sigma}^2)^T = \left( 1, \sum_{i=1}^{100} (y_i - 1)^2 / 100 \right)^T.$$

## Example 2.19 (continued)

For the Wald test, since  $R = (1, 0)$ , we first calculate

$$[\text{Cov}(\hat{\xi}(1))]^{-1} = [R I_n(\hat{\xi})^{-1} R^T]^{-1} = 100/\hat{\sigma}^2.$$

Thus, the Wald test statistic is given by

$$W_n(\hat{\xi}) = (\hat{\mu} - 1)^2 \times 100/\hat{\sigma}^2.$$

## Example 2.19 (continued)

For the score test statistic, we have

$$\partial_{\xi} \ell_n(\tilde{\xi}) = (100(\bar{y} - 1)/\tilde{\sigma}^2, 0)^T$$

and

$$I_n(\tilde{\xi}) = E[-\partial_{\xi}^2 \ell_n(\tilde{\xi})] = \begin{pmatrix} 100/\tilde{\sigma}^2 & 0 \\ 0 & 0.5n/\tilde{\sigma}^4 \end{pmatrix}.$$

Thus,  $SC_n$  is given by

$$SC_n = (\bar{y} - 1)^2 \times 100/\tilde{\sigma}^2.$$

# Elementary Likelihood Theory: Hypothesis Testing

For the likelihood ratio test statistic, since

$2\ell_n(\tilde{\xi}) = -100 - 50 \log(2\pi\tilde{\sigma}^2)$  and  $2\ell_n(\hat{\xi}) = -100 - 50 \log(2\pi\hat{\sigma}^2)$ ,  
we get

$$LRT_n = 50 \log(\tilde{\sigma}^2/\hat{\sigma}^2).$$

## Elementary Likelihood Theory: The Delta Method

Suppose that  $g(\xi_*)$  is a parameter of interest. A natural estimator for  $g(\xi_*)$  is  $g(\hat{\xi})$ . One question is whether we can use the asymptotic properties of  $\hat{\xi}$  to derive the asymptotic properties of  $g(\hat{\xi})$ ?

**Consistency:**  $g(\hat{\xi})$  converges to  $g(\xi_*)$  in probability;

**Asymptotic normality:** as  $n \rightarrow \infty$ ,

$$\sqrt{n}[g(\hat{\xi}) - g(\xi_*)] \xrightarrow{L} N(\mathbf{0}, \partial_{\xi}g(\xi_*)I(\xi_*)^{-1}\partial_{\xi}g(\xi_*)^T).$$

## Theorem 2.6

We have the following results:

- (a) If  $g : R^q \rightarrow R^{q'}$  is continuous at  $\xi_*$  and  $\hat{\xi}$  converges to  $\xi_*$  in probability, then  $g(\hat{\xi})$  converges to  $g(\xi_*)$  in probability;
- (b) If  $g$  is differentiable at  $\xi_*$  and  $\sqrt{n}(\hat{\xi} - \xi_*) \xrightarrow{L} N(0, I(\xi_*)^{-1})$ , then  $\sqrt{n}(g(\hat{\xi}) - g(\xi_*)) \xrightarrow{L} N(0, \partial_{\xi}g(\xi_*)I(\xi_*)^{-1}\partial_{\xi}g(\xi_*)^T)$ .

## Elementary Likelihood Theory: The Delta Method

Proof. (a) Because  $g(\xi)$  is continuous at  $\xi_*$ , for any  $\epsilon > 0$ , there exists  $\delta > 0$  such that  $\|g(\xi) - g(\xi_*)\| \leq \epsilon$  as  $\|\xi - \xi_*\| < \delta$ . Thus,  $\{\|\hat{\xi} - \xi_*\| < \delta\} \subset \{\|g(\hat{\xi}) - g(\xi_*)\| \leq \epsilon\}$  and

$$1 \leftarrow P(\|\hat{\xi} - \xi_*\| < \delta) \leq P(\|g(\hat{\xi}) - g(\xi_*)\| \leq \epsilon) \leq 1,$$

which leads to the desired result.

# Elementary Likelihood Theory: The Delta Method

(b) The function  $g(\cdot)$  is differentiable at  $\xi_*$ , that is,

$$g(\xi_* + \mathbf{h}) - g(\xi_*) = \partial_\xi g(\xi_*) \mathbf{h} + o(||\mathbf{h}||), \quad \mathbf{h} \rightarrow 0.$$

Since  $\sqrt{n}(\hat{\xi} - \xi_*)$  converges to a normal distribution,  $\hat{\xi} - \xi_*$  converges to zero in probability as  $n \rightarrow \infty$ . Thus,

$$g(\hat{\xi}) - g(\xi_*) = \partial_\xi g(\xi_*)(\hat{\xi} - \xi_*) + o_P(||\hat{\xi} - \xi_*||).$$

Multiplying the left and right hand sides with  $\sqrt{n}$ , and noting that  $o_P(\sqrt{n}||\hat{\xi} - \xi_*||) = o_p(1)$ , we get

$$\sqrt{n}[g(\hat{\xi}) - g(\xi_*)] = \partial_\xi g(\xi_*)\sqrt{n}(\hat{\xi} - \xi_*) + o_p(1).$$

Applying the continuous-mapping theorem and Slutsky's lemma then completes the proof of (b).

# Elementary Likelihood Theory: The Delta Method

## Example 2.20

Let  $y_1, \dots, y_n$  be a random sample from a discrete distribution, where  $P(Y = j) = \pi_j$  for  $j = 1, \dots, I$ . Let  $n_j$  be the number of  $y_i = j$ , that is  $n_j = \sum_{i=1}^n \mathbf{1}(y_i = j)$ , where  $\mathbf{1}(\cdot)$  is the indicator function. We have

$$n^{-1}(n_1, \dots, n_I) - \pi = n^{-1} \sum_{i=1}^n [(\mathbf{1}(y_i = 1), \dots, \mathbf{1}(y_i = I)) - E(\mathbf{1}(Y_i = 1), \dots, \mathbf{1}(Y_i = I))]$$

converges to zero in probability, where  $\pi = (\pi_1, \dots, \pi_I)$ . Furthermore,  $\sqrt{n}(n_1/n - \pi_1, \dots, n_I/n - \pi_I)$  can be written as

$$n^{-1/2} \sum_{i=1}^n [(\mathbf{1}(y_i = 1), \dots, \mathbf{1}(y_i = I)) - E(\mathbf{1}(Y_i = 1), \dots, \mathbf{1}(Y_i = I))]$$

## Example (2.20) (continued)

and converges to a  $N(\mathbf{0}, \Sigma)$  distribution as  $n \rightarrow \infty$ , where

$$\Sigma = \begin{pmatrix} \pi_1(1 - \pi_1) & -\pi_1\pi_2 & \cdots & -\pi_1\pi_I \\ -\pi_2\pi_1 & \pi_2(1 - \pi_2) & \cdots & -\pi_2\pi_I \\ \vdots & \vdots & \ddots & \vdots \\ -\pi_I\pi_1 & -\pi_I\pi_2 & \cdots & \pi_I(1 - \pi_I) \end{pmatrix}.$$

## Example (2.20) (continued)

Since  $E(\mathbf{1}(Y_i = j)) = P(Y_i = j) = \pi_j$  and  $E[(\mathbf{1}(Y_i = j))\mathbf{1}(Y_i = k)] = \pi_j$  if  $j = k$  and 0 otherwise, we have  $E(\mathbf{1}(Y_i = 1), \dots, \mathbf{1}(Y_i = I)) = \boldsymbol{\pi}$  and

$$\text{Cov}[(\mathbf{1}(Y_i = 1), \dots, \mathbf{1}(Y_i = I))] = \boldsymbol{\Sigma}.$$

Letting  $\boldsymbol{\Gamma} = \text{diag}(\pi_1, \dots, \pi_I)$ , we can then use the delta method to prove that

$$\sqrt{n}\boldsymbol{\Gamma}^{-1/2}(n_1/n - \pi_1, \dots, n_I/n - \pi_I) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Gamma}^{-1/2}\boldsymbol{\Sigma}\boldsymbol{\Gamma}^{-1/2}).$$

We note here that  $\boldsymbol{\Sigma} = \boldsymbol{\Gamma} - \boldsymbol{\pi}\boldsymbol{\pi}^T$ ,  $\boldsymbol{\Gamma}^{-1/2}\boldsymbol{\Sigma}\boldsymbol{\Gamma}^{-1/2} = \mathbf{I}_I - \sqrt{\boldsymbol{\pi}}^{\otimes 2}$ , where  $\sqrt{\boldsymbol{\pi}} = (\sqrt{\pi_1}, \dots, \sqrt{\pi_I})^T$ .

## Example (2.20) (continued)

Moreover,  $\Gamma^{-1/2}\Sigma\Gamma^{-1/2}$  is a projection matrix and its trace equals  $I - 1$ , since

$$\Gamma^{-1/2}\Sigma\Gamma^{-1/2}\Gamma^{-1/2}\Sigma\Gamma^{-1/2} = \Gamma^{-1/2}\Sigma\Gamma^{-1}\Sigma\Gamma^{-1/2} = \Gamma^{-1/2}\Sigma\Gamma^{-1/2}$$

and

$$\text{tr}[\Gamma^{-1/2}\Sigma\Gamma^{-1/2}] = \text{tr}[\Sigma\Gamma^{-1}] = \text{tr}[\Gamma\Gamma^{-1}] - \text{tr}[\pi\pi^T\Gamma^{-1}] = I - 1.$$

The Pearson chi-square statistic is defined as

$$\chi^2 = n \sum_{j=1}^I \left( \frac{n_j}{n} - \pi_j \right)^2 / \pi_j = [\sqrt{n}\Gamma^{-1/2}(n_1/n - \pi_1, \dots, n_I/n - \pi_I)]^{\otimes 2},$$

which converges to a  $\chi^2(I - 1)$  distribution as  $n \rightarrow \infty$ .

## Example 2.21

Suppose that  $y_1, \dots, y_n$  are independently and identically distributed as  $D(\theta, \phi)$ . Since  $\hat{\mu} = \bar{y} = n^{-1} \sum_{i=1}^n y_i$ , we can directly apply large sample theory to show that  $\hat{\mu}$  converges to  $\mu_* = \mu(\theta_*)$  almost surely and  $\sqrt{n}(\hat{\mu} - \mu_*)$  converges to a normal distribution with mean zero and variance  $\phi^{-1} \partial_\theta^2 b(\theta_*)$ . Since  $\hat{\theta} = \mu^{-1}(\bar{y}) = \theta(\bar{y})$ , we can use the delta method to derive the asymptotic distribution of  $\hat{\theta}$ .

- Definitions
- Estimation methods
- Likelihood theory
- Deviance

# Generalized Linear Models (GLMs): Definition

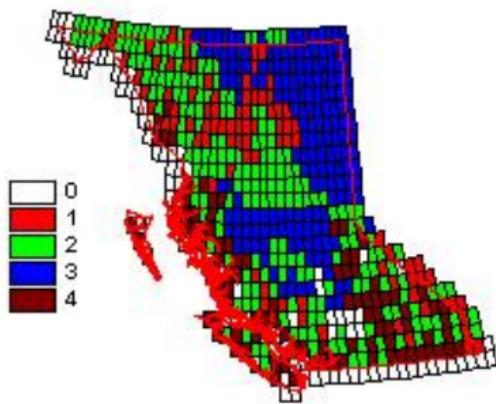
We consider data that is composed of a **response**  $y_i$  and a  $p \times 1$  **covariate** vector  $\mathbf{x}_i$  for  $i = 1, \dots, n$ .

- **Responses** may be **continuous** observations as in classical linear models, such as age, weight, income, or they may be **discrete or ordinal** observations, such as differing severity of diseases and disease status (sick vs. healthy subjects).
- **Covariates** may be **quantitative**, such as age, or **qualitative**, such as gender, race, or presence of risk factors (yes/no).

## Example 3.1

*The data we used were a subset of a large data set from a previous study that addressed the vegetation distribution of all of Canada from an ecophysiological perspective (Lenihan and Neilson, 1993). Our goal was to model the relationship between the distribution of a specific vegetation and the climate variables.*

# Generalized Linear Models: Definition



**Figure:** Figure 2.1: Vegetation distribution in British Columbia, Canada:  
0=background, 1=low arctic shrub tundra, 2=subarctic evergreen woodland,  
3=boreal evergreen forest, and 4=temperate evergreen forest.

## Generalized Linear Models: Definition

The entirety of BC is divided into a lattice of 707 cells with cell size =  $0.5^\circ \times 0.5^\circ$  (Figure 2.1). In each cell, there were records for five climatic variables:

- absolute minimum temperature ( $^{\circ}\text{C}$ ) for the coldest month ( $X_1$ ),
- annual degree-days with base temperature =  $0^{\circ}\text{C}$  ( $X_2$ ),
- total actual evapotranspiration (mm) for summer months ( $X_3$ ),
- annual soil moisture deficit (mm) ( $X_4$ ),
- annual snowpack (mm) ( $X_5$ ).

# Generalized Linear Models: Definition

These climate variables were calculated from two fundamental climatic factors: **monthly temperature** and **monthly precipitation over a 30 year period** (Lenihan and Neilson, 1993).

Some interesting questions in ecology include,

- how to predict the distribution of V2 using the climate variables  $X_1, \dots, X_5$ ?
- whether  $X_1$  is an important predictor for the distribution of V2?

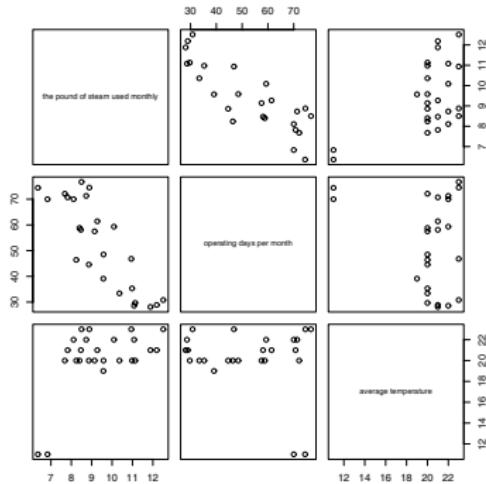
## Example 3.2

We consider a dataset consisting of 25 observations taken from Draper and Smith (1981, p. 205). Each observation includes **the pounds of steam used monthly ( $y_i$ )**, **the operating days per month ( $x_{i2}$ )**, and **the average atmospheric temperature ( $x_{i3}$ )**. Some interesting questions include,

- how to predict  $y_i$  using  $x_{i2}$  and  $x_{i3}$ ?
- whether the operating days per month strongly influence the pounds of steam used monthly?

# Generalized Linear Models: Definition

Figure: Figure 2.2: Pair plots of the pounds of steam used monthly, the operating days per month, and the average atmospheric temperature.



## Example 3.3

Suppose that a new drug was developed for migraine headaches. A double-blind three-arm parallel randomized clinical trial was conducted to evaluate the efficacy of the **new drug (treatment B)** compared to a **marketed drug (treatment C)** and a placebo (**treatment A**). The **pain relief scores** were obtained at 60 minutes after the administration of treatment based on a visual pain relief scale ranging from 0 (no relief from pain) to 10 (complete relief from pain). The pain relief scores for each treatment are included in Table 2.1.

# Generalized Linear Models: Definition

Some interesting questions include whether the new drug is more efficacious than the marketed drug and the placebo?

Table: Table 2.1: Pain Relief Scores for Each Treatment

Treatment	Pain relief	$n_i$	$y_{i\cdot}$	$\bar{y}_{i\cdot}$
A	0.0, 1.0	2	1.0	0.50
B	3.1, 2.7, 3.8	3	9.6	3.20
C	2.3, 3.5, 2.8, 2.5	4	11.1	2.78

# Generalized Linear Models: Definition

## Definition 3.1

*Generalized linear models (GLMs)* are defined as follows:

- the components of  $\mathbf{y} = (y_1, \dots, y_n)^T$  are mutually independent, and the conditional density of  $y_i$  given  $\mathbf{x}_i$  is  $D(\theta_i, \phi/\omega_i)$ , which is a member of the exponential family in (2.1), where  $\theta_i$  is a function of  $\mathbf{x}_i$  and  $\omega_i$  is a weight;
- $\mu_i$  (or equivalently  $\theta_i$ ) is related to  $\mathbf{x}_i$  by

$$g(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta} = \eta_i \quad (3.1)$$

for  $i = 1, \dots, n$ , where  $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$  is defined in a subset  $\mathcal{B}$  of  $R^p$  ( $p < n$ ) and  $g(\cdot)$  is a known *monotonic link function*.

## Generalized Linear Models: Definition

Since  $\mu_i = \dot{b}(\theta_i)$  and  $g(\mu_i) = g(\dot{b}(\theta_i)) = \mathbf{x}_i^T \beta$ , we have

$$\begin{aligned}\mu_i &= g^{-1}(\mathbf{x}_i^T \beta), \\ \theta_i &= \dot{b}^{-1} \circ g^{-1}(\mathbf{x}_i^T \beta) = k(\mathbf{x}_i^T \beta),\end{aligned}$$

where  $\dot{b}(\theta) = \partial_\theta b(\theta)$  and  $\circ$  denotes the product of two functions (as mappings). If  $\theta_i = \eta_i = \mathbf{x}_i^T \beta$ , then  $\dot{b}^{-1} \circ g^{-1}(\cdot) = k(\cdot)$  is an identity function.

## Generalized Linear Models: Definition

When  $g(\cdot) = b^{-1}(\cdot)$ , we call it a **canonical link function**. For example, the canonical link functions for the well-known linear model, loglinear (Poisson) model, and logistic model are, respectively, given by

$$\mu = \mathbf{x}_i^T \boldsymbol{\beta}, \quad \log \lambda = \mathbf{x}_i^T \boldsymbol{\beta}, \quad \text{and} \quad \log \left( \frac{p}{1-p} \right) = \mathbf{x}_i^T \boldsymbol{\beta}.$$

# Generalized Linear Models: Definition

In general, there are **two steps** for identifying the canonical link function:

- Step 1. Identify the canonical parameter for  $D(\theta_i, \phi/\omega_i)$  in the distributional assumption of the generalized linear model (GLM).
- Step 2. Set  $\theta_i = \eta_i = \mathbf{x}_i^T \beta$ .

If  $y_i$  follows a Poisson ( $\lambda_i$ ) distribution, then the canonical parameter for the Poisson( $\lambda_i$ ) distribution is  $\theta_i = \log \lambda_i$  and the canonical link corresponds to  $\theta_i = \log \lambda_i = \eta_i = \mathbf{x}_i^T \beta$ .

## Generalized Linear Models: Definition

For the canonical link, the log-likelihood function of  $(\beta, \phi)$  is given by

$$\ell_n(\beta, \phi) = \sum_{i=1}^n \phi[y_i \mathbf{x}_i^T \beta - b(\mathbf{x}_i^T \beta) - c(y_i)] - \frac{1}{2} \sum_{i=1}^n s(y_i, \phi). \quad (3.2)$$

The quantity  $\sum_{i=1}^n y_i \mathbf{x}_i^T$  is a sufficient statistic for  $\beta$ . Finally, although generalized linear models with canonical links lead to mathematical and statistical convenience, this should **not be the reason for choosing them in applications.**

## Example 3.4

*The classical linear model is defined by the following two assumptions:*

- (i)  $y_i$  given  $\mathbf{x}_i$  follows a  $N(\mu_i, \sigma^2)$  distribution for  $i = 1, \dots, n$ ;
- (ii)

$$\mu_i = \mathbf{x}_i^T \boldsymbol{\beta} = \eta_i \quad \text{for } i = 1, \dots, n. \quad (3.3)$$

# Generalized Linear Models: Estimation Methods

The log-likelihood function of  $\xi = (\beta, \phi)$ , denoted by  $\ell_n(\xi)$ , is given by

$$\ell_n(\xi) = \sum_{i=1}^n \phi[y_i k(\mathbf{x}_i^T \beta) - b(k(\mathbf{x}_i^T \beta)) - c(y_i)] - \frac{1}{2} \sum_{i=1}^n s(y_i, \phi). \quad (3.4)$$

The ML estimate of  $\xi$ , denoted  $\hat{\xi} = (\hat{\beta}, \hat{\phi})$ , can be found in two steps:

- Compute  $\hat{\beta} = \operatorname{argmax}_{\beta} \sum_{i=1}^n [y_i k(\mathbf{x}_i^T \beta) - b(k(\mathbf{x}_i^T \beta)) - c(y_i)];$
- Compute  $\hat{\phi} = \operatorname{argmax}_{\phi} \{\phi \sum_{i=1}^n [y_i k(\mathbf{x}_i^T \hat{\beta}) - b(k(\mathbf{x}_i^T \hat{\beta})) - c(y_i)] - \frac{1}{2} \sum_{i=1}^n s(y_i, \phi)\}.$

# Generalized Linear Models: Estimation Methods

A Newton-Raphson algorithm can be used for computing  $\hat{\beta}$ . We have

$$\dot{\ell}_n(\beta) = \partial_{\beta} \ell_n(\beta) = \phi \sum_{i=1}^n [y_i - \dot{b}(k(\mathbf{x}_i^T \beta))] \dot{k}(\mathbf{x}_i^T \beta) \mathbf{x}_i \quad \text{and} \quad (3.5)$$

$$\begin{aligned} \ddot{\ell}_n(\beta) &= \partial_{\beta}^2 \ell_n(\beta) = -\phi \sum_{i=1}^n \ddot{b}(k(\mathbf{x}_i^T \beta)) \dot{k}(\mathbf{x}_i^T \beta)^2 \mathbf{x}_i \mathbf{x}_i^T \\ &\quad + \phi \sum_{i=1}^n [y_i - \dot{b}(k(\mathbf{x}_i^T \beta))] \ddot{k}(\mathbf{x}_i^T \beta) \mathbf{x}_i \mathbf{x}_i^T, \end{aligned} \quad (3.6)$$

in which  $\dot{b}(\theta_i) = \partial_{\theta} b(\theta)|_{\theta=\theta_i}$ ,  $\dot{k}(\eta) = \partial_{\eta} k(\eta)$ , and  $\ddot{k}(\eta) = \partial_{\eta}^2 k(\eta)$ .

# Generalized Linear Models: Estimation Methods

Let  $\dot{\theta}_i = \partial_{\beta} \theta_i = \dot{k}(\mathbf{x}_i^T \beta) \mathbf{x}_i$  and  $\ddot{\theta}_i = \partial_{\beta}^2 \theta_i = \ddot{k}(\mathbf{x}_i^T \beta) \mathbf{x}_i \mathbf{x}_i^T$ . Let  $\mu_i = \dot{b}(\theta_i)$  and  $v_i = \ddot{b}(\theta_i)$ . Thus,

$$\dot{\ell}_n(\beta) = \phi \sum_{i=1}^n e_i \dot{\theta}_i \quad \text{and} \quad \ddot{\ell}_n(\beta) = -\phi \sum_{i=1}^n v_i \dot{\theta}_i^{\otimes 2} + \phi \sum_{i=1}^n e_i \ddot{\theta}_i, \quad (3.7)$$

where  $e_i = y_i - \mu_i$  and  $\mathbf{a}^{\otimes 2} = \mathbf{a}\mathbf{a}^T$  for any vector  $\mathbf{a}$ . Furthermore, the observed information matrix and the Fisher information matrix of  $\beta$  are, respectively, given by  $-\ddot{\ell}_n(\beta)$  and

$$E[-\ddot{\ell}_n(\beta)] = \phi \sum_{i=1}^n v_i \dot{\theta}_i^{\otimes 2}.$$

# Generalized Linear Models: Estimation Methods

For simplicity, we temporarily drop  $i$  from  $\theta_i$  and  $\mu_i$ . Since  $\mu = \partial_\theta b(\theta) = \dot{b}(\theta)$ , we have

$$\partial_\theta \mu = \partial_\theta^2 b(\theta) = \ddot{b}(\theta) \quad \text{and} \quad \partial_\theta^2 \mu = \partial_\theta^3 b(\theta) = \dddot{b}(\theta).$$

Using the fact that  $\partial_\mu \mu = \partial_\theta \mu \partial_\mu \theta = 1$  and

$$\partial_\mu^2 \mu = \partial_\theta^2 \mu (\partial_\mu \theta)^2 + \partial_\theta \mu \partial_\mu^2 \theta = 0,$$

we have  $\partial_\mu \theta = (\partial_\theta \mu)^{-1} = \ddot{b}(\theta)^{-1}$  and

$$\partial_\mu^2 \theta = -\partial_\theta^2 \mu (\partial_\mu \theta)^3 = -\ddot{b}(\theta) \ddot{b}(\theta)^{-3}. \quad (3.8)$$

# Generalized Linear Models: Estimation Methods

Based on the results in equation (3.8), we get

$$\begin{aligned}\dot{\theta}_i &= \partial_{\beta} \mu_i \partial_{\mu_i} \theta_i = \partial_{\beta} \mu_i [\ddot{b}(\theta_i)]^{-1} \text{ and} \\ \ddot{\theta}_i &= (\partial_{\mu_i}^2 \theta_i) (\partial_{\beta} \mu_i)^{\otimes 2} + \partial_{\mu_i} \theta_i (\partial_{\beta}^2 \mu_i) \\ &= -\ddot{b}(\theta_i) \ddot{b}(\theta_i)^{-3} (\partial_{\beta} \mu_i)^{\otimes 2} + [\ddot{b}(\theta_i)]^{-1} (\partial_{\beta}^2 \mu_i).\end{aligned}\quad (3.9)$$

# Generalized Linear Models: Estimation Methods

We can obtain  $\dot{\ell}_n(\beta)$  and  $E[-\ddot{\ell}_n(\beta)]$  in terms of  $\mu_i$  and its derivatives with respect to  $\beta$  and  $\theta_i$  as follows:

$$\dot{\ell}_n(\beta) = \phi \sum_{i=1}^n \frac{e_i \partial_\beta \mu_i}{\ddot{b}(\theta_i)} = \sum_{i=1}^n \frac{(y_i - \mu_i) \partial_\beta \mu_i}{\text{var}(y_i)} = \sum_{i=1}^n \frac{(y_i - \mu_i) \mathbf{x}_i \partial_{\eta_i} \mu_i}{\text{var}(y_i)},$$

$$E[-\ddot{\ell}_n(\beta)] = \phi \sum_{i=1}^n \frac{[\partial_\beta \mu_i]^{\otimes 2}}{\ddot{b}(\theta_i)} = \sum_{i=1}^n \frac{[\partial_\beta \mu_i]^{\otimes 2}}{\text{var}(y_i)} = \sum_{i=1}^n \frac{\mathbf{x}_i^{\otimes 2}}{\text{var}(y_i)} (\partial_{\eta_i} \mu_i)^2.$$

# Generalized Linear Models: Estimation Methods

Let  $V(\beta) = \text{diag}(v_1(\beta), \dots, v_n(\beta))$ ,  
 $\mathbf{e}(\beta) = (y_1 - \mu_1(\beta), \dots, y_n - \mu_n(\beta))^T$ ,  
 $D_\theta(\beta)^T = (\partial_\beta \theta_1(\beta), \dots, \partial_\beta \theta_n(\beta))_{p \times n}$ ,  
 $D(\beta)^T = (\partial_\beta \mu_1(\beta), \dots, \partial_\beta \mu_n(\beta))_{p \times n}$ .

## Lemma 3.1

For the GLM defined in (3.1), the score function and the Fisher information of  $\beta$  can be written as

$$\dot{\ell}_n(\beta) = \phi D_\theta(\beta)^T \mathbf{e}(\beta) = \phi D(\beta)^T V(\beta)^{-1} \mathbf{e}(\beta) \quad (3.10)$$

$$E[-\ddot{\ell}_n(\beta)] = \phi D_\theta(\beta)^T V D_\theta(\beta) = \phi D(\beta)^T V^{-1} D(\beta). \quad (3.11)$$

# Generalized Linear Models: Estimation Methods

## Example 3.5

The logistic regression model is defined as follows:

- $y_i | \mathbf{x}_i \sim \text{Binomial}(m_i, p_i)$  for  $i = 1, \dots, n$ ;

- 

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1 - p_i}\right) = \mathbf{x}_i^T \boldsymbol{\beta} = \eta_i, \quad \text{for } i = 1, \dots, n. \quad (3.12)$$

The log-likelihood function of  $\boldsymbol{\beta}$  for the logistic regression model is given by

$$\ell_n(\boldsymbol{\beta}) = \sum_{i=1}^n \left\{ y_i \mathbf{x}_i^T \boldsymbol{\beta} - m_i \log(1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})) \right\}.$$

# Generalized Linear Models: Estimation Methods

Since  $\theta_i = \mathbf{x}_i^T \beta$  and  $k(\eta_i) = \eta_i$ , we get  $\dot{k}(\eta_i) = 1$ ,  $\ddot{k}(\eta_i) = 0$ ,  $\dot{\theta}_i = \mathbf{x}_i$ , and  $\ddot{\theta}_i = 0$ . Moreover,  $b(k(\mathbf{x}_i^T \beta)) = m_i \log(1 + \exp(\mathbf{x}_i^T \beta))$ ,  $\mu_i = b(k(\mathbf{x}_i^T \beta)) = m_i p_i$ ,  $\partial_\beta \mu_i = \mathbf{x}_i m_i p_i (1 - p_i)$ , and  $v_i = \ddot{b}(k(\mathbf{x}_i^T \beta)) = m_i p_i (1 - p_i)$ . Therefore, we have

$$\dot{\ell}_n(\beta) = \sum_{i=1}^n (y_i - m_i p_i) \mathbf{x}_i \quad \text{and} \quad -\ddot{\ell}_n(\beta) = E[-\ddot{\ell}_n(\beta)] = \sum_{i=1}^n m_i p_i (1 - p_i) \mathbf{x}_i^{\otimes 2}.$$

# Generalized Linear Models: Estimation Methods

The Newton-Raphson algorithm for obtaining the ML estimate  $\hat{\beta}$  in a GLM is given by

$$\begin{aligned}\beta^{k+1} &= \beta^k + \{E[-\ddot{\ell}_n(\beta^k)]\}^{-1} \dot{\ell}_n(\beta^k), \\ &= \beta^k + \{(D^T V^{-1} D)^{-1} D^T V^{-1} \mathbf{e}\}_{\beta^k} \quad i = 0, 1, 2, \dots\end{aligned}\quad (3.13)$$

where  $\beta^k$  is the estimate of  $\beta$  at the  $k$ -th iteration.

# Generalized Linear Models: Estimation Methods

Let  $G(\beta) = \{D^T(\beta)V^{-1}(\beta)D(\beta)\}^{-1}D^T(\beta)V^{-1}(\beta)\mathbf{e}(\beta)$ . We define the modified Newton-Raphson algorithm as follows:

- (a) choose an initial value  $\beta^0$  and compute  $G^0 = G(\beta^0)$  and find a  $0 < \lambda^0 < 1$  such that

$$\ell_n(\beta^0 + \lambda^0 G^0) > \ell_n(\beta^0);$$

- (b) let  $\beta^1 = \beta^0 + \lambda^0 G^0$  and compute  $G^1 = G(\beta^1)$  and find a  $0 < \lambda^1 < 1$  such that

$$\ell_n(\beta^1 + \lambda^1 G^1) > \ell_n(\beta^1);$$

- (c) set  $\beta^2 = \beta^1 + \lambda^1 G^1, \dots$

To choose the  $\lambda$ 's, we can use  $\lambda = 2^{-1}, 2^{-2}, 2^{-3}, \dots$ .

## Example 3.1 (continued)

We consider **the relationship between the distribution of V2 and the climate variables  $X_1, \dots, X_5$** . Let  $\mathcal{D}$  be the collection of cells in BC and  $(k, l)$  denotes the coordinate of a particular cell in  $\mathcal{D}$ . Assume that  $y_{k,l}$  given  $\mathbf{x}_{k,l}$  follows the logistic regression model, given by

$$\ell_n(\beta) = \sum_{(k,l) \in D} \{y_{k,l} \mathbf{x}_{k,l}^T \beta - \log[1 - \exp(\mathbf{x}_{k,l}^T \beta)]\}, \quad (3.14)$$

and thus

$$\dot{\ell}_n(\beta) = \sum_{(k,l) \in D} [y_{k,l} - p_{k,l}(\beta)] \mathbf{x}_{k,l} \quad \text{and} \quad -\ddot{\ell}_n(\beta) = \sum_{(k,l) \in D} p_{k,l}(\beta)[1 - p_{k,l}(\beta)] \mathbf{x}_{k,l}^{\otimes 2},$$

where  $p_{k,l}(\beta) = 1/[1 + \exp(-\mathbf{x}_{k,l}^T \beta)]$ .

## Example 3.1 (continued)

The Newton-Raphson algorithm is given by

$$\beta^{i+1} = \beta^i + \left\{ \sum_{(k,l) \in D} p_{k,l}(\beta^i)[1 - p_{k,l}(\beta^i)]\mathbf{x}_{k,l}^{\otimes 2} \right\}^{-1} \sum_{(k,l) \in D} [y_{k,l} - p_{k,l}(\beta^i)]\mathbf{x}_{k,l}.$$

Setting  $\beta^0 = \mathbf{0}$ , the Newton-Raphson algorithm converged in 6 iterations. The  $\beta^i$ 's in each iteration and  $\hat{\beta}$  and its standard error (SE) are summarized in Table 2.2.

# Generalized Linear Models: Estimation Methods

**Table:** Table 2.2: Iterations of the Newton-Raphson algorithm. SE denotes the standard error of  $\hat{\beta}$ .

	Iter	$\beta_1$	$\beta_2$	$\beta_3$	$\beta_4$	$\beta_5$	$\beta_6$	$\ell_n(\theta)$
$\beta^1$	1	1.332	0.001496	-0.000819	-0.004521	-0.002736	-0.000898	-490.0551
$\beta^2$	2	1.568	0.001389	-0.001130	-0.003205	-0.003197	-0.001087	-399.6731
$\beta^3$	3	1.552	0.001099	-0.001175	-0.002703	-0.003229	-0.001096	-396.7247
$\beta^4$	4	1.550	0.001087	-0.001176	-0.002691	-0.003230	-0.001096	-396.6825
$\hat{\beta}$	5	1.550	0.001087	-0.001176	-0.002691	-0.003230	-0.001096	-396.6825
SE		1.104	0.021404	0.000476	0.006883	0.002785	0.000291	

# Generalized Linear Models: Estimation Methods

With some simple calculations, we obtain

$$\partial_\phi \ell_n(\phi) = \partial_\phi \ell_n(\phi, \hat{\beta}) = \sum_{i=1}^n d(y_i, \hat{\mu}_i) - \frac{1}{2} \sum_{i=1}^n \dot{s}(y_i, \phi),$$

where  $d(y_i, \hat{\mu}_i) = y_i k(\mathbf{x}_i^T \hat{\beta}) - b(k(\mathbf{x}_i^T \hat{\beta})) - c(y_i)$  and  
 $\dot{s}(y_i, \phi) = \partial_\phi s(y_i, \phi)$ .

# Generalized Linear Models: Estimation Methods

- The observed information matrix and the Fisher information matrix of  $\phi$  are, respectively, given by

$$-\partial_{\phi}^2 \ell_n(\phi) = 0.5 \sum_{i=1}^n \ddot{s}(y_i, \phi) \text{ and } J_{\phi}(\mathbf{y}) = 0.5 \sum_{i=1}^n E[\ddot{s}(y_i, \phi)] \quad (3.15)$$

where  $\ddot{s}(y_i, \phi) = \partial_{\phi}^2 s(y_i, \phi)$ .

- Finally, a Newton-Raphson algorithm for  $\phi$  is given by

$$\hat{\phi}^{k+1} = \hat{\phi}^k + [0.5 \sum_{i=1}^n \ddot{s}(y_i, \hat{\phi}^k)]^{-1} [\sum_{i=1}^n d(y_i, \hat{\mu}_i) - \frac{1}{2} \sum_{i=1}^n \dot{s}(y_i, \hat{\phi}^k)]. \quad (3.16)$$

# Generalized Linear Models: Likelihood Theory

**Consistency:**  $\hat{\xi} = (\hat{\beta}, \hat{\phi})$  converges to  $\xi_* = (\beta_*, \phi_*)$  (the 'true value') in probability (or almost surely);

**Asymptotic Normality:**  $\sqrt{n}(\hat{\xi} - \xi_*)$  converges to a  $N(\mathbf{0}, I(\xi_*)^{-1})$  distribution, where  $I(\xi_*) = \lim_{n \rightarrow \infty} n^{-1} I_n(\xi_*)$  denotes the limit of the average Fisher information matrix based on  $y_1, \dots, y_n$  at  $\xi_*$ . Thus, we can use  $I_n(\xi_*)$  to approximate the covariance matrix of  $\hat{\xi} - \xi_*$ , where  $I_n(\xi_*)$  is given by

$$I_n(\xi_*) \approx \begin{pmatrix} \phi D_\theta(\hat{\beta})^T V(\hat{\beta}) D_\theta(\hat{\beta}) & \mathbf{0} \\ \mathbf{0} & 0.5 \sum_{i=1}^n \ddot{s}(y_i, \hat{\phi}) \end{pmatrix}. \quad (3.17)$$

# Generalized Linear Models: Likelihood Theory

We consider the following hypotheses:

$$H_0 : \beta_1 = b_0 \text{ v.s. } H_1 : \beta_1 \neq b_0, \quad (3.18)$$

where  $\beta_1$  is the first  $r \times 1$  subvector of  $\beta$ .

- The Wald test statistic for generalized linear models is given by

$$W_n = (\hat{\beta}_1 - b_0)^T \left[ R I_n(\hat{\xi})^{-1} R^T \right]^{-1} (\hat{\beta}_1 - b_0), \quad (3.19)$$

where  $R = [\mathbf{I}_r, \mathbf{0}]$ .

- The likelihood ratio test statistic for testing (3.18) is given by

$$LRT_n = 2[\ell_n(\hat{\xi}) - \ell_n(\tilde{\xi})]. \quad (3.20)$$

# Generalized Linear Models: Likelihood Theory

Score test statistic:

- We calculate  $\tilde{\xi} = (\tilde{b}_0, \tilde{\beta}_2, \tilde{\phi})$  where  $\beta_2$  is the last  $p - r$  subvector of  $\beta$  and  $\tilde{\beta}_2$  is the maximizer of  $\sum_{i=1}^n [y_i k(\mathbf{x}_i^T \beta) - b(k(\mathbf{x}_i^T \beta))]$ , in which  $\beta_1 = b_0$ .
- We calculate  $\partial_{\xi} \ell_n(\tilde{\xi}) = \begin{pmatrix} \tilde{\phi} D_{\theta}(\tilde{\beta})^T \mathbf{e}(\tilde{\beta}) \\ 0 \end{pmatrix}$ .
- We calculate  $I_n(\tilde{\xi}) = \begin{pmatrix} \tilde{\phi} D_{\theta}(\tilde{\beta})^T V(\tilde{\beta}) D_{\theta}(\tilde{\beta}) & 0 \\ 0 & * \end{pmatrix}$ .
- From (2.34), it follows that

$$SC_n = SC_n(b_0) = \tilde{\phi}^2 \mathbf{e}(\tilde{\beta})^T D_{\theta}(\tilde{\beta}) I_n(\tilde{\xi})^{-1} D_{\theta}(\tilde{\beta})^T \mathbf{e}(\tilde{\beta}). \quad (3.21)$$

# Generalized Linear Models: Likelihood Theory

## Example 3.6

Consider a simple linear model  $y_i = \mathbf{x}_i^T \beta + \epsilon_i$  for  $i = 1, \dots, n$ , in which  $\mathbf{x}_i = (1, z_i)^T$ ,  $\beta = (\beta_1, \beta_2)^T$ , and  $\sum_{i=1}^n z_i = 0$  represents that the  $z_i$ 's are centered. We are interested in testing the hypothesis  $H_0 : \beta_2 = 0$ .

### Wald Statistic

- In Step 1, calculate the maximum likelihood estimate  $\hat{\xi} = (\hat{\beta}_1, \hat{\beta}_2, \hat{\sigma}^2)^T$  as  $\hat{\beta}_1 = \bar{y} = n^{-1} \sum_{i=1}^n y_i$ ,  $\hat{\beta}_2 = \sum_{i=1}^n y_i z_i / \sum_{i=1}^n z_i^2$ , and  $n\hat{\sigma}^2 = \sum_{i=1}^n (y_i - \mathbf{x}_i^T \hat{\beta})^2$ .
- In Step 2, calculate the Fisher information matrix as

$$I_n(\hat{\xi}) = \begin{pmatrix} \hat{\sigma}^{-2} n & 0 & 0 \\ 0 & \hat{\sigma}^{-2} \sum_{i=1}^n z_i^2 & 0 \\ 0 & 0 & 0.5n\hat{\sigma}^{-4} \end{pmatrix}.$$

## Example 3.6 (continued)

- In Step 3, we calculate the Wald test statistic as

$$W_n = \hat{\beta}_2^2 \left[ R I_n(\hat{\xi})^{-1} R^T \right]^{-1} = \frac{\hat{\sigma}^{-2} (\sum_{i=1}^n y_i z_i)^2}{\sum_{i=1}^n z_i^2},$$

where  $R = (0, 1, 0)$ .

## Example 3.6 (continued)

### Score Statistic

- In Step 1, calculate the maximum likelihood estimate  $\tilde{\xi} = (\tilde{\beta}_1, \tilde{\beta}_2, \tilde{\sigma}^2)^T$  as  $\tilde{\beta}_1 = \bar{y} = n^{-1} \sum_{i=1}^n y_i$ ,  $\tilde{\beta}_2 = 0$ , and  $\tilde{\sigma}^2 = \sum_{i=1}^n (y_i - \bar{y})^2/n$ .
- In Step 2, we have  $\partial_{\xi} \ell_n(\tilde{\xi}) = (0, \sum_{i=1}^n (y_i - \bar{y}) z_i / \tilde{\sigma}^2, 0)^T$ .
- In Step 3, calculate the Fisher information matrix as
$$I_n(\tilde{\xi}) = \begin{pmatrix} \tilde{\sigma}^{-2} n & 0 & 0 \\ 0 & \tilde{\sigma}^{-2} \sum_{i=1}^n z_i^2 & 0 \\ 0 & 0 & 0.5n\tilde{\sigma}^{-4} \end{pmatrix}.$$
- In Step 4, calculate the score test statistic as

$$SC_n = \frac{\tilde{\sigma}^{-2} [\sum_{i=1}^n y_i z_i]^2}{\sum_{i=1}^n z_i^2}.$$

# Generalized Linear Models: Deviance

Most data analyses aim at **predicting/interpreting** a set of data points  $y_1, \dots, y_n$  using a set of fitted values

$\mu_1 = E[y_1|\mathbf{x}_1], \dots, \mu_n = E[y_n|\mathbf{x}_n]$  from a statistical model having a small number of parameters  $\xi$ .

An important question is how to **measure the goodness-of-fit of the proposed statistical model.**

# Generalized Linear Models: Deviance

Consider testing the hypotheses

$$H_0 : \mu = \mu(\beta) \text{ vs. } H_1 : \mu \neq \mu(\beta). \quad (3.22)$$

- The null hypothesis in (3.22), called **the null model**, corresponds to the statistical model proposed to fit the data. Let  $\hat{\beta}$  be the maximum likelihood estimate of  $\beta$  under the null hypothesis and  $\hat{\mu} = \mu(\hat{\beta})$  denotes the ML estimate of  $\mu$ .
- The alternative hypothesis corresponds to **the full model** having  $n$  parameters, one per observation, and thus the ML estimate of  $\mu$  under the full model, denoted by  $\tilde{\mu}$ , equals  $\mathbf{y}$ .

# Generalized Linear Models: Deviance

- The null model for generalized linear models assumes that  $y_i$  given  $\mathbf{x}_i$  follows  $D(\theta_i, \phi)$  and

$$\mu_i = E[y_i | \mathbf{x}_i] = b(\theta_i) = g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta}).$$

Given  $\hat{\boldsymbol{\beta}}$ , we can use  $\mu_i(\hat{\boldsymbol{\beta}}) = E[y_i | \mathbf{x}_i] = g^{-1}(\mathbf{x}_i^T \hat{\boldsymbol{\beta}})$  to predict  $y_i$  for each  $i$ .

- The full model only assumes that  $y_i$  given  $\mathbf{x}_i$  follows  $D(\theta_i, \phi)$  and  $\mu_i = E[y_i | \mathbf{x}_i]$ . There are  $n$  parameters, one per observation, and therefore,  $\tilde{\boldsymbol{\mu}} = \mathbf{y}$ .

# Generalized Linear Models: Deviance

The likelihood ratio statistic for testing the hypotheses in (3.22) is given as follows:

$$\begin{aligned} LRT_n &= 2\{\ell_n(\tilde{\mu}, \mathbf{y}) - \ell_n(\mu(\hat{\beta}), \mathbf{y})\} \\ &= 2\phi \sum_{i=1}^n \{y_i \theta_i(\mu_i) - b(\theta_i(\mu_i))\}_{y_i} - 2\phi \sum_{i=1}^n \{y_i \theta_i(\mu_i) - b(\theta_i(\mu_i))\}_{\mu_i(\hat{\beta})} \\ &= 2\phi \sum_{i=1}^n \{y_i q(y_i) - b(q(y_i))\} - 2\phi \sum_{i=1}^n \{y_i q(\mu_i(\hat{\beta})) - b(q(\mu_i(\hat{\beta})))\}, \end{aligned}$$

where  $\theta = b^{-1}(\mu) = q(\mu)$  for all  $i = 1, \dots, n$ .

# Generalized Linear Models: Deviance

The **deviance** for all  $n$  observations is defined as

$$Dv(\mathbf{y}; \hat{\mu}) = \phi^{-1} LRT_n = \sum_{i=1}^n Dv_i, \quad (3.23)$$

where  $Dv_i = 2\{y_i q(y_i) - b(q(y_i))\} - 2\{y_i q(\mu_i(\hat{\beta})) - b(q(\mu_i(\hat{\beta})))\}$  is the deviance for the  $i$ th observation.

# Generalized Linear Models: Deviance

## Example 3.7

The deviances for the normal, Poisson, binomial, and gamma regression models are, respectively, given by

normal	$\sum_{i=1}^n (y_i - \hat{\mu}_i)^2,$
Poisson	$2 \sum_{i=1}^n \{y_i \log(y_i/\hat{\mu}_i) - (y_i - \hat{\mu}_i)\},$
binomial	$2 \sum_{i=1}^n \{y_i \log(y_i/\hat{\mu}_i)$ $+ (m_i - y_i) \log[(m_i - y_i)/(m_i - \hat{\mu}_i)]\},$
gamma	$2 \sum_{i=1}^n \{-\log(y_i/\hat{\mu}_i) + (y_i - \hat{\mu}_i)/\hat{\mu}_i\},$

where  $\hat{\mu}_i = \mu_i(\hat{\beta})$ .

# Generalized Linear Models: Deviance

## Example 3.8

The full model in the normal linear model has  $n$  parameters and

$$\ell_n(\mu, \sigma^2) = -0.5\sigma^{-2} \sum_{i=1}^n (y_i - \mu_i)^2 - 0.5n \log \sigma^2.$$

It can be shown that  $\partial_{\mu_i} \ell_n(\mu, \sigma^2) = y_i - \mu_i = 0$  and  $\tilde{\mu}_i = y_i$  for  $i = 1, \dots, n$ . Thus,  $\ell_n(\tilde{\mu}, \sigma^2) = -0.5n \log \sigma^2$ . Under the null model,  $\mu_i(\beta) = \mathbf{x}_i^T \beta$  and thus  $\hat{\beta} = (\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T)^{-1} \sum_{i=1}^n \mathbf{x}_i y_i$ . Thus, we have  
 $\ell_n(\mu(\hat{\beta}), \sigma^2) = -0.5\sigma^{-2} \sum_{i=1}^n (y_i - \mathbf{x}_i^T \hat{\beta})^2 - 0.5n \log \sigma^2$ , and therefore

$$Dv(\mathbf{y}; \hat{\mu}) = \sum_{i=1}^n (y_i - \hat{\mu}_i)^2.$$

# Generalized Linear Models: Deviance

## Example 3.9

For the Poisson model, since  $\lambda = \mu$ ,

$$\ell_n(\mu) = \sum_{i=1}^n [y_i \log \mu_i - \mu_i - \log(y_i!)].$$

Thus,  $\partial_{\mu_i} \ell_n(\mu) = y_i/\mu_i - 1 = 0$  leads to  $\tilde{\mu}_i = y_i$ . Therefore,

$\ell_n(\tilde{\mu}) = \sum_{i=1}^n [y_i \log y_i - y_i - \log(y_i!)]$ . Under the null model, suppose that we consider  $\mu_i = \lambda_i = \exp(\mathbf{x}_i^T \beta)$  and  $\hat{\beta}$  is the ML estimate of  $\beta$ . Thus,

$$\ell_n(\mu(\hat{\beta})) = \sum_{i=1}^n [y_i \log \mu_i(\hat{\beta}) - \mu_i(\hat{\beta}) - \log(y_i!)].$$

Finally, we get  $Dv(\mathbf{y}; \hat{\mu})$  as shown above.

# Generalized Linear Models: Deviance

Another important application of the deviance is to **select an 'optimal' model from a sequence of nested models**. Consider two nested models defined by  $\mu = \mu_A(\beta)$  and  $\mu = \mu_B(\beta)$ , respectively. For model  $\mu = \mu_A(\beta)$ , the deviance equals  $Dv(\mathbf{y}; \hat{\mu}_A)$ , while the deviance equals  $Dv(\mathbf{y}; \hat{\mu}_B)$  for model  $\mu = \mu_B(\beta)$ . Therefore, if  $A \subset B$ , then

$$Dv(\mathbf{y}; \hat{\mu}_A) - Dv(\mathbf{y}; \hat{\mu}_B) = 2\phi^{-1}\{\ell_n(\mu_B(\hat{\beta}), \mathbf{y}) - \ell_n(\mu_A(\hat{\beta}), \mathbf{y})\},$$

which is close to the likelihood ratio test statistic except for the factor  $\phi$ .

## Generalized Linear Models: Deviance

If the difference of the deviances is large, then model  $A$  does not fit the data very well compared to model  $B$ . In later chapters, we will discuss the use of the deviance to form an analysis-of-deviance table, which is closely related to the analysis-of-variance table for normal linear models. However, if  $\phi$  is unknown, then we have to estimate  $\phi$  under models  $A$  and  $B$ . In this case, **the deviance difference between model A and model B cannot be interpreted as a scaled likelihood ratio test statistic.** Instead, we can fix  $\phi$  at its estimate from model  $B$ , denoted by  $\hat{\phi}_B$ . Then, we have

$$Dv(\mathbf{y}; \hat{\mu}_A) - Dv(\mathbf{y}; \hat{\mu}_B) = 2\hat{\phi}_B^{-1} \{ \ell_n(\hat{\phi}_B, \mu(\hat{\beta}_B), \mathbf{y}) - \ell_n(\hat{\phi}_B, \mu(\hat{\beta}_A), \mathbf{y}) \}.$$

# Generalized Linear Models: Deviance

Generally, we consider a sequence of nested models defined by  $\{\mu(\beta_{A_i}) : A_1 \subset \cdots \subset A_m, i = 1, \dots, m\}$ .

- (i) Under each model  $A_i$ , we can calculate the maximum likelihood estimators  $\hat{\beta}_{A_i}$  and  $\hat{\phi}_{A_i}$ .
- (ii) To select the best model, we fix  $\phi$  at its estimate from the largest model  $A_m$ , denoted by  $\hat{\phi}_{A_m}$ .
- (iii) We compute the deviance between any two nested models as

$$Dv(\mathbf{y}; \hat{\mu}_{A_k}) - Dv(\mathbf{y}; \hat{\mu}_{A_l}) = 2\hat{\phi}_{A_m}^{-1}\{\ell_n(\hat{\phi}_{A_m}, \mu(\hat{\beta}_{A_k}), \mathbf{y}) - \ell_n(\hat{\phi}_{A_m}, \mu(\hat{\beta}_{A_l}), \mathbf{y})\}.$$

## Example 3.10

*In Example 3.3, a clinical trial was used to evaluate the efficacy of the new drug (treatment B) as compared to a marketed drug (treatment C) and a placebo (treatment A). We can set up an analysis-of-variance model and test the null hypothesis that there is no difference among the new drug, the marketed drug, and the placebo.*

# Generalized Linear Models: Deviance

## Example 3.3 (continued)

We use R to calculate the ANOVA table as follows.

```
res<-lm(bb[,2]~factor(bb[,1]), data=bb)
```

```
anova(res)
```

Analysis of Variance Table

Response: bb[, 2]

Df Sum Sq Mean Sq F value Pr(>F)

Df	Sum Sq	Mean Sq	F value	Pr(>F)
factor(bb[, 1])	2	9.7014	4.8507	14.944 0.004673 **
Residuals	6	1.9475	0.3246	

In R, the command '**lm**' is the function for the linear model and the notation  $Y \sim X$  denotes using the covariates  $X$  to predict  $Y$ . The variable '**res**' is used to save all the output from the linear model. The '**anova**' command creates the anova table based on '**res**'. The  $F$ -value = 14.944 with  $p$ -value = 0.0047 clearly indicates that the three treatments are significantly different from each other.

# Chapter 4: Generalized Linear Models for Continuous Responses

- Classical Linear Model
- Gamma Regression Model
- Inverse Gaussian Regression Model

## Definition 4.1

The *Classical linear model* consists of a

- (i) [Distributional assumption]  $y_i \sim N(\mu_i, \sigma^2)$  for  $i = 1, \dots, n$ ;
- (ii) [Structural assumption]  $\mu_i$  is related to  $\mathbf{x}_i$  by

$$\mu_i = \mathbf{x}_i^T \boldsymbol{\beta} = \sum_{k=1}^p x_{ik} \beta_k = \eta_i \quad (4.1)$$

for  $i = 1, \dots, n$ , where  $x_{i1} = 1$  and  $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$  is defined in a subset  $\mathcal{B}$  of  $R^p$  ( $p < n$ ).

## Covariates

- **Regression models** typically contain continuous covariates, such as age and weight.
- **Analysis-of-variance models** contain only qualitative factors, such as gender and disease status.
- **General Regression Models** may contain continuous covariates, qualitative covariates, and their interactions. There is no need to distinguish between analysis of variance and regression models.

# Gamma Regression

Gamma regression is mainly used in

- modeling continuous positive responses having a constant coefficient of variation, such as survival data, failure time data, insurance claims and amount of rainfall.
- Analyzing count data where the counts are relatively large.

## Definition 4.2

*Gamma regression assumes*

- (i)  $y_i|\mathbf{x}_i \sim G(\mu_i, \nu)$  for  $i = 1, \dots, n$ , where  $G(\mu, \nu)$  denotes the gamma distribution with mean  $\mu$  and shape parameter  $\nu > 0$ ;
- (ii)  $\mu_i$  is related to  $\mathbf{x}_i$  by

$$\mu_i = g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta}) \quad (4.2)$$

for  $i = 1, \dots, n$ , where  $g(\cdot)$  is a given monotone function and  $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$  is defined in a subset  $\mathcal{B}$  of  $R^p$  ( $p < n$ ).

# Gamma Regression

- The gamma distribution with mean  $\mu$  and shape parameter  $\nu > 0$  is given by

$$p(y|\mu, \nu) = \frac{1}{\Gamma(\nu)} \left(\frac{\nu}{\mu}\right)^\nu y^{\nu-1} \exp\left(-\frac{\nu}{\mu}y\right), \quad y \geq 0. \quad (4.3)$$

- The parameter  $\nu/\mu$  is also called the scale parameter.  $G(\mu, \nu)$  belongs to the exponential family with  $\phi = \nu$ ,  $\theta = -1/\mu$ ,  $b(\theta) = -\log(-\theta)$ ,  $c(y) = -\log(y)$ , and  $s(y, \phi) = 2\log(y) + 2\log\Gamma(\nu) - 2\nu\log(\nu)$ .
- $K_y(t) = -\nu \log(1 - \mu t/\nu)$  and  $k_r = (r-1)! \mu^r / \nu^{r-1}$ .
- Thus,  $k_1 = E(y) = \mu$ ,  $\text{var}(y) = \mu^2/\nu$ ,  $k_3 = E(y - \mu)^3 = 2\mu^3/\nu^3$ , and  $k_4 = 6\mu^4/\nu^3$ .

# Gamma Regression

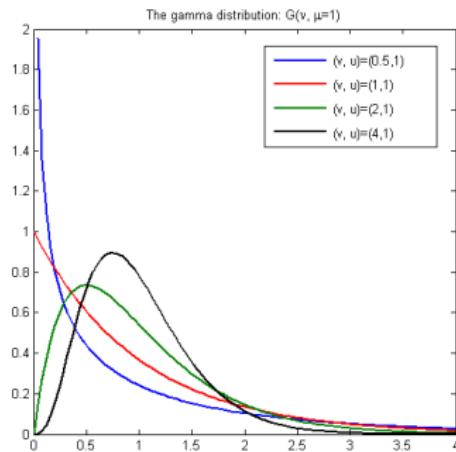


Figure: Figure 3.2: The gamma distribution:  $G(\mu = 1, \nu)$  for  $\nu = 0.5, 1.0, 2.0$  and  $4.0$ .

# Gamma Regression

- For  $\nu \in (0, 1)$ ,  $p(y|\mu, \nu)$  decreases monotonically with  $\nu$ .
- $p(y|\mu, \nu) \rightarrow \infty$  as  $y \rightarrow 0$  and  $p(y|\mu, \nu) \rightarrow 0$  as  $y \rightarrow \infty$ .
- For  $\nu = 1$ , the gamma distribution reduces to the exponential distribution.
- For  $\nu > 1$ , the gamma density has a mode at  $\mu - \mu/\nu$  and is positively skewed.
- As  $\nu \rightarrow \infty$ , the gamma density is close to a normal density.

# Gamma Regression

Let  $\eta = \mathbf{x}^T \boldsymbol{\beta}$ . Three popular link functions include

- $g_1(\mu) = \mu = \eta$  (identity link),
- $g_2(\mu) = -\mu^{-1} = \eta$  (canonical link),
- $g_3(\mu) = \log(\mu) = \eta$  (log link).

# Gamma Regression

- The log-likelihood function of  $(\beta, \nu)$  is given by

$$\ell_n(\beta, \nu) = \sum_{i=1}^n \nu \left\{ -\frac{1}{\mu_i(\beta)} y_i - \log \mu_i(\beta) + \log(y_i) \right\} - \sum_{i=1}^n \log(y_i) - n \log \Gamma(\nu) + n\nu \log(\nu),$$

where  $\mu_i(\beta) = g^{-1}(\mathbf{x}_i^T \beta)$ , that is,  $g(\mu_i(\beta)) = \mathbf{x}_i^T \beta$ .

- The Newton-Raphson algorithm in Chapter 3 can be applied here to estimate  $\beta$ .
- $\partial_\beta \ell_n(\xi) = \sum_{i=1}^n \frac{(y_i - \mu_i(\beta))}{\text{Var}(y_i)} \partial_\beta \mu_i(\beta) = \sum_{i=1}^n \frac{(y_i - \mu_i(\beta))}{\mu_i(\beta)^2 / \nu} \partial_\beta \mu_i(\beta) = \mathbf{0}$ .

# Gamma Regression

- To estimate  $\nu$ , we consider
  - The moment estimator of  $\nu$  ( $\text{var}(y) = \mu^2/\nu$ ) given by

$$\hat{\nu}^{-1} = \sum_{i=1}^n ((y_i - \mu_i(\hat{\beta}))/\mu_i(\hat{\beta}))^2/(n-p). \quad (4.4)$$

- The maximum likelihood estimate of  $\nu$  given by

$$2n\{\log \nu - \partial_\nu \Gamma(\nu)/\Gamma(\nu)\} = \sum_{i=1}^n \left\{ -\frac{1}{\mu_i(\beta)} y_i - \log \mu_i(\beta) + \log(y_i) \right\}. \quad (4.5)$$

The ML estimate of  $\nu$  is sensitive to small values of  $y_i$ .

# Gamma Regression

The **deviance function** for the gamma regression model can be written as

$$Dv(\mathbf{y}; \hat{\mu}) = 2 \sum_{i=1}^n \left\{ -\log \left( \frac{y_i}{\mu_i(\hat{\beta})} \right) + \frac{(y_i - \mu_i(\hat{\beta}))}{\mu_i(\hat{\beta})} \right\}.$$

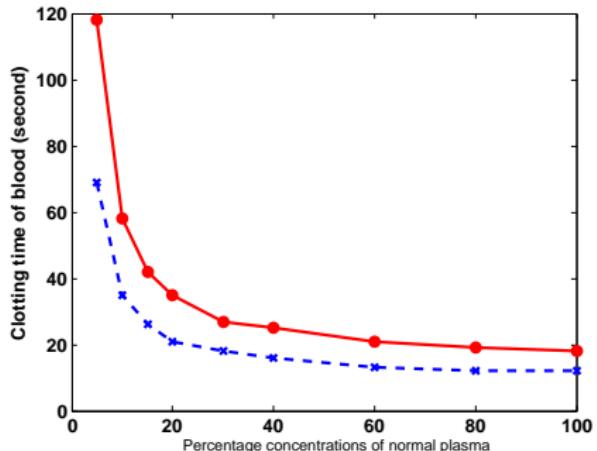
We can use  $Dv(\mathbf{y}; \hat{\mu})$  to construct an analysis-of-deviance table and select an optimal model.

## Example 4.1

We consider a dataset on the clotting time of blood from Hurn et al. (1945) (see Table 8.4 in McCullagh and Nelder (1989)). Each observation includes the clotting time of blood ( $y_i$ ) (unit: second), the percentage concentrations with prothrombin-free plasma ( $x_{i,1}$ ), and the two lots of thromboplastin ( $x_{i,2}$ ) (see Figure 3.3). We analyze the dataset using the gamma regression model with inverse link. We also calculate the analysis of deviance table.

# Gamma Regression

Figure: Figure 3.3: Plot of the clotting time data.



# Gamma Regression

```
clotting <- data.frame( u = c(5,10,15,20,30,40,60,80,100,
5,10,15,20,30,40,60,80,100), lot = c(118,58,42,35,27,25,21,19,18,
69,35,26,21,18,16,13,12,12), cat = c(c(1:9)*0+1, c(1:9)*0))
> summary(glm(lot ~ log(u)*cat, data=clotting, family=Gamma))
> small<-update(full, .~.-log(u):cat)
> summary(small)
> small2<-update(small, .~.-cat)
> summary(small2)
```

In *R*, the command ‘`glm`’ is the function for fitting a generalized linear model to the data, in which ‘`family=Gamma`’ denotes the gamma distribution. The command ‘`update`’ denotes fitting the same model with a new update. For instance, ‘`-log(u):cat`’ denotes the removal of the interaction of  $\log(u)$  and `cat`.

# Gamma Regression

```
Call: glm(formula = lot ~ log(u) * cat, family = Gamma, data = clotting)
```

Coefficients:

Estimate Std. Error t value Pr(> |t|)

(Intercept) -0.0239085 0.0014375 -16.632 1.29e-10 \*\*\*

log(u) 0.0235992 0.0006251 37.754 1.73e-15 \*\*\*

cat 0.0073541 0.0016779 4.383 0.000625 \*\*\*

log(u):cat -0.0082561 0.0007353 -11.229 2.18e-08 \*\*\*

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Gamma family taken to be 0.002129676)

Null deviance: 7.708667 on 17 degrees of freedom

Residual deviance: 0.029401 on 14 degrees of freedom AIC: 63.195

# Gamma Regression

Call: `glm(formula = lot ~ log(u) + cat, family = Gamma, data = clotting)`

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
--	----------	------------	---------	----------

(Intercept) -0.010581 0.002998 -3.529 0.00304 \*\*

$\log(u)$  0.017756 0.001023 17.361 2.43e-11 \*\*\*

cat -0.010868 0.001950 -5.574 5.32e-05 \*\*\*

(Dispersion parameter for Gamma family taken to be 0.01958558)

Null deviance: 7.70867 on 17 degrees of freedom

Residual deviance: 0.30042 on 15 degrees of freedom AIC: 103.07

# Gamma Regression

Call: `glm(formula = lot ~ log(u), family = Gamma, data = clotting)`

Coefficients:

Estimate Std. Error t value Pr(> |t|)

(Intercept) -0.019635 0.004127 -4.757 0.000214 \*\*\*

$\log(u)$  0.018609 0.001827 10.187 2.12e-08 \*\*\*

(Dispersion parameter for Gamma family taken to be 0.06219636)

Null deviance: 7.7087 on 17 degrees of freedom Residual deviance:  
1.0183 on 16 degrees of freedom AIC: 123.17

# Inverse Gaussian Regression

## Definition 4.3

*Inverse Gaussian regression assumes*

- (i)  $y_i \sim IG(\mu_i, \lambda)$  for  $i = 1, \dots, n$ , where  $IG(\mu, \lambda)$  denotes the inverse Gaussian distribution with mean  $\mu$  and shape parameter  $\lambda > 0$ ;
- (ii)  $\mu_i$  is related to  $\mathbf{x}_i$  by

$$\mu_i = g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta}) \quad (4.6)$$

*for  $i = 1, \dots, n$ , where  $g(\cdot)$  is a given monotone function and  $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$  is defined in a subset  $\mathcal{B}$  of  $R^p$  ( $p < n$ ).*

# Inverse Gaussian Regression

- $p(y|\mu, \lambda) = \left(\frac{\lambda}{2\pi y^3}\right)^{1/2} \exp\left[-\frac{\lambda(y-\mu)^2}{2\mu^2 y}\right], \quad y \geq 0.$
- $\text{IG}(\mu, \lambda)$  is a two-parameter exponential family with canonical parameters  $(-\lambda/(2\mu^2), -\lambda/2)$  and sufficient statistics  $(y, 1/y)$ .
- $K_y(t) = \frac{\lambda}{\mu} \left(1 - \sqrt{1 - \frac{2\mu^2 t}{\lambda}}\right),$   
 $k_1 = E(y) = \mu, \quad \text{var}(y) = \mu^3/\lambda, \quad k_3 = 3\sqrt{\mu/\lambda}, \text{ and}$   
 $k_4 = 3 + 15\mu/\lambda.$
- The variance function of the inverse Gaussian distribution increases more rapidly with the mean than the gamma distribution.

# Inverse Gaussian Regression

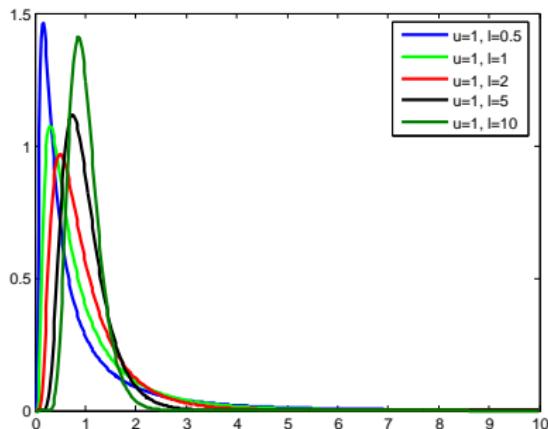


Figure: Figure 3.4: The inverse Gaussian distribution:  $IG(\mu = 1, \lambda)$  for  $\lambda = 0.5, 1, 2, 5$ , and  $10$ .

# Inverse Gaussian Regression

- The log-likelihood function of  $(\beta, \lambda)$  is given by

$$\ell_n(\beta, \lambda) = \sum_{i=1}^n \left\{ -\lambda[y_i/(2\mu_i(\beta)^2) - 1/\mu_i(\beta)] + \lambda/(-2y_i) - 0.5 \log(2\pi y_i^3/\lambda) \right\}.$$

- The Newton-Raphson algorithm can be used to estimate  $\beta$ .
- Since  $\text{var}(y) = \mu^3/\lambda$ , the moment estimator of  $\lambda$  is given by  
$$\hat{\lambda}^{-1} = \sum_{i=1}^n ((y_i - \mu_i(\hat{\beta}))^2 / \mu_i(\hat{\beta})^3) / (n - p).$$
- The maximum likelihood estimate of  $\lambda$  is given by  
$$\hat{\lambda}^{-1} = 2 \sum_{i=1}^n [y_i/(2\hat{\mu}_i^2) - 1/\hat{\mu}_i - 0.5/y_i].$$

# Inverse Gaussian Regression

## Example 4.2

We consider sales data on a range of products from Whitmore (1986). Each observation includes the projected,  $x_i$ , and actual  $y_i$  (see Figure 3.5). Because the sales range from small to large, a normal error can be unreasonable because  $y_i$  is positive. We first fit a normal model. Furthermore, we consider the inverse Gaussian GLM with an identity link  $\mu_i = x_i\beta$ . There is a difference in the estimates of the slope (Figure 3.5). Inspecting the residual plot reveals that the variance of the residuals decreases with  $\log(\mu_i)$ . This indicates that the inverse Gaussian distribution may be not appropriate for these data.

# Inverse Gaussian Regression

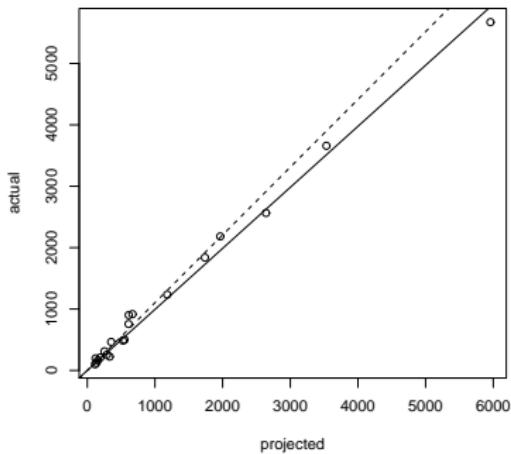


Figure: Figure 3.5: Whitemore's data: The linear model fit (solid line) and the inverse Gaussian GLM fit (dotted line)

# Inverse Gaussian Regression

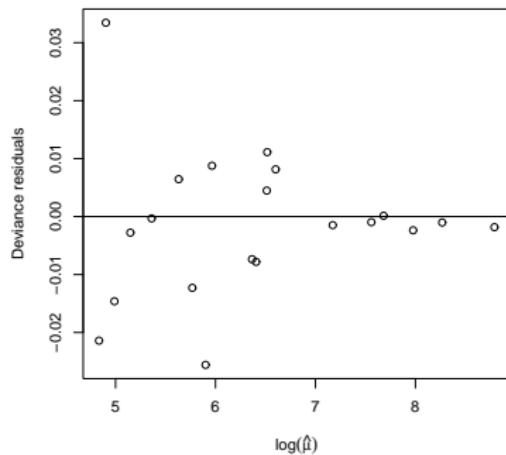


Figure: Figure 3.6: residuals vs. fitted plot for the inverse Gaussian GLM.

# Inverse Gaussian Regression

```
> aa<-read.table('XXX//cpd.txt');  
> lmod<-lm(actual~projected-1, aa)  
> summary(lmod)
```

Residuals:

Min 1Q Median 3Q Max

-250.39 -32.66 39.83 120.61 292.66

Coefficients:

Estimate Std. Error t value Pr(> |t|)

projected 0.99402 0.01718 57.87 <2e-16 \*\*\*

Residual standard error: 139 on 19 degrees of freedom

Multiple R-Squared: 0.9944, Adjusted R-squared: 0.9941

F-statistic: 3349 on 1 and 19 DF, p-value: < 2.2e-16

```
> igmod<-glm(actual~projected-1,  
family=inverse.gaussian(link="identity"), aa)
```

# Inverse Gaussian Regression

```
> summary(igmod)
Deviance Residuals:
Min 1Q Median 3Q Max
-0.025603 -0.007481 -0.001257 0.004970 0.033446
Coefficients:
Estimate Std. Error t value Pr(> |t|)    
projected 1.10358 0.06143 17.96 2.22e-13 ***
(Dispersion parameter for inverse.gaussian family taken to be 0.0001701237)
Null deviance: Inf on 20 degrees of freedom
Residual deviance: 0.0030616 on 19 degrees of freedom
AIC: 268.13
```

# Chapter 5: Generalized Linear Models for Categorical Response: I

- Categorical Response Data
- Distributions for Categorical Variables
- Contingency Tables and Sampling Methods
- Models for Binary and Binomial responses

# Categorical Response Data

Categorical variables are common in the social, biomedical, and behavioral sciences. A **categorical variable** is a measurement scale having a set of categories.

For example, categorical variables include the severity of breast cancer (normal, benign, probably benign, suspicious, and malignant), gender (male and female), the status for a particular disease (presence and absence), and many others.

# Categorical Response Data

## Example 5.1

Scientists were interested in how the distribution of a particular vegetation species is associated with the climate variables, such as absolute minimum temperature for the coldest month in British Columbia. The entirety of BC was divided into a lattice of 707 cells with cell size =  $0.5^\circ \times 0.5^\circ$  (Figure 4.1). In each cell,  $\mathbf{x}_i$  is the climate variables vector and  $y_i = 1$  denotes the presence of a particular species and  $y_i = 0$  denotes absence. We model the absence and presence of a particular species as a function of the covariates  $x_i$ , denoted

$$P(y_i = 0|\mathbf{x}_i) = 1 - \pi(\mathbf{x}_i) \quad \text{and} \quad P(y_i = 1|\mathbf{x}_i) = \pi(\mathbf{x}_i)$$

for  $i = 1, \dots, n$ .

# Categorical Response Data

## Example 5.2

*Neter et al. (1996) presents a consumer dataset from a survey of customers of the Miller Lumber Company. The information includes the total number of customers from each tract of a metropolitan area census and relevant demographic information for each tract. Neter et al. (1996) hypothesize that the **number of customers**  $y_i$  from each of the  $n = 110$  census tracts is associated with the number of housing units  $x_1$ , the average income in dollars  $x_2$ , the average housing unit age in years  $x_3$ , the distance to the nearest competitor in miles  $x_4$ , and the distance to the store in miles  $x_5$ .*

# Categorical Response Data

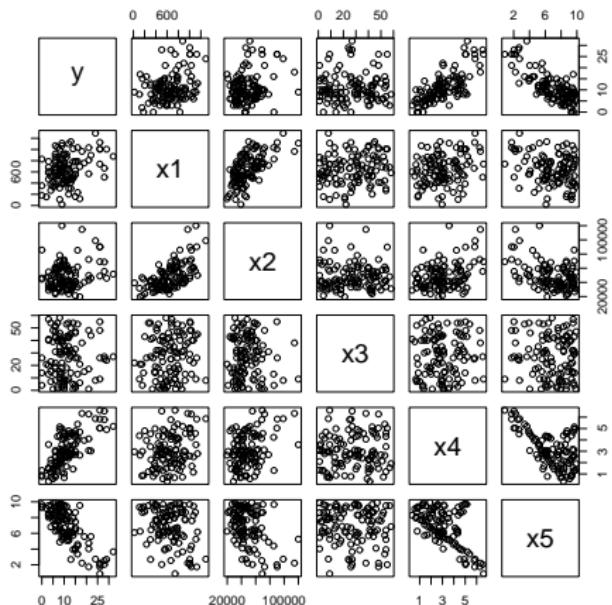


Figure: A pair plots of the consumer dataset

# Categorical Response Data

## Example 5.3

We are interested in the analysis of the association and interaction patterns among alcohol use, cigarette use, and marijuana use based on a substance abuse dataset

Table: Table 4.1: Substance abuse data

Alcohol Use	Cigarette Use	Marijuana Use	
		Yes	No
Yes	Yes	911	538
	No	44	456
No	Yes	3	43
	No	2	279

# Categorical Response Data

Categorical variables have five primary types of scales:

- Dichotomous variables have 2 possible outcomes;
- Nominal variables are qualitative and do not have a natural ordering. An example is gender (male, female);
- Ordinal variables are quantitative and have ordered categories: social class (upper, middle, lower), severity of disease (good, fair, serious, critical);
- Discrete counts;
- Grouped survival data.

# Distributions for Categorical Data

- The **binomial** distribution arises naturally in various contexts where the observations are the number of “successes” and “failures” in a fixed number of experiments.
- Let  $Y_1, \dots, Y_n$  be responses of  $n$  independent and identical trials such that  $P(Y_i = 1) = \pi$  and  $P(Y_i = 0) = 1 - \pi$ . The total number of “successes”, or  $Y = \sum_{i=1}^n Y_i$ , has a  $B(n, \pi)$  distribution.
- $P(Y = y) = \binom{n}{y} \pi^y (1 - \pi)^{n-y}, \quad y = 0, 1, \dots, n$ , where  $\binom{n}{y} = n!/[y!(n - y)!]$ .
- $E(Y) = n\pi$  and  $\text{Var}(Y) = n\pi(1 - \pi)$ .

# Distributions for Categorical Data

- $M_{Y_1}(t) = 1 - \pi + \pi \exp(t)$  and  $K_{Y_1}(t) = \log M_{Y_1}(t) = \log\{1 - \pi + \pi \exp(t)\}.$
- $M_Y(t) = \{1 - \pi + \pi \exp(t)\}^n$  and  
 $K_Y(t) = \log M_Y(t) = n \log\{1 - \pi + \pi \exp(t)\}.$
- $k_1 = n\pi, k_2 = n\pi(1 - \pi), k_3 = n\pi(1 - \pi)(1 - 2\pi), k_4 = n\pi(1 - \pi)\{1 - 6\pi(1 - \pi)\}.$

# Distributions for Categorical Data

The **binomial distribution** is closely related to other distributions.

- If  $Y_1 \sim \text{Poisson}(\mu_1)$  and  $Y_2 \sim \text{Poisson}(\mu_2)$  are independent, then  $Y_1$  given  $Y_1 + Y_2 = n$  follows a  $B(n, \pi)$  distribution, where  $\pi = \mu_1 / (\mu_1 + \mu_2)$ .
- The central hypergeometric distribution.
- As  $n \rightarrow \infty$ ,  $Z = (Y - n\pi) / \sqrt{n\pi(1 - \pi)} \rightarrow N(0, 1)$   
 $\Pr(Y \geq y) \approx 1 - \Phi(z^-)$  and  $\Pr(Y \leq y) \approx \Phi(z^+)$ , where  
 $z^- = (y - n\pi - 0.5) / \sqrt{n\pi(1 - \pi)}$  and  
 $z^+ = (y - n\pi + 0.5) / \sqrt{n\pi(1 - \pi)}$ .

## Empirical logistic transformation

- The canonical parameter  $\theta$  of the  $B(n, \pi)$  distribution is  $\log(\pi) - \log(1 - \pi)$ .  $\pi/(1 - \pi)$  is called the *odds* and  $\theta$  is called the *log odds*.
- The ML estimate is  $\hat{\theta} = \log(\hat{\pi}/(1 - \hat{\pi})) = \log(Y/(n - Y))$ . The asymptotic bias of  $\hat{\theta}$  is  $O(n^{-1})$ .
- The **empirical logistic transformation** (estimator)  $\tilde{\theta} = \log((Y + 0.5)/(n - Y + 0.5))$  has  $O(n^{-2})$  asymptotic bias (Cox, 1970).
- The above arguments support the fact that *adding 0.5 to  $Y$  and  $n - Y$  can reduce the bias in estimating the log odds*.

# Distributions for Categorical Data

Let  $Y = n\pi + \sqrt{n\pi(1-\pi)}Z$ , where  $Z = O_p(1)$  for large  $n$  since  $E(Z) = 0$  and  $\text{Var}(Z) = 1$ . Then, we get

$$\log(Y + 0.5) = \log(n\pi + \sqrt{n\pi(1-\pi)}Z + 0.5) = \log(n\pi) + \log\left(1 + \frac{\sqrt{1-\pi}}{\sqrt{n\pi}}Z + \frac{0.5}{n\pi}\right).$$

Then, we use the formula  $\log(1 + x) = x - 0.5x^2 + x^3/3 + O(x^4)$  to obtain

$$E\{\log(Y + 0.5)\} = \log(n\pi) + \frac{0.5}{n\pi} - \frac{1-\pi}{2n\pi} + O(n^{-2}).$$

Similarly, we can obtain an approximation for  $E\{\log(n - Y + 0.5)\}$  given by

$$E\{\log(n - Y + 0.5)\} = \log(n(1-\pi)) + \frac{0.5}{n(1-\pi)} - \frac{\pi}{2n(1-\pi)} + O(n^{-2}).$$

Finally, we can show that  $E[\tilde{\theta}] = \theta + O(n^{-2})$ .

# Distributions for Categorical Data

- Suppose that  $Y_1, \dots, Y_n$  are iid and  $Y_i$  has  $I$  categories. Write  $Y_i = (y_{i1}, \dots, y_{il})^T$  and  $\sum_{j=1}^I y_{ij} = 1$ . If the  $i$ th trial has the  $k$ th outcome, then  $y_{ik} = 1$  and  $y_{ij} = 0$  for  $j \neq k$ .
- Then  $Y = (Y^1, \dots, Y^l) = \sum_{i=1}^n Y_i = (\sum_{i=1}^n y_{i1}, \dots, \sum_{i=1}^n y_{il})^T$  and  $n_j = \sum_{i=1}^n y_{ij}$  denotes the number of trials having the  $j$ th category.
- Let  $\pi_j = P(y_{ij} = 1)$ , then  $(n_1, \dots, n_l)$  have the **multinomial distribution**, denoted by  $Y \sim Multi(n; \pi_1, \dots, \pi_l)$ ,

$$P(n_1, \dots, n_{l-1}) = \frac{n!}{n_1! \cdots n_l!} \pi_1^{n_1} \cdots \pi_l^{n_l}.$$

# Distributions for Categorical Data

- The moment generating function of the  $Multi(n; \pi_1, \dots, \pi_I)$  distribution is  $M_Y(t) = \{\sum_{j=1}^I \pi_j \exp(t_j)\}^n$ . Thus,  
$$K_Y(t) = n \log\{\sum_{j=1}^I \pi_j \exp(t_j)\}.$$
- $E(n_j) = n\pi_j$ ,  $\text{Var}(n_j) = n\pi_j(1 - \pi_j)$ , and  
 $\text{Cov}(n_j, n_k) = -n\pi_j\pi_k$  for  $j \neq k$ .

# Distributions for Categorical Data

- The Poisson distribution has been widely used to analyze count data, such as predicting the probability of observing any number of events.

$$P(Y = y|\mu) = \frac{e^{-\mu}\mu^y}{y!}; \quad y = 0, 1, 2, \dots,$$

where  $\mu$  is a parameter. We denote  $Y \sim \text{Poisson}(\mu)$ .

- If  $Y$  denotes the number of events, then  $P(Y = 0) = e^{-\mu}$  denotes the probability of no events.
- The cumulant generating function of  $Y$  is given by  $\mu(e^t - 1)$ . Thus, the mean, variance, and all other cumulants of  $Y$  equal  $\mu$ .

# Distributions for Categorical Data

- As  $\mu \rightarrow \infty$ ,

$$(Y - \mu)/\sqrt{\mu} \sim N(0, 1) + O_p(\mu^{-1/2}).$$

- The Poisson distribution is closed under addition. If  $Y_1, \dots, Y_n$  are independent Poisson random variables with means  $\mu_1, \dots, \mu_n$ , respectively, then  
 $S = Y_1 + \dots + Y_n \sim \text{Poisson}(\mu_1 + \dots + \mu_n)$ .
- The Poisson distribution is also a limit of the binomial distribution.
- The multinomial distribution is also related to the Poisson distribution.

# Contingency Tables

A sample of  $n$  subjects is observed to characterize the relationship between two categorical variables  $X$  and  $Y$  and each cell contains the frequency count of a particular outcome for this sample.

Table: Table 4.2: A sample contingency table

		Y				Total
		1	2	...	J	
X	1	$n_{11}$	$n_{12}$	...	$n_{1,J}$	$n_{1\cdot}$
	2	$n_{21}$	$n_{22}$	...	$n_{2,J}$	$n_{2\cdot}$
:	:	:	:	:	:	:
I	$n_{I,1}$	$n_{I,2}$	...	$n_{I,J}$	$n_{I\cdot}$	
Total	$n_{\cdot 1}$	$n_{\cdot 2}$	...	$n_{\cdot J}$	$n$	

# Contingency Tables

We primarily focus on the relationship between two categorical variables using either their joint, marginal, or conditional distributions.

- **joint distribution** of  $(X, Y)$ :  $\pi_{i,j} = P(X = i, Y = j)$ .
- **marginal distribution** of  $X$  or  $Y$ :  
 $\{\pi_{\cdot,j} = \sum_{i=1}^I \pi_{i,j} : j = 1, \dots, J\}$  and  
 $\{\pi_{i,\cdot} = \sum_{j=1}^J \pi_{i,j} : i = 1, \dots, I\}$ .
- **conditional distribution** of  $Y$  given  $X$ :  $\pi_{j|X=i} = \pi_{i,j}/\pi_{i,\cdot}$  for  $j = 1, \dots, J$  and  $i = 1, \dots, I$ .

# Contingency Tables

Table: Table 4.3: A sample joint distribution of two categorical variables

		Y				Total
		1	2	...	J	
X	1	$\pi_{1,1}$	$\pi_{1,2}$	...	$\pi_{1,J}$	$\pi_{1,\cdot}$
	2	$\pi_{2,1}$	$\pi_{2,2}$	...	$\pi_{2,J}$	$\pi_{2,\cdot}$
	:	:	:	:	:	:
I	$\pi_{I,1}$	$\pi_{I,2}$	...	$\pi_{I,J}$	$\pi_{I,\cdot}$	
Total	$\pi_{\cdot,1}$	$\pi_{\cdot,2}$	...	$\pi_{\cdot,J}$	1	

# Contingency Tables

## Example 5.4

For example, in a study of the **association between birthweight and maternal age**, 200 deliveries in a given hospital are randomly selected from the medical records. Then, these 200 deliveries are cross-classified by the weight of the offspring ( $B = \text{birthweight} \leq 2500 \text{ grams}$ ,  $\bar{B} = \text{birthweight} > 2500 \text{ grams}$ ) and by the age of the mother ( $A = \text{age} \leq 20 \text{ years}$ ,  $\bar{A} = \text{age} > 20$ ) in a contingency table given in Table 4.4.

# Contingency Tables

Table: Table 4.4: Association between birthweight and maternal ages: cross-sectional study

Maternal Age= $X$	Birthweight= $Y$			Total
	$B$	$\bar{B}$		
$A$	10	40		50
$\bar{A}$	15	135		150
Total	25	175		200

# Contingency Tables

## Example 5.5

Table: Table 4.5: Substance abuse data

	Peptic ulcer		Control	
	Group O	Group A	Group O	Group A
London	911	579	4578	4219
Manchester	361	246	4532	3775
Newcastle	396	219	6598	5261

Cox and Snell (1989) analyzed a dataset on the incidence of peptic ulcers from three cities. We are interested in investigating the possible effect of blood group on the incidence of having peptic ulcer.

# Contingency Tables

A sample of  $n$  subjects is observed to characterize the relationship between two categorical variables  $X$  and  $Y$ .

Table: Table 4.6: A sample contingency table

		Y				Total
		1	2	...	J	
X	1	$n_{11}$	$n_{12}$	...	$n_{1,J}$	$n_{1\cdot}$
	2	$n_{21}$	$n_{22}$	...	$n_{2,J}$	$n_{2\cdot}$
	:	:	:	:	:	:
I	$n_{I,1}$	$n_{I,2}$	...	$n_{I,J}$	$n_{I\cdot}$	
Total	$n_{\cdot 1}$	$n_{\cdot 2}$	...	$n_{\cdot J}$	$n$	

# Contingency Tables

Whether we can consistently estimate  $\pi_{i,j}$  or not in a contingency table depends on the sampling model that is used in a particular study.

# Sampling Methods

There are three types of sampling methods:

- cross-sectional, naturalistic, or multinomial sampling
- purposive sampling
  - comparative prospective (cohort)
  - retrospective (case-control) studies.
- Sampling method III

# Sampling Methods

- Cross-sectional sampling starts from selecting a total of  $n$  subjects from a larger population and then it determines the frequency count of each possible outcome of  $(X, Y)$ .
- For cross-sectional sampling, we can estimate all the joint probabilities of  $(X, Y)$ , marginal probabilities of  $X$  and  $Y$ , and conditional probabilities of  $X$  given  $Y$  and  $Y$  given  $X$ .

# Sampling Methods

To test whether  $X$  and  $Y$  are associated with each other, we consider the hypothesis

$$H_0 : \pi_{i,j} = \pi_{i,\cdot} \pi_{\cdot,j}.$$

- The Pearson chi-square test statistic is given by

$$\chi^2 = n \sum_{i=1}^2 \sum_{j=1}^2 \frac{|\hat{\pi}_{ij} - \hat{\pi}_{i,\cdot} \hat{\pi}_{\cdot,j}|^2}{\hat{\pi}_{i,\cdot} \hat{\pi}_{\cdot,j}}. \quad (5.1)$$

- As  $n \rightarrow \infty$ ,  $\chi^2 \xrightarrow{L} \chi^2(1)$ . A correction of  $\chi^2$  is given by

$$\chi^2 = n \sum_{i=1}^2 \sum_{j=1}^2 \frac{|\hat{\pi}_{ij} - \hat{\pi}_{i,\cdot} \hat{\pi}_{\cdot,j}| - 1/(2n)|^2}{\hat{\pi}_{i,\cdot} \hat{\pi}_{\cdot,j}}. \quad (5.2)$$

# Sampling Methods

Table: Table 4.7: Association between birthweight and maternal ages: cross-sectional study

Maternal Age= $X$	Birthweight= $Y$		
	$B$	$\bar{B}$	Total
$A$	10	40	50
$\bar{A}$	15	135	150
Total	25	175	200

# Sampling Methods

Table: Table 4.8: Estimated Proportions between birthweight and maternal ages: cross-sectional study

Maternal Age= $X$	Birthweight= $Y$			Total
	$B$	$\bar{B}$		
$A$	0.05	0.2		0.25
$\bar{A}$	0.075	0.675		0.75
Total	0.125	0.875		1.0
	$\chi^2 = 2.58$			

# Sampling Methods

For the  $2 \times 2$  table, the relative risk of  $Y = 2$  over  $Y = 1$  for  $X = 2$  can be measured by

$$odds(P(Y = 2|X = 2)) = \frac{P(Y = 2|X = 2)}{P(Y = 1|X = 2)} \approx \frac{n_{22}/n_{2\cdot}}{n_{21}/n_{2\cdot}} = \frac{n_{22}}{n_{21}}.$$

Similarly, for  $X = 1$ , the relative risk of  $Y = 2$  over  $Y = 1$  can be measured by

$$odds(P(Y = 2|X = 1)) = \frac{P(Y = 2|X = 1)}{P(Y = 1|X = 1)} \approx \frac{n_{12}/n_{1\cdot}}{n_{11}/n_{1\cdot}} = \frac{n_{12}}{n_{11}}.$$

# Sampling Methods

- The **odds ratio** of two odds,  $O(X = 2)$  and  $O(X = 1)$ , is defined and estimated by

$$R = \frac{\text{odds}(P(Y = 2|X = 2))}{\text{odds}(P(Y = 2|X = 1))} \approx \hat{R} = \frac{n_{22}n_{11}}{n_{21}n_{12}}. \quad (5.3)$$

- The **standard error** of  $\hat{R}$  can be estimated by

$$\text{s.e.}(\hat{R}) = \frac{\hat{R}}{\sqrt{n}} \sqrt{\frac{1}{\hat{\pi}_{11}} + \frac{1}{\hat{\pi}_{12}} + \frac{1}{\hat{\pi}_{21}} + \frac{1}{\hat{\pi}_{22}}}, \quad (5.4)$$

where  $\hat{\pi}_{i,j} = n_{ij}/n$  for  $i, j = 1, 2$ .

# Sampling Methods

If  $R = 1$ , this indicates that  $X$  and  $Y$  are independent.



$$odds(P(Y = 2|X = 2)) = odds(P(Y = 2|X = 1)) = c_0,$$

where  $c_0$  is a constant.

- $\pi_{2,2} = c_0\pi_{2,1}$  and  $\pi_{1,2} = c_0\pi_{1,1}$ .
- $\pi_{2,2} + \pi_{1,2} = \pi_{\cdot,2} = c_0\pi_{\cdot,1}$ .
- Since  $\pi_{\cdot,1} + \pi_{\cdot,2} = 1$ , we have  $\pi_{\cdot,1} = 1/(c_0 + 1)$  and  $\pi_{\cdot,2} = c_0/(c_0 + 1)$ .
- Similarly,  $\pi_{2,1} + \pi_{2,2} = \pi_{2,\cdot} = (c_0 + 1)\pi_{2,1}$  and  $\pi_{1,2} + \pi_{1,1} = \pi_{1,\cdot} = (c_0 + 1)\pi_{1,1}$ .
- $\pi_{i,j} = \pi_{i,\cdot}\pi_{\cdot,j}$  for  $i, j = 1, 2$ .

# Sampling Methods

Another estimate of  $R$  is given by

$$R \approx \tilde{R} = \frac{(n_{22} + 0.5)(n_{11} + 0.5)}{(n_{21} + 0.5)(n_{12} + 0.5)}. \quad (5.5)$$

The odds ratio  $R$  proposed by Cornfield (1951) is an important measure of the degree of association between  $X$  and  $Y$ .

# Sampling Methods

- We consider  $\log \hat{R}$  (or  $\log \tilde{R}$ ) instead of  $\hat{R}$ , because  $\log \hat{R}$  converges rapidly to a normal distribution compared to  $\hat{R}$ .
- An estimated standard error of  $\log \hat{R}$  is

$$\text{s.e.}(\log \hat{R}) = \frac{1}{\sqrt{n}} \sqrt{\frac{1}{\hat{\pi}_{11}} + \frac{1}{\hat{\pi}_{12}} + \frac{1}{\hat{\pi}_{21}} + \frac{1}{\hat{\pi}_{22}}}. \quad (5.6)$$

- We can construct a confidence interval for  $\log R$  given by

$$[\log \hat{R} - z_{\alpha/2} \text{s.e.}(\log \hat{R}), \log \hat{R} + z_{\alpha/2} \text{s.e.}(\log \hat{R})],$$

where  $z_{\alpha/2}$  is the the normal quantile corresponding to the  $(1 - \alpha)$  confidence level.

# Sampling Methods

For multinomial sampling, the likelihood function for a  $2 \times 2$  contingency table is given by

$$\prod_{i=1}^2 \prod_{j=1}^2 \pi_{ij}^{n_{ij}}, \quad \pi_{ij} \geq 0 \text{ and } \sum_{i=1}^2 \sum_{j=1}^2 \pi_{ij} = 1. \quad (5.7)$$

Let  $\theta = (\pi_{11}, \pi_{12}, \pi_{21})^T$ . Then the log-likelihood function is given by

$$\ell_n(\theta) = n_{11} \log \pi_{11} + n_{12} \log \pi_{12} + n_{21} \log \pi_{21} + n_{22} \log(\pi_{22}),$$

where  $\sum_{i=1}^2 \sum_{j=1}^2 \pi_{ij} = 1$ .

# Sampling Methods

The maximum likelihood estimates of  $\pi_{ij}$  are given by  $\hat{\pi}_{ij} = n_{ij}/n$ . Let  $\pi = (\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22})^T$ . Since

$$\log \hat{R} = g(\hat{\pi}) = \log \hat{\pi}_{11} + \log \hat{\pi}_{22} - \log \hat{\pi}_{21} - \log \hat{\pi}_{12},$$

the variance of  $g(\hat{\pi})$  can be approximated by

$$[\partial_\pi g(\pi)]^T n^{-1} [\text{diag}(\pi) - \pi\pi^T] [\partial_\pi g(\pi)] = \sum_{i=1}^2 \sum_{j=1}^2 \frac{1}{n\pi_{ij}}.$$

## Example 5.6

For Table 4.7, we can estimate  $\widehat{\text{odds}}(P(Y = 2|X = 1) = 4)$  and  $\widehat{\text{odds}}(P(Y = 2|X = 2)) = 9$ . Thus,  $\hat{R} = 9/4 = 2.25$  and the standard error of  $\hat{R}$  is 0.7808. The p-value for testing independence between  $X$  and  $Y$  is 0.3368. Similarly, we can define the relative risk of  $X = 2$  over  $X = 1$  for  $Y = 2$  (or  $Y = 1$ ).

Particularly,  $\widehat{\text{odds}}(P(X = 2|Y = 1)) = 1.5$  and  $\widehat{\text{odds}}(P(X = 2|Y = 2)) = 3.375$ . How about the odds ratio of two odds:  $\widehat{\text{odds}}(P(X = 2|Y = 2))$  and  $\widehat{\text{odds}}(P(X = 2|Y = 1))$ ?

# Sampling Methods

In **purposive sampling** studies, we have a predetermined  $n_1$  subjects having a particular characteristic and a predetermined  $n_2$  subjects not having a particular characteristic. Within each group, we measure another categorical variable for all subjects.

Table: Table 4.9: Purposive Sampling

		Y=Disease Status		Total
		1=No	2=Yes	
X=Exposure Status	1=No	$n_{11}$	$n_{12}$	$n_{1\cdot}$
	2=Yes	$n_{21}$	$n_{22}$	$n_{2\cdot}$
Total		$n_{\cdot 1}$	$n_{\cdot 2}$	n

# Sampling Methods

- A comparative prospective study or cohort, or a forward-going, or follow-up study is characterized by identifying two (or more) study samples on the basis of an antecedent factor (e.g., exposure status) and by predicting the development of disease (or condition) using the antecedent factor under study.
- In a prospective study, a group of  $n_2$  exposed subjects is selected together with a group of  $n_1$  unexposed subjects. Then, both groups are followed up for a prolonged period. At the end of study, the proportions of diseased subjects are compared in the two groups.

# Sampling Methods

For a prospective study, we can only estimate  $P(Y|X)$  and measure the risk of the presence of the disease in the exposed subjects as follows:

$$odds(P(Y = 2|X = 2)) = \frac{P(Y = 2|X = 2)}{P(Y = 1|X = 2)} \approx \frac{n_{22}/n_2}{n_{21}/n_2} = \frac{n_{22}}{n_{21}}.$$

Similarly, in the unexposed subjects, the risk of the presence of the disease can be estimated by

$$odds(P(Y = 2|X = 1)) = \frac{P(Y = 2|X = 1)}{P(Y = 1|X = 1)} \approx \frac{n_{12}/n_1}{n_{11}/n_1} = \frac{n_{12}}{n_{11}}.$$

# Sampling Methods

$$R = \frac{\text{odds}(P(Y=2|X=2))}{\text{odds}(P(Y=2|X=1))} \approx \hat{R} = \frac{n_{22}n_{11}}{n_{21}n_{12}}. \quad (5.8)$$

The standard error of  $\hat{R}$  is given by

$$\text{s.e.}(\hat{R}) \approx \hat{R} \sqrt{\frac{1}{n_1 \cdot \hat{\pi}_{Y=2|X=1} \hat{\pi}_{Y=1|X=1}} + \frac{1}{n_2 \cdot \hat{\pi}_{Y=2|X=2} \hat{\pi}_{Y=1|X=2}}}, \quad (5.9)$$

where  $\hat{\pi}_{Y=i|X=j} = n_{i,j}/n_j$ , for all  $i, j = 1, 2$ .

# Sampling Methods

For retrospective studies, we can only estimate  $P(X|Y)$  and only measure the risk of exposure in the disease subjects as

$$odds(P(X = 2|Y = 2)) = \frac{P(X = 2|Y = 2)}{P(X = 1|Y = 2)} \approx \frac{n_{22}/n_{\cdot 2}}{n_{12}/n_{\cdot 2}} = \frac{n_{22}}{n_{12}}.$$

Similarly, in the disease-free subjects, the risk of exposure can be estimated by

$$odds(P(X = 2|Y = 1)) = \frac{P(X = 2|Y = 1)}{P(X = 1|Y = 1)} \approx \frac{n_{21}/n_{\cdot 1}}{n_{11}/n_{\cdot 1}} = \frac{n_{21}}{n_{11}}.$$

# Sampling Methods

The odds ratio of two odds,  $odds(P(X = 2|Y = 2))$  and  $odds(P(X = 2|Y = 1))$  is defined and estimated by

$$R = \frac{odds(P(X = 2|Y = 2))}{odds(P(X = 2|Y = 1))} \approx \hat{R} = \frac{n_{22}n_{11}}{n_{21}n_{12}}. \quad (5.10)$$

The standard error of  $\hat{R}$  is given by

$$\text{s.e.}(\hat{R}) \approx \hat{R} \sqrt{\frac{1}{n_{\cdot,1}\hat{\pi}_{X=2|Y=1}\hat{\pi}_{X=1|Y=1}} + \frac{1}{n_{\cdot,2}\hat{\pi}_{X=2|Y=2}\hat{\pi}_{X=1|Y=2}}}, \quad (5.11)$$

where  $\hat{\pi}_{X=i|Y=j} = n_{i,j}/n_{\cdot,j}$  for all  $i, j = 1, 2$ .

An important property of the odds ratio is that it is invariant across the three kinds of studies: cross-sectional, prospective, and retrospective studies. Moreover, the sample odds ratio  $\hat{R} = n_{22}n_{11}/(n_{21}n_{12})$  estimates  $R$  and has the same standard error in each case.

# Sampling Methods

Table: Table 4.10: Retrospective Sampling

		Y=Blood Type		Total
		O	A	
X=Exposure Status	Peptic ulcer	911	579	1490
	Control	4578	4219	8797
Total		5489	4798	10287

## Example 5.7

*Consider the incidence of peptic ulcers from London. We reorganized the data and presented them in Table 4.10. The sample odd ratio  $R$  is 1.45 and its associated standard deviation is 0.083. Thus, the  $p$ -value for testing independence between  $X$  and  $Y$  is smaller than 0.0001.*

# Sampling Methods

```
London<-cbind(c(911, 579), c(4578, 4219))
chisq.test(London, correct=TRUE)
chisq.test(London, correct=FALSE)
```

Pearson's Chi-squared test with Yates' continuity correction data:

London X-squared = 42.0369, df = 1, p-value = 8.957e-11

Pearson's Chi-squared test data: London X-squared = 42.4018, df = 1, p-value = 7.432e-11

In R, the command 'chisq.test' is used to carry out Pearson's test and the 'correct' denotes the option of applying Yates' continuity correction or not. The results showed that **exposure status is significantly associated with blood type in London.**

# Sampling Methods

```
lor <- oddsratio(London)
```

```
summary(lor)
```

```
Log Odds Ratio Std. Error z value Pr(> |z|)
```

```
[1] 0.371576 0.057255 6.4899 4.296e-11 *** —
```

In R, the command 'oddsratio' is used to calculate the logarithm of odd ratio and associated statistics. Thus, exposure status is significantly associated with blood type in London.

# Sampling Methods

```
Blood<-c("Group O","Group A")
City<-c("London", "Manchester", "Newcastle")
Substance<-c("PepticUlcer", "control")
datalabel<-list(Blood, Substance, City)
table.4.5<-expand.grid(Blood, Substance, City)
data<-c(911, 579, 4578, 4219, 361, 246, 4532, 3775, 396, 219, 6598, 5261)
table.4.5<-cbind(table.4.5, freq=data)
names(table.4.5)<-c('Blood', 'Substance', 'City', 'freq')
newtable<-xtabs(freq~Blood+Substance+City, data=table.4.5)
lor<-oddsratio(newtable)
confint(lor)
plot(lor, xlab="City", main="Substance and Blood")
```

# Sampling Methods

summary(lor)

Log Odds Ratio Std. Error z value Pr(> |z|)

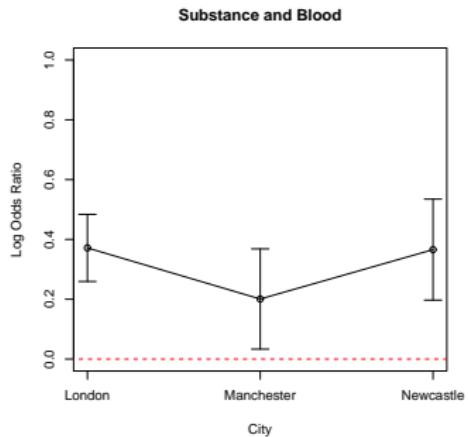
London 0.371576 0.057255 6.4899 4.296e-11 \*\*\*

Manchester 0.200783 0.085490 2.3486 0.009422 \*\*

Newcastle 0.365897 0.086136 4.2479 1.079e-05 \*\*\*

In London, Manchester, and Newcastle, exposure status is significantly associated with blood type.

# Sampling Methods



# Logistic Regression

Suppose that  $Z_i$  is a binary response taking values in  $\{0, 1\}$  and  $\mathbf{x}_i$  is its associated covariate vector for  $i = 1, \dots, n$ .

- Logistic regression models the binary response as a function of the covariates:  $\pi(\mathbf{x}_i) = P(Z_i = 1|\mathbf{x}_i)$ .
- In many studies, there are  $N$  of possible combinations  $\mathbf{x}_i$  and we observe  $n_i$   $Z_i$ 's for each combination of  $\mathbf{x}_i$ . Then, we obtain the number of successes  $\sum_{i=1}^{n_i} Z_i = Y_i$  and the total number of binary responses  $n_i$  for a given  $\mathbf{x}_i$ . We obtain grouped data or binomial responses, called covariate classes.

## Definition 1

Logistic regression assumes

- (i)  $y_i | \mathbf{x}_i \sim B(n_i, \pi(\mathbf{x}_i))$  for  $i = 1, \dots, N$ ;
- (ii)  $\pi_i = \pi(\mathbf{x}_i)$  is related to  $\mathbf{x}_i$  by

$$g_1(\pi_i) = \text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1 - \pi_i}\right) = \mathbf{x}_i^T \boldsymbol{\beta} \quad (5.12)$$

for  $i = 1, \dots, N$ , where  $\mathbf{x}_i$  is a  $p \times 1$  covariate vector and  $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$  is defined in a subset  $\mathcal{B}$  of  $R^p$  ( $p < n$ ).

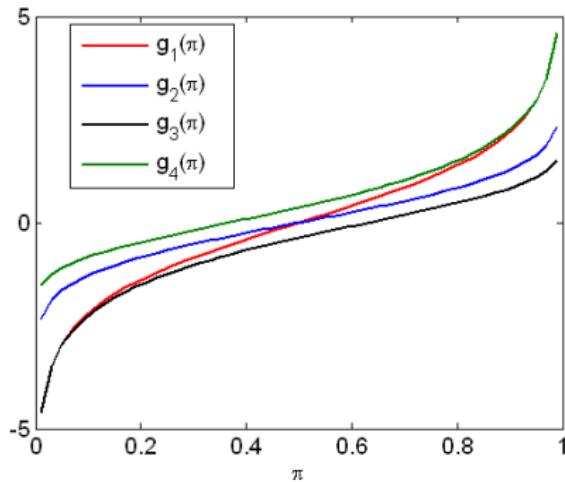
# Logistic Regression

We consider four link functions:

- $g_1(\pi) = \log(\pi/(1 - \pi))$  (logit link),
- $g_2(\pi) = \Phi^{-1}(\pi)$  (probit link or inverse Normal function),
- $g_3(\pi) = \log\{-\log(1 - \pi)\}$  (complementary log-log function),
- $g_4(\pi) = -\log\{-\log(\pi)\}$  (log-log function).

# Logistic Regression

Figure: Figure 4.2: A graphical comparison of four link functions



## Comparisons

- The logit and probit links are almost linearly related over the interval [0.1, 0.9].
- The regression coefficients of the logistic regression model can be interpreted as the logarithm of the odds ratio.
- the use of logit link leads to a simple analysis for data from retrospective studies.

# Logistic Regression

For logistic regression,  $\ell_n(\beta)$  is given by

$$\ell_n(\beta) = \sum_{i=1}^N \left[ y_i \mathbf{x}_i^T \beta - n_i \log(1 + \exp(\mathbf{x}_i^T \beta)) \right].$$

The first and second derivatives of  $\ell_n(\beta)$  are, respectively, given by

$$\partial_\beta \ell_n(\beta) = \sum_{i=1}^N (y_i - n_i \pi_i) \mathbf{x}_i \quad \text{and} \quad -\partial_\beta^2 \ell_n(\beta) = \sum_{i=1}^N n_i \pi_i (1 - \pi_i) \mathbf{x}_i^{\otimes 2},$$

where  $\pi_i = \exp(\mathbf{x}_i^T \beta) / [1 + \exp(\mathbf{x}_i^T \beta)]$ . The Newton-Raphson algorithm is given by

$$\beta^{k+1} = \beta^k + \{-\partial_\beta^2 \ell_n(\beta^k)\}^{-1} \partial_\beta \ell_n(\beta^k). \quad (5.13)$$

# Logistic Regression

The deviance function is given by

$$D(\mathbf{y}; \hat{\pi}) = 2 \sum_{i=1}^N \left\{ y_i \log \left( \frac{y_i}{\hat{\pi}_i} \right) + (n_i - y_i) \log \left( \frac{n_i - y_i}{n_i - \hat{\pi}_i} \right) \right\}, \quad (5.14)$$

where  $\hat{\pi}_i = 1/(1 + \exp(-\mathbf{x}_i^T \hat{\beta}))$ . Finally, we get

$$\left\{ n^{-1} \sum_{i=1}^N n_i \pi_i(\beta_*) [1 - \pi_i(\beta_*)] \mathbf{x}_i^{\otimes 2} \right\}^{-1/2} \sqrt{n} (\hat{\beta} - \beta_*) \xrightarrow{L} N(0, \mathbf{I}_p).$$

# Logistic Regression

In logistic regression,  $Y$  can be regarded as the response and  $X$  is the only covariate, say exposure status. Thus,

$$\text{logit}(P(Y = 1|X)) = \beta_1 + \beta_2 X.$$

Furthermore, the log-odds for  $X = 1$  and  $X = 0$  are given by

$$\text{logit}(P(Y = 1|X = 1)) = \beta_1 + \beta_2 \times 1$$

and  $\text{logit}(P(Y = 1|X = 0)) = \beta_1 + \beta_2 \times 0$ , respectively. It follows that

$$R_{X=1 \text{ vs. } X=0} = \frac{\exp(\beta_1 + \beta_2)}{\exp(\beta_1)} = \exp(\beta_2).$$

The odds ratio comparing the two categories of the covariate can be obtained by exponentiating the coefficient of the covariate in the logistic regression model.

# Logistic Regression

We can define  $X_A = (x_{A1}, \dots, x_{Ap})^T$  and  $X_B = (x_{B1}, \dots, x_{Bp})^T$  for groups  $A$  and  $B$ , respectively. Then, the odds ratio for comparing groups  $A$  and  $B$  is given by

$$R_{A \text{ vs. } B}(\beta) = \frac{\text{odds}(P(Y = 1|X_A))}{\text{odds}(P(Y = 1|X_B))} = \exp((X_A - X_B)^T \beta). \quad (5.15)$$

The standard deviation of  $R_{A \text{ vs. } B}(\hat{\beta})$  is approximated by

$$s.d.(R_{A \text{ vs. } B}) = R_{A \text{ vs. } B}(\beta_*) \sqrt{(X_A - X_B)^T \text{Cov}(\hat{\beta})(X_A - X_B)}.$$

# Logistic Regression

Suppose  $X_A = (x_{A1}, x_{A2}, x_{A3}, \dots, x_{Ap})$  and  
 $X_B = (x_{A1}, x_{A2} + c, x_{A3}, \dots, x_{Ap})$ , where  $c \neq 0$  is a constant.  
Then,  $R_A \text{ vs. } B(\beta) = \exp(-c\beta_2)$  and  $\beta_2 = -c^{-1} \log\{R_A \text{ v } B(\beta)\}$ .

## Example 5.8

We consider an example as follows (Kleinbaum, Kupper, Muller, and Nizam, 1998). If  $p = 4$ ,  $x_1$  is the intercept,  $x_2$  is smoking status (1=yes, 0=no),  $x_3$  is age (continuous), and  $x_4$  is race (1=black, 0=white). We specify  $X_A = (1, 0, 30, 1)$  and  $X_B = (1, 0, 30, 0)$ . Thus,  $X_A$  denotes the group of 30-year-old black nonsmokers, and  $X_B$  denotes the group of 30-year-old white nonsmokers. We have

$$R_{A \text{ vs. } B} = \exp(\hat{\beta}_4) \text{ and } s.d.(R_{A \text{ vs. } B}) \approx \exp(\hat{\beta}_4) \sqrt{\text{var}(\hat{\beta}_4)}.$$

## Model Selection

- **Selection Criteria:** Goodness of fit + Complexity
  - Deviance
  - Akaike Information Criterion (AIC)
  - Bayesian Information Criterion (BIC)
- **Selection Methods**
  - forward selection
  - backward elimination
  - exhaustive search

## Example 5.9

*Consider the vegetation data in Example 3.4. First, we are interested in using logistic regression with different link functions to model the relationship between the distribution of V2 and the climate variables  $X_1, \dots, X_5$ . Using the `glm` function in R, the probit link gives the best fit. Furthermore, we eliminate covariates at significance level  $\alpha = 5\%$ .*

# Logistic Regression

```
outbim<-glm(V3~V4+V5+V6+V7+V8, family=binomial, data=aa)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
--	----------	------------	---------	----------

(Intercept)	1.5509809	1.1044570	1.404	0.160232
-------------	-----------	-----------	-------	----------

V4	0.0010865	0.0214041	0.051	0.959515
----	-----------	-----------	-------	----------

V5	-0.0011755	0.0004758	-2.471	0.013486 *
----	------------	-----------	--------	------------

V6	-0.0026909	0.0068830	-0.391	0.695832
----	------------	-----------	--------	----------

V7	-0.0032304	0.0027847	-1.160	0.246029
----	------------	-----------	--------	----------

V8	-0.0010959	0.0002910	-3.766	0.000166 ***
----	------------	-----------	--------	--------------

Null deviance: 861.89 on 706 degrees of freedom

Residual deviance: 793.36 on 701 degrees of freedom AIC: 805.36

# Logistic Regression

```
outbim<-glm(V3~V4+V5+V6+V7+V8,  
family=binomial(link=probit),data=aa)
```

Coefficients:

Estimate Std. Error z value Pr(> |z|)

(Intercept) 1.0890300 0.6482115 1.680 0.09295 .

V4 0.0032950 0.0125820 0.262 0.79341

V5 -0.0007700 0.0002803 -2.747 0.00601 \*\*

V6 -0.0013974 0.0040450 -0.345 0.72974

V7 -0.0017399 0.0016176 -1.076 0.28213

V8 -0.0006887 0.0001700 -4.051 5.1e-05 \*\*\*

Null deviance: 861.89 on 706 degrees of freedom

Residual deviance: 790.96 on 701 degrees of freedom AIC: 802.96

# Logistic Regression

```
outbim<-glm(V3~V4+V5+V6+V7+V8,  
family=binomial(link=cloglog),data=aa)
```

Coefficients:

Estimate Std. Error z value Pr(> |z|)

(Intercept) 0.3475258 0.9146781 0.380 0.70399

V4 -0.0053276 0.0178872 -0.298 0.76582

V5 -0.0008145 0.0003917 -2.079 0.03758 \*

V6 -0.0015372 0.0057070 -0.269 0.78766

V7 -0.0028326 0.0023577 -1.201 0.22958

V8 -0.0007582 0.0002382 -3.182 0.00146 \*\*

Null deviance: 861.89 on 706 degrees of freedom

Residual deviance: 798.36 on 701 degrees of freedom AIC: 810.36

# Logistic Regression

- In **retrospective sampling**, one starts from a group of cases ( $Y = 1$ ) and controls ( $Y = 0$ ), and then follows up the case and control groups independently to observe the covariates  $\mathbf{x}$ . We can occasionally model  $P(\mathbf{x}|Y)$  instead of  $P(Y|\mathbf{x})$ .
- In **cross-sectional and prospective studies**, the logistic regression model can be used to model  $P(Y = 1|\mathbf{x})$ .
- One often **ignores the case-control sampling scheme** and just fits the logistic regression model for  $P(Y = 1|\mathbf{x})$ .

An important issue is whether directly applying the logistic regression model and its associated inference (e.g., estimation and standard errors) to the retrospective study is correct.

- Prentice and Pyke (1979): one can ignore the study design and use estimation and inference based on a prospective study.
- Carroll, Wang, and Wang (1995): generalize this prospective formulation of case-control studies to the cases of multiplicative models, stratification, missing data, measurement errors, and robustness.

# Logistic Regression

The key point is that the coefficient of a particular covariate is associated with the odds ratio of the covariate, which is invariant with prospective and retrospective studies.

- Let  $S$  be the selection indicator, where  $S = 1$  denotes the inclusion of a subject in the case-control study. The likelihood function for the retrospective study is  $p(\mathbf{x}|Y, S = 1)$ .
- To avoid biased sampling, we assume that  $P(S = 1|X, Y) = P(S = 1|Y)$ .
- $p(\mathbf{x}|Y, S = 1) = \frac{P(S=1|\mathbf{x}, Y)p(\mathbf{x}|Y)}{P(S=1|Y)} = p(\mathbf{x}|Y)$ .

# Logistic Regression

$$\begin{aligned} p(\mathbf{x}|Y, S=1) &= \frac{P(Y|\mathbf{x}, S=1)p(\mathbf{x}|S=1)}{P(Y|S=1)} \\ p(Y|\mathbf{x}, S=1) &= \frac{p(X|Y, S=1)p(Y, S=1)}{p(X, S=1)} = p(Y|\mathbf{x}) \frac{p(S=1|Y)}{p(S=1|\mathbf{x})}. \end{aligned}$$

We have

$$p(\mathbf{x}|Y, S=1) = \frac{P(Y|\mathbf{x})p(\mathbf{x}|S=1)P(S=1|Y)}{P(Y|S=1)p(S=1|\mathbf{x})} = P(Y|\mathbf{x})H_1(Y)H_2(\mathbf{x}),$$

in which the ratio  $p(\mathbf{x}|S=1)/p(S=1|\mathbf{x})$  only depends on  $\mathbf{x}$ ,  $P(S=1|Y)$  and  $P(Y|S=1)$  are determined by the sampling scheme.

# Logistic Regression

Thus,

$$\text{odds}[p(\mathbf{x}|Y=1, S=1)] = \text{odds}[P(Y=1|\mathbf{x})] \left[ \frac{H_1(Y=1)}{H_1(Y=0)} \right]. \quad (5.16)$$

If one considers two specifications of  $\mathbf{x}$ :  $\mathbf{x}_A$  and  $\mathbf{x}_B$ , then

$$R_{A \text{ vs. } B} = \frac{\text{odds}[p(\mathbf{x}_A|Y=1, S=1)]}{\text{odds}[p(\mathbf{x}_B|Y=1, S=1)]} = \frac{\text{odds}[P(Y=1|\mathbf{x}_A)]}{\text{odds}[P(Y=1|\mathbf{x}_B)]}. \quad (5.17)$$

# Logistic Regression

If we specify a parametric model  $P(y_i|\mathbf{x}_i, \beta)$  and  $H_1(Y)H_2(\mathbf{x})$  is independent of  $\beta$ , then we have

$$\prod_{i=1}^n p(\mathbf{x}_i|y_i, S_i = 1) = \prod_{i=1}^n P(y_i|\mathbf{x}_i, \beta) \prod_{i=1}^n [H_1(y_i)H_2(\mathbf{x}_i)]. \quad (5.18)$$

If a logistic regression model is assumed for  $P(y_i|\mathbf{x}_i, \beta)$ , then  $\hat{\beta}$  can be directly applied to the retrospective study. In addition, we have

$$R_A \text{ vs. } B = \frac{\text{odds}[p(\mathbf{x}_A|Y = 1, S = 1)]}{\text{odds}[p(\mathbf{x}_B|Y = 1, S = 1)]} = \exp((\mathbf{x}_A - \mathbf{x}_B)^T \hat{\beta}).$$

# Chapter 6: Generalized Linear Models for Categorical Responses: II

- Models for Polytomous and ordinal responses
  - generalized logistic regression models
  - proportional odds models
- Poisson regression
- Loglinear models for contingency tables

# Models for Polytomous and Ordinal Responses

We are interested in using a set of covariates of interest to model polytomous or ordinal responses.

- Ordinal Responses: tumor grade (well differentiated, moderately differentiated, poor differentiated); disease symptoms (absent, mild, moderate, severe).
- Polytomous Responses: breast cancer subtypes; drug types (placebo, drug A, drug B).

## Modeling Set-up

- $Z_i \in \{1, \dots, I\}$  for  $i = 1, \dots, n$ ;
- $Z_i \Rightarrow (z_{i1}, \dots, z_{il})$  and  $Z_i = j \Leftrightarrow z_{ij} = 1, z_{ik} = 0$  for all  $k \neq j$ ;
- $\pi(\mathbf{x}_i) = (E(z_{i1}|\mathbf{x}_i), \dots, E(z_{il}|\mathbf{x}_i))^T = (\pi_1(\mathbf{x}_i, \beta), \dots, \pi_l(\mathbf{x}_i, \beta))^T$ , where  $\sum_{j=1}^l \pi_j(\mathbf{x}_i, \beta) = 1$  for all  $\beta \in \mathcal{B} \in R^p$ ;
- We can group  $Z_i$  according to  $N$  possible combinations of  $\mathbf{x}_i$  into  $Y_i = \sum_{k=1}^{n_i} (z_{k1}, \dots, z_{kI})$ .
- $Y_i | \mathbf{x}_i \sim Multi(n_i; \pi(\mathbf{x}_i))$  for  $i = 1, \dots, N$ .

## Definition 6.1

*Multicategorical logit models* assume

- (i)  $Y_i | \mathbf{x}_i \sim \text{Multi}(n_i; \pi_1(\mathbf{x}_i), \dots, \pi_I(\mathbf{x}_i))$  for  $i = 1, \dots, N$ ;
- (ii)  $\pi_{ij} = \pi_j(\mathbf{x}_i)$  is related to  $\mathbf{x}_i$  by

$$\pi_j(\mathbf{x}_i, \beta) = \frac{\exp(\mathbf{x}_i^T \beta_j)}{\sum_{j=1}^I \exp(\mathbf{x}_i^T \beta_j)} \quad (6.1)$$

for  $j = 1, \dots, I$  and  $i = 1, \dots, N$ , where  $\mathbf{x}_i$  is a  $p \times 1$  covariate vector, and  $\beta_j = (\beta_{j1}, \dots, \beta_{jp})^T \in \mathcal{B} \subset \mathcal{R}^p$  ( $p < n$ ) for  $j = 1, \dots, I$ .

# Models for Polytomous and Ordinal Responses

- Choose a referent category and compare other categories with the referent category.
- Parameter identification:  $\log \frac{P(Z=j|\mathbf{x}_i)}{P(Z=1|\mathbf{x}_i)} = \mathbf{x}_i^T (\beta_j - \beta_1)$  for  $j = 2, \dots, I$ .
- $\beta_j - \beta_1$ , ( $j = 2, \dots, I$ ) are estimable.
- Identifiability: we set  $\beta_1 = 0$ .

# Models for Polytomous and Ordinal Responses

- For  $j = 2, \dots, I$ , we have  $\log \frac{P(Z=j|\mathbf{x}_i)}{P(Z=1|\mathbf{x}_i)} = \mathbf{x}_i^T \beta_j$ , in which  $\beta_j$  determines the **log odds for category  $j$  with respect to the referent category 1**. Thus, we can obtain  $I - 1$  odds ratios for the polytomous response.
- If we compare two groups with  $\mathbf{x}_A$  and  $\mathbf{x}_B$  for all odds ratios, then in the  $j$ th category, we have

$$R_{A \text{ vs. } B}(j) = \frac{P(Z = j|\mathbf{x}_A)P(Z = 1|\mathbf{x}_B)}{P(Z = 1|\mathbf{x}_A)P(Z = j|\mathbf{x}_B)} = \exp((\mathbf{x}_A - \mathbf{x}_B)^T \beta_j). \quad (6.2)$$

# Models for Polytomous and Ordinal Responses

## Estimation of $\beta$

- $\ell_n(\beta) = \sum_{i=1}^N \left[ \sum_{j=2}^I y_{ij} \mathbf{x}_i^T \beta_j - n_i \log(1 + \sum_{j=2}^I \exp(\mathbf{x}_i^T \beta_j)) \right].$
- $\sum_{i=1}^N y_{ij} \mathbf{x}_i$  is a sufficient statistic for  $\beta$ .
- $\partial_{\beta_j} \ell_n(\beta) = \sum_{i=1}^N (y_{ij} - n_i \pi_{ij}) \mathbf{x}_i$
- $-\partial_{\beta_j}^2 \ell_n(\beta) = \sum_{i=1}^N n_i \pi_{ij} (1 - \pi_{ij}) \mathbf{x}_i^{\otimes 2},$   
 $-\partial_{\beta_j \beta_k}^2 \ell_n(\beta) = \sum_{i=1}^N n_i \pi_{ij} \pi_{ik} \mathbf{x}_i^{\otimes 2}, \text{ for } j \neq k, \text{ where}$   
 $\pi_{ij} = \exp(\mathbf{x}_i^T \beta_j) / [1 + \sum_{j=2}^I \exp(\mathbf{x}_i^T \beta_j)].$

Disadvantages of dichotomizing a polytomous or ordinal outcome into multiple binary outcomes:

- we wind up fitting multiple dichotomous logistic regression models
- we reduce the power in detecting significant covariates, particularly with the presence of continuous covariates
- it affects all possible comparisons between any two different categories.

## Example 6.1

*Consider a study of factors influencing the primary food choice of alligators. The data consist of 219 alligators. For each alligator, the response is the primary food type, which includes five categories (fish, invertebrate, reptile, bird, other) found in its stomach. The covariates of interest include  $L$  = lake of capture (Hancock, Oklawaha, Trafford, George),  $G$  = gender (male, female), and  $S$  = size ( $\leq 2.3$  meters long,  $> 2.3$  meters long).*

# Models for Polytomous and Ordinal Responses

```
> food.labs<-factor(c("fish","invert","rep","bird","other"),
  levels=c("fish","invert", "rep", "bird","other"))
> size.labs<-factor(c("<2.3",">2.3"),levels=c(">2.3","<2.3"))
> gender.labs<-factor(c("m","f"),levels=c("m","f"))
> lake.labs<-factor(c("hancock","oklawaha","trafford","george"),
  levels=c("george", "hancock", "oklawaha","trafford"))
> table.7.1<-expand.grid(food=food.labs,size=size.labs, gender=gender.labs,
  lake=lake.labs)
> temp<-c(7,1,0,0,5,4,0,0,1,2,16,3,2,2,3,3,0,1,2,3,2,2,0,0,1,
  13,7,6,0,0,3,9,1,0,2,0,1,0,1, 0,3,7,1,0,1,8,6,6,3,5,2,4,1,1, 4,0,1,0,0,0,13,10,
  0,2,2,9,0,0,1,2,3,9,1,0,1,8,1,0,0,1)
> table.7.1<-structure(.Data=table.7.1[rep(1:nrow(table.7.1),temp),],
  row.names=1:219)
```

# Models for Polytomous and Ordinal Responses

```
> fitS<-multinom(food~lake*size*gender,data=table.7.1)
> fit0<-multinom(food~1,data=table.7.1)
> fit1<-multinom(food~gender,data=table.7.1)
> fit2<-multinom(food~size,data=table.7.1)
> fit3<-multinom(food~lake,data=table.7.1)
> fit4<-multinom(food~size+lake,data=table.7.1)
> fit5<-multinom(food~size+lake+gender,data=table.7.1)
> deviance(fit1)-deviance(fitS) [1] 114.6571
> deviance(fit2)-deviance(fitS) [1] 101.6116
> deviance(fit3)-deviance(fitS) [1] 73.56589
> deviance(fit4)-deviance(fitS) [1] 52.47848
> deviance(fit5)-deviance(fitS) [1] 50.26368
> deviance(fit0)-deviance(fitS) [1] 116.7611
```

# Models for Polytomous and Ordinal Responses

```
> fitS<-multinom(food~lake*size,data=table.7.1)
> fit0<-multinom(food~1,data=table.7.1)
> fit1<-multinom(food~size,data=table.7.1)
> fit2<-multinom(food~lake,data=table.7.1)
> fit3<-multinom(food~size+lake,data=table.7.1)
> deviance(fit1)-deviance(fitS) [1] 66.2129
> deviance(fit2)-deviance(fitS) [1] 38.16724
> deviance(fit3)-deviance(fitS) [1] 17.07983
> deviance(fit0)-deviance(fitS) [1] 81.36247
```

# Models for Polytomous and Ordinal Responses

```
> library(MASS)
> summary(fit3, cor = F)
(Intercept) size<2.3 lakehancock lakeoklawaha laketrafford
invert -1.549021 1.4581457 -1.6581178 0.937237973 1.122002
rep -3.314512 -0.3512702 1.2428408 2.458913302 2.935262
bird -2.093358 -0.6306329 0.6954256 -0.652622721 1.088098
other -1.904343 0.3315514 0.8263115 0.005792737 1.516461
invert 0.4249185 0.3959418 0.6128466 0.4719035 0.4905122
rep 1.0530577 0.5800207 1.1854031 1.1181000 1.1163844
bird 0.6622972 0.6424863 0.7813123 1.2020025 0.8417085
other 0.5258313 0.4482504 0.5575446 0.7765655 0.6214371
```

# Models for Polytomous and Ordinal Responses

- $Z_i \in \{1, \dots, I\}$  for  $i = 1, \dots, n$ ;
- $Z_i \Rightarrow (z_{i1}, \dots, z_{il})$  and  $Z_i = j \Leftrightarrow z_{ij} = 1, z_{ik} = 0$  for all  $k \neq j$ ;
- $\pi(\mathbf{x}_i) = (E(z_{i1}|\mathbf{x}_i), \dots, E(z_{il}|\mathbf{x}_i))^T = (\pi_1(\mathbf{x}_i, \beta), \dots, \pi_l(\mathbf{x}_i, \beta))^T$ , where  $\sum_{j=1}^I \pi_j(\mathbf{x}_i, \beta) = 1$  for all  $\beta \in \mathcal{B} \in R^p$ ;

## Definition 6.2

The *ordinal regression model* assumes

- (i)  $Z_i | \mathbf{x}_i \sim \text{Multi}(1; \pi_1(\mathbf{x}_i), \dots, \pi_I(\mathbf{x}_i))$  for  $i = 1, \dots, n$ ;
- (ii)  $P(Z_i \leq j | \mathbf{x}_i) = \sum_{k=1}^j \pi_k(\mathbf{x}_i)$  is related to  $\mathbf{x}_i$  by

$$g(P(Z_i \leq j | \mathbf{x}_i)) = \alpha_j + \mathbf{x}_i^T \beta \quad (6.3)$$

for  $j = 1, \dots, I - 1$  and  $i = 1, \dots, n$ , where  $\alpha_1 \leq \dots \leq \alpha_{I-1}$ ,  $g(\cdot)$  is an increasing link function,  $\mathbf{x}_i$  is a  $p \times 1$  covariate vector, and  $\beta = (\beta_1, \dots, \beta_p)^T$  is defined in a subset  $\mathcal{B}$  of  $R^p$  ( $p < n$ ).

## Four link functions:

- $g_1(t) = \log(t/(1-t))$  (logit link) **proportional odds model**
- $g_2(t) = \Phi^{-1}(t)$  (probit link or inverse Normal function)
- $g_3(t) = \log\{-\log(1-t)\}$  (complementary log-log link)
- $g_4(t) = -\log\{-\log(t)\}$  (log-log link)

# Ordinal Regression Models

The proportional odds model assumes that

$$\text{logit}[P(Z \leq j|\mathbf{x})] = \alpha_j + \mathbf{x}^T \boldsymbol{\beta}, \text{ for } j = 1, \dots, I - 1.$$

- Each cumulative logit has its own intercept
- $\alpha_1 \leq \dots \leq \alpha_{I-1}$ ,  $P(Z \leq 1|\mathbf{x}) \leq \dots \leq P(Z \leq I|\mathbf{x})$ .
- The odds ratio is invariant to how we dichotomize the ordinal response  $\frac{\text{odds}(P(Z \leq j|\mathbf{x}_1))}{\text{odds}(P(Z \leq j|\mathbf{x}_2))} = \exp((\mathbf{x}_1 - \mathbf{x}_2)^T \boldsymbol{\beta})$  hold for all  $j$ .

## Example 6.2

*Consider the  $3 \times 2$  contingency table given in Table 5.1. The two variables are, respectively, Tumor Grade (Well, Good, and Poor) and Gender (Male and Female). Our interest is to use Gender to predict Tumor Grade. The 'Tumor Grade' is an ordinal response. To use the proportional odds model, we must consider the possible  $2 \times 2$  tables that preserve the natural ordering of 'Tumor Grade'. In Table 5.1, we calculated the two odds ratios as 2.0 and 2.06, which are close to each other. Thus, the assumption of a proportional odds model may be reasonable.*

# Ordinal Regression Models

Table: Table 5.1: Data

Tumor		x	
Grade	Male	Female	Total
Well	200	300	500
Good	300	700	1000
Poor	250	1000	1250
Total	750	2000	2750

# Ordinal Regression Models

Table: Table 5.2: Collapsed  $2 \times 2$  tables

Tumor Grade	Male	Female	x Total
Well or Good	500	1000	1500
Poor	250	1000	1250
Total	750	2000	2750

Tumor Grade	Male	Female	x Total
Well	200	300	500
Good or Poor	550	1700	2250
Total	750	2000	2750

# Ordinal Regression Models

## Latent variable models:

- $Z_i^* = \text{a continuous latent variable for each } Z_i.$
- $Z_i = j \text{ if } \alpha_{j-1} < Z_i^* \leq \alpha_j, \text{ for } j = 1, \dots, I \text{ and } i = 1, \dots, n,$  where  $-\infty = \alpha_0 < \alpha_1 < \dots < \alpha_I = \infty.$
- $Z_i^* = \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i,$  where  $\epsilon_i$  denotes a random error with mean zero and cumulative distribution function  $G(\cdot)$
- $P(Z_i \leq j | \mathbf{x}) = P(Z_i^* \leq \alpha_j | \mathbf{x}) = G(\alpha_j - \mathbf{x}^T \boldsymbol{\beta}).$

## Example 6.3

*Consider a mental health dataset consisting of 40 subjects. For each subject, the response is a mental impairment variable including four levels: Well, Mild, Moderate, and Impaired. The two covariates of interest are a life events index ( $x_1$ ) and socioeconomic status ( $x_2=SES$ ).*

# Ordinal Regression Models

```
table.7.5<-read.table("your directory/Mental.txt",col.names=c("mental",  
"ses", "life"))  
table.7.5$mental<-ordered(table.7.5$mental, levels=1:4, labels=c("well",  
"mild", "moderate", "impaired"))  
fit.polr<- polr(mental ~ ses + life, data = table.7.5)  
summary(fit.polr)  
Value Std. Error t value  
ses 1.1112270 0.6108459 1.819161  
life -0.3188574 0.1209897 -2.635411  
well|mild -0.2819 0.6422 -0.4389  
mild|moderate 1.2128 0.6607 1.8356  
moderate|impaired 2.2094 0.7210 3.0645
```

# Ordinal Regression Models

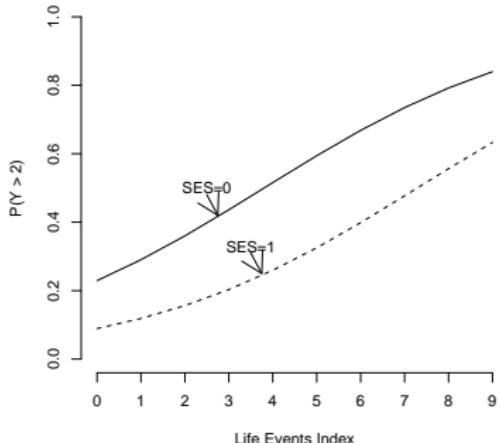


Figure: Figure 1:  $P(Y > 2)$  based on the estimated parameters.

# Poisson Regression

## Definition 6.3

*Poisson regression assumes*

- (i) *the components of  $\mathbf{y} = (y_1, \dots, y_n)^T$  are mutually independent, and  $y_i | \mathbf{x}_i \sim \text{Poisson}(\mu(\mathbf{x}_i))$  for  $i = 1, \dots, n$ ;*
- (ii)  *$\mu(\mathbf{x}_i)$  is related to  $\mathbf{x}_i$  by*

$$g(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta} \quad (6.4)$$

*for  $i = 1, \dots, n$ , where  $\mathbf{x}_i$  is a  $p \times 1$  covariate vector and  $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$  is defined in a subset  $\mathcal{B}$  of  $R^p$  ( $p < n$ ). If  $g(\mu) = \log(\mu)$ , then we obtain a **loglinear model**.*

# Poisson Regression

Poisson regression is a regression technique for **modeling count data, such as colony counts for bacteria or viruses and accidents, as a function of a set of covariates.**

- $\ell_n(\beta) = \sum_{i=1}^n [-\exp(\mathbf{x}_i^T \beta) + y_i \mathbf{x}_i^T \beta - \log(y_i!)]$ .
- $\partial_\beta \ell_n(\beta) = \sum_{i=1}^n [y_i - \mu_i(\beta)] \mathbf{x}_i = \sum_{i=1}^n \frac{y_i - E(y_i)}{\text{Var}(y_i)} \partial_\beta \mu_i(\beta)$
- $-\partial_\beta^2 \ell_n(\beta) = \sum_{i=1}^n \mu_i(\beta) \mathbf{x}_i^{\otimes 2}$ .
- $\beta^{k+1} = \beta^k + \{\sum_{i=1}^n \mu_i(\beta^k) \mathbf{x}_i^{\otimes 2}\}^{-1} \sum_{i=1}^n [y_i - \mu_i(\beta^k)] \mathbf{x}_i$ .

# Poisson Regression

- G-statistic:  $D(\mathbf{y}; \hat{\mu}) = 2 \sum_{i=1}^n \{y_i \log(y_i/\hat{\mu}_i) - (y_i - \hat{\mu}_i)\}$ .
- If  $\mathbf{x}_i$  includes an intercept, then  $\sum_{i=1}^n [y_i - \mu_i(\hat{\beta})] = 0$  and  $D(\mathbf{y}; \hat{\mu}) = 2 \sum_{i=1}^n \{y_i \log(y_i/\hat{\mu}_i)\}$ .
- If two models  $M_1 : \mu_i(\beta_{(1)}) = g_1^{-1}(\mathbf{x}_{i,(1)}^T \beta_{(1)})$  and  $M_2 : \mu_i = g_1^{-1}(\mathbf{x}_i^T \beta)$  contain an intercept,

$$\begin{aligned} G^2(M_1|M_2) &= 2 \sum_{i=1}^n \left\{ y_i \log(y_i/\mu_i(\hat{\beta}_{(1)})) - y_i \log(y_i/\mu_i(\hat{\beta})) \right\} \\ &= 2 \sum_{i=1}^n \left\{ y_i \log(\mu_i(\hat{\beta})/\mu_i(\hat{\beta}_{(1)})) \right\}, \end{aligned}$$

where  $\mathbf{x}_i = (\mathbf{x}_{i,(1)}, \mathbf{x}_{i,(2)})$  and  $\beta = (\beta_{(1)}, \beta_{(2)})$ .

# Poisson Regression

- $G^2(M_1|M_2)$  is asymptotically  $\chi^2(df = \dim(\beta) - \dim(\beta_{(1)}))$  as  $n \rightarrow \infty$ .
- For log-linear model,

$$G^2(M_1|M_2) = 2 \sum_{i=1}^n \left\{ \mu_i(\hat{\beta}_{(2)}) \log(\mu_i(\hat{\beta}_{(1)})/\mu_i(\hat{\beta}_{(2)})) \right\}.$$

# Poisson Regression

Poisson regression is useful for modeling **event rates as a function of a set of covariates.**

- Model incidence densities in epidemiologic studies where events are occurrences of rare diseases for subpopulations of different sizes
- Model incidence densities in epidemiologic studies where events are occurrences of rare diseases for individuals with possibly differing amounts of exposure to risk
- $y_i$ =the observed number of events,  $N_i$ =the total length of follow-up time for all subjects in that subgroup,  $\lambda(\mathbf{x}_i, \beta)$ =the event rate for the  $i$ -th subgroup
- $E(y_i) = \mu_i = N_i \lambda(\mathbf{x}_i, \beta), \quad i = 1, \dots, n.$
- $\lambda(\mathbf{x}_i, \beta) = \exp(\mathbf{x}_i^T \beta)$

## Example 6.4

Consider a damage incidents dataset consisting of 34 observations. For each observation, the response is the number of damage incidents, and the covariates of interest include ship type ( $x_{i1}$ ), year of construction ( $x_{i2}$ ), period of operation ( $x_{i3}$ ), and logarithm of aggregate months of service ( $x_{i4}$ ). We are interested in knowing how the risk of damage is associated with the three classification factors  $x_{i1}$ ,  $x_{i2}$  and  $x_{i3}$ .

# Poisson Regression

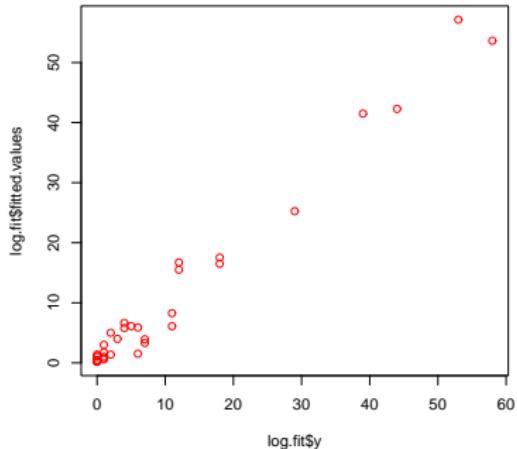


Figure: Figure 2: Plot of fitted values and responses.

# Poisson Regression

```
> log.fit<-glm(Response~log(Months)
+factor(Period)+factor(Year)*factor(Ship),
family=poisson(link=log),data=table62)
> log.fit<-update(log.fit,.~-factor(Year):factor(Ship))
> log.fit<-update(log.fit,.~-factor(Ship))
Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.22294 0.48260 -10.822 < 2e-16 ***
log(Months) 0.83107 0.04602 18.059 < 2e-16 ***
factor(Period)2 0.35465 0.11685 3.035 0.00240 **
factor(Year)2 0.67352 0.15029 4.481 7.42e-06 ***
factor(Year)3 0.79669 0.17019 4.681 2.85e-06 ***
factor(Year)4 0.39785 0.23375 1.702 0.08875.
Null deviance: 614.54 on 33 degrees of freedom
Residual deviance: 50.37 on 28 degrees of freedom AIC: 160.24
```

Loglinear models have been widely used for the analysis of **contingency tables**.

- Modeling categorical variables from cross-sectional sampling.
- Assessing statistical independence and dependence of multiple categorical variables.
- Modeling cell counts in contingency tables.
- The expected count in each cell depends on both levels of the categorical variables and associations and interactions among them.
- The framework and flavor of loglinear modeling is similar to analysis of variance modeling for continuous responses.

# Loglinear Models for Contingency Tables

**Table:** Table 5.3: Cell counts, probabilities and expected counts of a  $2 \times 2$  table

X	Y		Total
	1	2	
1	$n_{11}$	$n_{12}$	$n_{1\cdot}$
2	$n_{21}$	$n_{22}$	$n_{2\cdot}$
Total	$n_{\cdot 1}$	$n_{\cdot 2}$	$n$

X	Y		Total
	1	2	
1	$\pi_{11}$	$\pi_{12}$	$\pi_{1\cdot}$
2	$\pi_{21}$	$\pi_{22}$	$\pi_{2\cdot}$
Total	$\pi_{\cdot 1}$	$\pi_{\cdot 2}$	1

# Loglinear Models for Contingency Tables

The loglinear model for a  $2 \times 2$  table is

$$\log(\mu_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}, \quad i, j = 1, 2. \quad (6.5)$$

- Identifiability:  $\lambda_1^X + \lambda_2^X = 0$ ,  $\lambda_1^Y + \lambda_2^Y = 0$ ,  $\lambda_{11}^{XY} = \lambda_{22}^{XY}$ ,  
 $\lambda_{11}^{XY} + \lambda_{12}^{XY} = 0$ ,  $\lambda_{21}^{XY} + \lambda_{22}^{XY} = 0$ .
- Four identifiable parameters  $\lambda$ ,  $\lambda_1^X$ ,  $\lambda_1^Y$ , and  $\lambda_{11}^{XY}$ .
- $\log R_{XY} = 4\lambda_{11}^{XY}$ .
- as  $n \rightarrow \infty$ ,  $G^2 = 2 \sum_{i=1}^2 \sum_{j=1}^2 n_{ij} \log(n_{ij}/\tilde{\mu}_{ij}) \rightarrow \chi^2(1)$  where  
 $\tilde{\mu}_{ij} = n_{i\cdot}n_{\cdot j}/n$ .

# Loglinear Models for Contingency Tables

		Y	
X		1	2
1	1	$\exp(\lambda + \lambda_1^X + \lambda_1^Y + \lambda_{11}^{XY})$	$\exp(\lambda + \lambda_1^X - \lambda_1^Y - \lambda_{11}^{XY})$
	2	$\exp(\lambda - \lambda_1^X + \lambda_1^Y - \lambda_{11}^{XY})$	$\exp(\lambda - \lambda_1^X - \lambda_1^Y + \lambda_{11}^{XY})$

## Example 6.5

*Consider a bicycle dataset observed from a cross-sectional study with 100 subjects (Stokes et al., 2005). We are interested in establishing the relationship between ‘Wearing a Helmet’ and ‘Bicycle Type’. We calculate  $G^2 = 4.56$  with  $p$ -value 0.033. Thus, we reject the hypothesis that ‘Wearing a Helmet’ and ‘Bicycle Type’ are independent at the 5% significance level.*

# Loglinear Models for Contingency Tables

Table: Table 5.4: Bicycle dataset

		Wearing a Helmet		
Bicycle Type		Yes	No	Total
Mountain		34	32	66
other		10	24	34
	Total	44	56	100

# Loglinear Models for Contingency Tables

Consider an  $I \times J$  contingency table and  $X$  and  $Y$ , respectively, represent the row and column categorical variables.

- A sample of  $n$  subjects.
- The cell frequency counts  $y_{ij} = n_{ij}$  for  $i = 1, \dots, I; j = 1, \dots, J$ .
- The cell probabilities are  $\{\pi_{ij}\}$ .
- The expected frequencies are  $\{\mu_{ij} = n\pi_{ij}\}$ .
- All  $IJ$  cell counts  $y_{ij}$  are independent and  $y_{ij} \sim \text{Poisson}(\mu_{ij})$  for all cells.

# Loglinear Models for Contingency Tables

If  $X$  and  $Y$  are independent, then  $\pi_{ij} = \pi_{i\cdot}\pi_{\cdot j}$  for all  $i, j$ .

- $\mu_{ij} = n\pi_{i\cdot}\pi_{\cdot j}$  and  $\log \mu_{ij}$  can be written as

$$\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y, \quad (6.6)$$

where  $\lambda_i^X$  denotes a row effect and  $\lambda_j^Y$  denotes a column effect.

- Using a reference cell coding scheme, we have

$$\log \mu_{ij} = \lambda + \sum_{a=1}^{I-1} x_{(i,j)a} \lambda_a^X + \sum_{b=1}^{J-1} y_{(i,j)b} \lambda_b^Y. \quad (6.7)$$

- $\log \mu_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta}$ .

# Loglinear Models for Contingency Tables

- Saturated model:

$$\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}, \quad (6.8)$$

where  $\lambda_{ij}^{XY}$  denote the interactions between  $X$  and  $Y$  and reflect deviations from independence.

- Using the reference cell coding scheme, we have

$$\log \mu_{ij} = \lambda + \sum_{a=1}^{I-1} x_{(i,j)a} \lambda_a^X + \sum_{b=1}^{J-1} y_{(i,j)b} \lambda_b^Y + \sum_{a=1}^{I-1} \sum_{b=1}^{J-1} \lambda_{ab}^{XY} x_{(i,j)a} y_{(i,j)b},$$

- $\log \mu_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta}.$

# Loglinear Models for Contingency Tables

The  $\lambda$  parameters in loglinear models have some meaningful interpretations.

- For model  $(X, Y)$ , we have

$$\begin{aligned}\log \frac{P(Y = j|X = i)}{P(Y = k|X = i)} &= \log \frac{\pi_{ij}}{\pi_{ik}} = \log \frac{\mu_{ij}}{\mu_{ik}} = \log \mu_{ij} - \log \mu_{ik} \\ &= (\lambda + \lambda_i^X + \lambda_j^Y) - (\lambda + \lambda_i^X + \lambda_k^Y) = \lambda_j^Y - \lambda_k^Y,\end{aligned}$$

which is independent of  $i$ .

- The ratio of  $P(Y = j|X = i)$  to  $P(Y = k|X = i)$  is identical at each level of  $X$ .

# Loglinear Models for Contingency Tables

- For model  $(XY)$ , the odds ratios of  $XY$  are given by  
$$\log R_{ij} = \log \frac{\pi_{ii}\pi_{jj}}{\pi_{ij}\pi_{ji}} = \log \frac{\mu_{ii}\mu_{jj}}{\mu_{ij}\mu_{ji}} = \lambda_{ii}^{XY} + \lambda_{jj}^{XY} - \lambda_{ij}^{XY} - \lambda_{ji}^{XY},$$
 where  $i = 1, \dots, I$  and  $j = 1, \dots, J.$
- $\{\lambda_{ij}^{XY}\}$  determine the association between  $X$  and  $Y.$

# Loglinear Models for Contingency Tables

- For an  $I \times J$  table, the likelihood function is given by

$$L(\mu) = \prod_{i=1}^I \prod_{j=1}^J \frac{\exp(-\mu_{ij}) \mu_{ij}^{n_{ij}}}{n_{ij}!}.$$



$$\ell_n(\mu) = \sum_{i=1}^I \sum_{j=1}^J n_{ij} \log \mu_{ij} - \sum_{i=1}^I \sum_{j=1}^J \mu_{ij}.$$

# Loglinear Models for Contingency Tables

- For the loglinear model (XY), this simplifies to

$$\begin{aligned}\ell_n(\mu) &= n\lambda + \sum_{i=1}^I n_{i+} \lambda_i^X + \sum_{j=1}^J n_{+j} \lambda_j^Y \\ &+ \sum_{i=1}^I \sum_{j=1}^J n_{ij} \lambda_{ij}^{XY} - \sum_{i=1}^I \sum_{j=1}^J \exp(\lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}),\end{aligned}\tag{6.9}$$

where  $n_{i+} = \sum_{j=1}^J n_{ij}$  and  $n_{+j} = \sum_{i=1}^I n_{ij}$ .

- The Poisson distribution is in the exponential family, and therefore, the coefficients of the parameters are the sufficient statistics.

# Loglinear Models for Contingency Tables

For loglinear models, we have

- $\ell_n(\beta) = \sum_{i=1}^I \sum_{j=1}^J n_{ij} \mathbf{x}_{ij}^T \beta - \sum_{i=1}^I \sum_{j=1}^J \exp(\mathbf{x}_{ij}^T \beta).$
- The sufficient statistic for the  $l$ th component of  $\beta$  is  $\sum_{i=1}^I \sum_{j=1}^J n_{ij} x_{ij;l}$ , where  $x_{ij;l}$  is the  $l$ -th component of  $\mathbf{x}_{ij}$ .
- $\sum_{i=1}^I \sum_{j=1}^J n_{ij} \mathbf{x}_{ij} = \sum_{i=1}^I \sum_{j=1}^J E[n_{ij}] \mathbf{x}_{ij} = \sum_{i=1}^I \sum_{j=1}^J \mu_{ij} \mathbf{x}_{ij}.$
- $-\partial_\beta^2 \ell_n(\beta) = \sum_{i=1}^I \sum_{j=1}^J \mu_{ij} \mathbf{x}_{ij}^{\otimes 2}.$
- $G^2 = 2 \sum_{i=1}^I \sum_{j=1}^J n_{ij} \log(n_{ij}/\tilde{\mu}_{ij})$ , where  $\tilde{\mu}_{ij} = n_i \cdot n_j / n$ . As  $n \rightarrow \infty$ ,  $G^2$  converges to a  $\chi^2((I-1)(J-1))$  distribution.

## Example 6.6

*Consider a malignant melanoma dataset observed from a cross-sectional study with 400 subjects having malignant melanoma (Roberts, et al., 1981). The malignant melanoma counts for each of the tumor sites and the histological types were recorded (see Table 5.5). We are interested in establishing the relationship between malignant melanoma counts, tumor site, and histological type. The likelihood ratio statistic  $G^2 = 51.80$  with 6 df ( $p < .0001$ ) provides strong evidence that tumor type and tumor site are not independent.*

# Loglinear Models for Contingency Tables

Table: Table 5.5: Malignant melanoma dataset

Tumor Type	Tumor Site			Total
	Head + Neck	Trunk	Extremities	
Hutchinson's melanotic freckle	22	2	10	34
Superficial spreading melanoma	16	54	115	185
Nodular	19	33	73	125
Indeterminate	11	17	28	56
Total	68	106	226	400

## $I \times I$ Contingency Tables

Square  $I \times I$  contingency tables commonly appear in comparing categorical responses for two samples when each observation in one sample is paired with an observation in the other. Such matched pairs are common in longitudinal studies.

**Table:** Table 5.6: Survey dataset

		Second Survey		Total
First		Approve	Disapprove	
Approve		794	150	944
Disapprove		86	570	656
Total		880	720	1600

## $I \times I$ Contingency Tables

Table: Table 5.7: Sample Square Table

		$Y_2$			
$Y_1$		1	$\dots$		Total
1	$n_{11}$	$\dots$	$n_{1I}$	$n_{1+}$	
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
	$n_{I1}$	$\dots$	$n_{II}$	$n_{I+}$	
$n_{+1}$	$\dots$	$n_{+I}$	$n$		

# Loglinear Models for Contingency Tables

- A square  $I \times I$  table  $\{n_{ab}\}$  consists of counts of possible sequences  $(a, b)$  of outcomes for  $(Y_1, Y_2)$ . If the cell probabilities are  $\{\pi_{ij}\}$ , the expected frequencies are  $\{\mu_{ij} = n\pi_{ij}\}$ .
- An  $I \times I$  joint distribution  $\{\pi_{ij}\}$  is *symmetric* if  $\pi_{ij} = \pi_{ji}$  for all  $i \neq j$ .
- We have  $\pi_{i+} = \pi_{+i}$  for all  $i$ , which is called *marginal homogeneity*.
- For  $I = 2$ , marginal homogeneity is equivalent to symmetry, whereas they are not equivalent for  $I > 2$ .

- The log-linear model for symmetry can be written as

$$\log \mu_{ij} = \lambda + \lambda_i + \lambda_j + \lambda_{ij}, \quad (6.10)$$

in which  $\lambda_{ij} = \lambda_{ji}$ .

- For identification purposes, we need to impose constraints, such as  $\lambda + \lambda_i + \lambda_j = 0$  for all  $i, j = 1, \dots, I$ .
- $\hat{\mu}_{ij} + \hat{\mu}_{ji} = n_{ij} + n_{ji}$ ,  $\hat{\mu}_{ii} = n_{ii}$ , for all  $i < j$  and  $i = 1, \dots, I$ .  
Thus,  $\hat{\mu}_{ij} = (n_{ij} + n_{ji})/2$ .

# Loglinear Models for Contingency Tables

- $G^2 = 2 \sum_{i \neq j} n_{ij} \log(2n_{ij}/(n_{ij} + n_{ji}))$ . As  $n \rightarrow \infty$ ,  $G^2$  converges to  $\chi^2(I(I - 1)/2)$ .
- $\chi^2 = \sum_{i < j} \frac{(observed - fitted)^2}{fitted} = \sum_{i < j} \frac{(n_{ij} - n_{ji})^2}{n_{ij} + n_{ji}}$ , which is asymptotically  $\chi^2$  distributed with  $df = I(I - 1)/2$ . When  $I = 2$ ,  $\chi^2$  is the well-known McNemar's test.

## $I \times I$ Contingency Tables

- Another popular model is *quasi-symmetry* defined by

$$\log \mu_{ij} = \lambda + \lambda_i^{Y_1} + \lambda_j^{Y_2} + \lambda_{ij},$$

in which  $\lambda_{ij} = \lambda_{ji}$  for all  $i < j$ .

- Quasi-symmetry reduces to symmetry when  $\lambda_i^{Y_1} = \lambda_i^{Y_2}$  for all  $i$ .
- We can rewrite quasi-symmetry in terms of  $\{\pi_{ij}\}$  as  $\pi_{ij} = \alpha_i \beta_j \gamma_{ij}$ , in which  $\gamma_{ij} = \gamma_{ji}$  for all  $i < j$  and all parameters are positive. If  $\alpha_i = \beta_i$  for all  $i$ , the quasi-symmetry model reduces to the symmetry model.

## Example 6.7

*Consider a poll of a random sample of 1600 voting-age British citizens in two surveys at different times. In the first survey, 944 supported the prime minister's performance, whereas 800 supported in the second survey. McNemar's test is given by*

$$\chi^2 = \frac{(86 - 150)^2}{86 + 150} = (4.17)^2,$$

*which indicates strong evidence of a drop in the approval rating.*

## $I \times J \times K$ Contingency Tables

- Three variables  $X$ ,  $Y$ , and  $Z$ , respectively, have  $I$ ,  $J$ , and  $K$  categories.
- The  $I \times J \times K$  table usually describes the association between two variables  $X$  and  $Y$ , while controlling for the third variable  $Z$ .
- $X$ ,  $Y$ , and  $Z$ , respectively, denote blood groups ( $O$  and  $A$ ), subject groups (peptic-ulcer and control), and cities (London, Manchester, Newcastle).

# Loglinear Models for Contingency Tables

We define **conditional and marginal odds ratios** to describe conditional and marginal associations.

- For  $2 \times 2 \times K$  tables, we can define the **conditional association** between  $X$  and  $Y$  for a fixed level of  $Z = k$  as follows:

$$R_{XY(k)} = \frac{\pi_{2|2k}\pi_{1|1k}}{\pi_{2|1k}\pi_{1|2k}} = \frac{\pi_{22k}\pi_{11k}}{\pi_{21k}\pi_{12k}} \approx \frac{n_{22k}n_{11k}}{n_{21k}n_{12k}}, \quad (6.11)$$

where  $\pi_{j|ik} = P(Y = j|X = i, Z = k)$ .

- Frequencies in the marginal  $XY$  table are  $n_{ij\cdot} = \sum_{k=1}^K n_{ijk\cdot}$ . These odds ratios  $R$  can be defined as

$$R_{XY} = \frac{\pi_{2|2}\pi_{1|1}}{\pi_{2|1}\pi_{1|2}} = \frac{n_{22\cdot}n_{11\cdot}}{n_{21\cdot}n_{12\cdot}}, \quad (6.12)$$

where  $\pi_{j|i} = P(Y = j|X = i)$ .

# Loglinear Models for Contingency Tables

Table: Table 5.8: Data for Example 6.8

Clinic	Treatment	Response	
		Success	Failure
1	A	18	12
	B	12	8
2	A	2	8
	B	8	32
Total	A	20	20
	B	20	40

## Example 6.8

Consider the  $2 \times 2 \times 2$  table in Table 5.8. The three variables  $X$ ,  $Y$ , and  $Z$  are, respectively, Treatment (A and B), Response (Success and Failure), and Clinic (1 and 2). The conditional  $XY$  odds ratios are

$$R_{XY(1)} = 1.0 \text{ and } R_{XY(2)} = 1.0.$$

However, the marginal odds ratio  $R_{XY}$  equals 2.0. Thus,  $X$  and  $Y$  are not marginally independent, but given  $Z$ , they are conditionally independent.

# Loglinear Models for Contingency Tables

For  $I \times J \times K$  tables, **conditional odds ratios** can be introduced as:

- At a fixed  $Z = k$ , we define the conditional  $XY$  association using  $(I - 1)(J - 1)$  consecutive odds ratios as follows:

$$R_{ij(k)} = \frac{\pi_{ijk}\pi_{i+1,j+1,k}}{\pi_{i,j+1,k}\pi_{i+1,j,k}}, \quad 1 \leq i \leq I - 1; \quad 1 \leq j \leq J - 1. \quad (6.13)$$

- If  $R_{ij(1)} = \dots = R_{ij(K)}$ , then this  $I \times J \times K$  table has *homogeneous XY association*.
- If  $X$  and  $Y$  are independent for the  $k$ th level of  $Z$ , then  $X$  and  $Y$  are *conditionally independent* at the  $k$ th level of  $Z$ .
- If  $X$  and  $Y$  are conditionally independent at every level of  $Z$ , then given  $Z$ ,  $X$  and  $Y$  are conditionally independent.

## Loglinear Models for $I \times J \times K$ Tables

- Consider an  $I \times J \times K$  contingency table and three categorical variables  $X$ ,  $Y$ , and  $Z$ .
- We have a sample of  $n$  subjects and the cell frequency counts are  $\{n_{ijk}\}$  for  $i = 1, \dots, I; j = 1, \dots, J; k = 1, \dots, K$
- The cell probabilities are  $\{\pi_{ijk}\}$  and the expected frequencies are  $\{\mu_{ijk} = n\pi_{ijk}\}$ .
- The loglinear model assumes that the  $IJK$  cell counts  $n_{ijk}$  are independent and  $n_{ijk} \sim Poisson(\mu_{ijk})$  for all cells.

# Loglinear Models for Contingency Tables

We consider several commonly used loglinear models for analyzing  $I \times J \times K$  contingency tables.

- Independence:  $\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z$ .
- The variable  $Z$  is jointly independent of  $X$  and  $Y$ ,  
 $\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY}$ .
- The **no three factor interaction** model is given by

$$\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_k^Z + \lambda_j^Y + \lambda_{ij}^{XY} + \lambda_{jk}^{YZ} + \lambda_{ik}^{XZ}. \quad (6.14)$$

- The saturated model is given by

$$\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_k^Z + \lambda_j^Y + \lambda_{ij}^{XY} + \lambda_{jk}^{YZ} + \lambda_{ik}^{XZ} + \lambda_{ijk}^{XYZ}. \quad (6.15)$$

# Loglinear Models for Contingency Tables

Define  $\{x_{(ijk)a} : a = 1, \dots, I - 1\}$ ,  $\{y_{(ijk)b} : b = 1, \dots, J - 1\}$ , and  $\{z_{(ijk)c} : c = 1, \dots, K - 1\}$  as three sets of dummy variables. Using the **reference cell coding scheme**,  $\log \mu_{ijk}$  can reexpressed as

$$\begin{aligned}\log \mu_{ijk} = & \lambda + \sum_{a=1}^{I-1} x_{(ijk)a} \lambda_a^X + \sum_{b=1}^{J-1} y_{(ijk)b} \lambda_b^Y + \sum_{c=1}^{K-1} z_{(ijk)c} \lambda_c^Z \\ & + \sum_{a=1}^{I-1} \sum_{b=1}^{J-1} \lambda_{ab}^{XY} x_{(ijk)a} y_{(ijk)b} + \sum_{a=1}^{I-1} \sum_{c=1}^{K-1} \lambda_{ac}^{XZ} x_{(ijk)a} z_{(ijk)c} + \\ & \sum_{b=1}^{J-1} \sum_{c=1}^{K-1} \lambda_{bc}^{YZ} y_{(ijk)b} z_{(ijk)c} + \sum_{b=1}^{J-1} \sum_{c=1}^{K-1} \sum_{a=1}^{I-1} \lambda_{abc}^{XYZ} x_{(ijk)a} y_{(ijk)b} z_{(ijk)c}.\end{aligned}$$

# Loglinear Models for Contingency Tables

The  $\lambda$  parameters in loglinear models have a strong connection with odds ratios and conditional odds ratios. For instance, for model (6.14),

$$\log R_{ij(k)} = \log \frac{\mu_{ijk}\mu_{i+1,j+1,k}}{\mu_{i+1,jk}\mu_{i,j+1,k}} = \lambda_{ij}^{XY} + \lambda_{i+1,j+1}^{XY} - \lambda_{i,j+1}^{XY} - \lambda_{i+1,j}^{XY}.$$

Because the right-hand side of the above equation is independent of  $k$ , the  $I \times J \times K$  table has homogeneous  $XY$  association based on the no three factor interaction model. Similarly, the  $I \times J \times K$  table also has homogeneous  $YZ$  and  $XZ$  associations based on the no three factor interaction model.

# Loglinear Models for Contingency Tables



$$L(\mu) = \prod_{i=1}^I \prod_{j=1}^J \prod_{k=1}^K \frac{\exp(-\mu_{ijk}) \mu_{ijk}^{n_{ijk}}}{n_{ijk}!}$$



$$\ell_n(\mu) = \sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^K n_{ijk} \log \mu_{ijk} - \sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^K \mu_{ijk}.$$

# Loglinear Models for Contingency Tables



$$\begin{aligned}\ell_n(\mu) &= n\lambda + \sum_{i=1}^I n_{i++} \lambda_i^X + \sum_{j=1}^J n_{+j+} \lambda_j^X + \sum_{k=1}^K n_{++k} \lambda_k^Z \quad (6.16) \\ &+ \sum_{i=1}^I \sum_{j=1}^J n_{ij+} \lambda_{ij}^{XY} + \sum_{i=1}^I \sum_{k=1}^K n_{i+k} \lambda_{ik}^{XZ} + \sum_{j=1}^J \sum_{k=1}^K n_{+jk} \lambda_{jk}^{YZ} \\ &+ \sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^K n_{ijk} \lambda_{ijk}^{XYZ} - \sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^K \exp(\lambda + \lambda_i^X + \dots + \lambda_{ijk}^{XYZ}).\end{aligned}$$

- The Poisson distribution is in the exponential family, and therefore, the coefficients of the parameters are the sufficient statistics.

# Loglinear Models for Contingency Tables

$$\log \mu_{ijk} = \mathbf{x}_{ijk}^T \boldsymbol{\beta}$$

- $\ell_n(\boldsymbol{\beta}) = \sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^K n_{ijk} \mathbf{x}_{ijk}^T \boldsymbol{\beta} - \sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^K \exp(\mathbf{x}_{ij}^T \boldsymbol{\beta}).$
- The sufficient statistic for the  $l$ th component of  $\boldsymbol{\beta}$  is  $\sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^K n_{ij} x_{ij;l}$ , where  $x_{ij;l}$  is the  $l$ -th component of  $\mathbf{x}_{ij}$ .
- The score equation for  $\boldsymbol{\beta}$  is  $\sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^K n_{ijk} \mathbf{x}_{ijk} = \sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^K \mu_{ijk} \mathbf{x}_{ijk}.$
- $-\partial_{\boldsymbol{\beta}}^2 \ell_n(\boldsymbol{\beta}) = \sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^K \mu_{ijk} \mathbf{x}_{ijk}^{\otimes 2}.$

We can use the deviance to compare the fitted cell counts obtained from a particular model to the sample counts. We can also compare two sets of fitted cell counts obtained from two particular models.

- For model  $(X, Y, Z)$ , we can calculate  $\hat{\beta}$  under the model  $(X, Y, Z)$  and then compute

$$D(\mathbf{y}; \hat{\mu}) = 2 \sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^K [n_{ijk} \log(n_{ijk}/\mu_{ijk}(\hat{\beta}))]. \quad (6.17)$$

- The residual degrees of freedom between the models  $(X, Y, Z)$  and  $(XYZ)$  equals

$$df = IJK - [1 + (I-1) + (J-1) + (K-1)] = IJK - I - J - K + 2,$$

in which model  $(X, Y, Z)$  has a single intercept and constraints such as  $\lambda_I^X = \lambda_J^Y = \lambda_K^Z = 0$ .

# Loglinear Models for Contingency Tables

- Consider two nested loglinear models  $M_1 : \log \mu_{ijk}^{(1)} = \mathbf{x}_{ijk}^{(1)T} \beta^{(1)}$  and  $M_2 : \log \mu_{ijk}^{(2)} = \mathbf{x}_{ijk}^{(2)T} \beta^{(2)}$ . Without loss of generality, we assume that  $M_1$  is nested in  $M_2$ .
- 

$$\begin{aligned} G^2(M_1 | M_2) \\ &= 2 \sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^K [n_{ijk} \log(n_{ijk}/\mu_{ijk}^{(1)}(\hat{\beta}^{(1)})) - n_{ijk} \log(n_{ijk}/\mu_{ijk}^{(2)}(\hat{\beta}^{(2)}))] \\ &= 2 \sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^K [\mu_{ijk}^{(2)}(\hat{\beta}^{(2)}) \log(\mu_{ijk}^{(2)}(\hat{\beta}^{(2)})/\mu_{ijk}^{(1)}(\hat{\beta}^{(1)}))], \end{aligned} \quad (6.18)$$

which is asymptotically  $\chi^2$  with  $df = df(M_1) - df(M_2)$  as  $n \rightarrow \infty$ .

# Loglinear Models for Contingency Tables

Table: Table 5.9: Residual degrees of freedom for  $I \times J \times K$  tables

Model	df
$(X, Y, Z)$	$IJK - I - J - K + 2$
$(XY, Z)$	$(K - 1)(IJ - 1)$
$(XZ, Y)$	$(J - 1)(IK - 1)$
$(X, YZ)$	$(I - 1)(JK - 1)$
$(XY, YZ)$	$J(I - 1)(K - 1)$
$(XZ, YZ)$	$K(I - 1)(J - 1)$
$(XY, XZ)$	$I(K - 1)(J - 1)$
$(XY, XZ, YZ)$	$(I - 1)(J - 1)(K - 1)$
$(XYZ)$	0

## Example 6.9

We analyzed the substance abuse dataset in Table 4.1 using different log-linear models. We also estimated the conditional and marginal odds ratios for each fitted model. Table 5.10 shows the fitted values for several loglinear models. The model (AC, AM, CM) fits the data reasonably well. We presented these odds ratios in Table 5.10. For instance, the AC conditional association equals 1.0 for model (AM, CM). The odds ratios for the observed data are reported for model (ACM).

# Loglinear Models for Contingency Tables

Table: Table 5.10: Estimated Odds Ratios for Example 6.8

Model	Conditional Association			Marginal Association		
	AC	AM	CM	AC	AM	CM
(A, C, M)	1.0	1.0	1.0	1.0	1.0	1.0
(AC, M)	17.7	1.0	1.0	17.7	1.0	1.0
(AM, CM)	1.0	61.9	25.1	2.7	61.9	25.1
(AC, AM, CM)	7.8	19.8	17.3	17.7	61.9	25.1
(ACM) level 1	13.8	24.3	17.5	17.7	61.9	25.1
(ACM) level 2	7.7	13.5	9.7			

# Loglinear Models for Contingency Tables

```
table.8.3<-data.frame(expand.grid( marijuana=factor(c("Yes","No"),  
levels=c("No","Yes")), cigarette=factor(c("Yes","No"), levels=c("No","Yes")),  
alcohol=factor(c("Yes","No"), levels=c("No","Yes"))),  
count=c(911,538,44,456,3,43,2,279))  
library(MASS)  
fitACM<-loglm(count~alcohol*cigarette*marijuana, data=table.8.3,  
param=T, fit=T)  
fitAC.AM.CM<-update(fitACM, .~. - alcohol:cigarette:marijuana)  
fitAM.CM<-update(fitAC.AM.CM, .~. - alcohol:cigarette)  
fitAC.M<-update(fitAC.AM.CM, .~. - alcohol:marijuana -cigarette:marijuana)  
fitA.C.M<-update(fitAC.M, .~. - alcohol:cigarette)
```

# Loglinear Models for Contingency Tables

marijuana cigarette alcohol ACM AC.AM.CM AM.CM AC.M  
A.C.M

1 Yes Yes Yes 279 279.61 179.84 162.47 64.87

2 No Yes Yes 2 1.38 0.24 118.52 47.32

3 Yes No Yes 43 42.383882 142.159 26.59 124.19

4 No No Yes 3 3.61 4.76 19.40 90.59

5 Yes Yes No 456 455.38 555.15 289.10 386.7

6 No Yes No 44 44.61 45.76 210.89 282.09

7 Yes No No 538 538.61 438.84 837.82 740.22

8 No No No 911 910.38 909.23 611.17 539.98

## Chapter 7: Eliminating Nuisance Parameters

- Nuisance parameters
- Hypergeometric distributions
- Conditional logistic regression
- Binary matched pairs
- Conditional likelihood

Statistical models involves a set of parameters  $\xi = (\psi, \lambda)$ , in which  $\psi$  is the **parameter of interest** and  $\lambda$  is a **nuisance parameter**.

## Example 7.1

Consider  $y_i = \mathbf{x}_i^T \beta + \epsilon_i$ , where  $\epsilon_i \sim N(0, \sigma^2)$ . In most applications,  $\beta$  is the **parameter of interest** and  $\sigma^2$  is a **nuisance parameter**. We may use the likelihood function  $\ell_n(\xi)$  to carry out statistical inference on  $\psi = \beta$  in the presence of  $\lambda = \sigma$ . However, since nuisance parameters can influence the inference on  $\psi$ , **specific methods are needed to eliminate nuisance parameters**.

## Example 7.2

*If the number of nuisance parameters increases with the sample size, then the estimates of the parameters under this approach may not be consistent (Neyman and Scott, 1948).*

# Hypergeometric Distributions

Table: Table 6.1:  $2 \times 2$  contingency table

		Y		Total
		0	1	
X	0	$n_{11}$	$n_{12}$	$n_{1\cdot}$
	1	$n_{21}$	$n_{22}$	$n_{2\cdot}$
Total		$n_{\cdot 1}$	$n_{\cdot 2}$	$n$

- We mainly discuss **exact methods** for testing hypotheses about the odds ratios, and then constructing confidence intervals for the odds ratio from a  $2 \times 2$  table under cross-sectional sampling.
- **Why do we need exact methods?**
- Asymptotic results are valid for large sample sizes. Specifically, the margins of the  $2 \times 2$  table need to be large and the **expected cell frequencies in the table need to be larger than five** for asymptotic results to be valid.

# Hypergeometric Distributions

- For a fixed sample size  $n$ , the joint distribution of the cell counts in the  $2 \times 2$  table is given by

$$\frac{n!}{n_{11}! n_{12}! n_{21}! n_{22}!} \pi_{11}^{n_{11}} \pi_{12}^{n_{12}} \pi_{21}^{n_{21}} \pi_{22}^{n_{22}}, \quad (7.1)$$

where  $\pi_{ij} = P(X = i - 1, Y = j - 1)$ .

- Let  $\psi = \pi_{11}\pi_{22}/(\pi_{21}\pi_{12})$  be the **parameter of interest** and  $\pi_{21}$  and  $\pi_{12}$  are the **nuisance parameters**.

# Hypergeometric Distributions

- Since  $n_{12} = n_{1\cdot} - n_{11}$ ,  $n_{21} = n_{\cdot 1} - n_{11}$ , and  $n_{22} = n - n_{11} - n_{12} - n_{21}$ , we have

$$p(n_{11}, n_{\cdot 1}, n_{1\cdot} | n) \\ = \frac{n!}{n_{11}! n_{12}! n_{21}! n_{22}!} \left( \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}} \right)^{n_{11}} \pi_{12}^{n_{1\cdot}} \pi_{21}^{n_{\cdot 1}} \pi_{22}^{n - n_{1\cdot} - n_{\cdot 1}}. \quad (7.2)$$

# Hypergeometric Distributions

- To eliminate the nuisance parameters  $(\pi_{12}, \pi_{21})$ , we can calculate the conditional distribution of  $n_{11}$  given  $(n_{\cdot 1}, n_{1\cdot}, n)$ .
- Since  $n_{\cdot 1} = n_{11} + n_{21} \geq n_{11}$  and  $n_{1\cdot} \geq n_{11}$ , we have  $n_{11} \leq \min(n_{\cdot 1}, n_{1\cdot})$ .
- Since  $n_{11} \geq 0$  and  $n_{22} = n - n_{11} - (n_{1\cdot} - n_{11}) - (n_{\cdot 1} - n_{11}) = n - n_{1\cdot} - n_{\cdot 1} + n_{11} \geq 0$ , we have  $n_{11} \geq \max(0, n_{1\cdot} + n_{\cdot 1} - n)$ .

# Hypergeometric Distributions

The marginal distribution of  $(n_{1\cdot}, n_{\cdot 1}|n)$  is given by

$$p(n_{1\cdot}, n_{\cdot 1}|n) = \sum_{n_{11}=\max(0, n_{1\cdot}+n_{\cdot 1}-n)}^{\min(n_{1\cdot}, n_{\cdot 1})} \frac{n!}{n_{11}! n_{12}! n_{21}! n_{22}!} \left( \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}} \right)^{n_{11}} \pi_{12}^{n_{1\cdot}} \pi_{21}^{n_{\cdot 1}} \pi_{22}^{n-n_{1\cdot}-n_{\cdot 1}}.$$

# Hypergeometric Distributions

This is the so-called **non-central hypergeometric distribution**.

$$P(n_{11}|n_{1\cdot}, n_{\cdot 1}, n, \psi) = \frac{P(n_{11}, n_{1\cdot}, n_{\cdot 1}|n)}{P(n_{1\cdot}, n_{\cdot 1}|n)} = \frac{\binom{n_{1\cdot}}{n_{11}} \binom{n - n_{1\cdot}}{n_{\cdot 1} - n_{11}} \psi^{n_{11}}}{P_0(\psi)}, \quad (7.3)$$

where  $P_0(\psi) = \sum_{x=\max(0, n_{1\cdot} + n_{\cdot 1} - n)}^{\min(n_{1\cdot}, n_{\cdot 1})} \binom{n_{1\cdot}}{x} \binom{n - n_{1\cdot}}{n_{\cdot 1} - x} \psi^x.$

# Hypergeometric Distributions

## Hypergeometric (HG) distribution

- Exponential family:

$P(n_{11}|n_{1\cdot}, n_{\cdot 1}, n, \psi) = \exp(n_{11} \log \psi - \log P_0(\psi) + const)$  with  $\theta = \log \psi$ ,  $\phi = 1$ , and  $b(\theta) = \log P_0(\psi) = \log P_0(\exp(\theta))$ .

- The canonical parameter is  $\log \psi$ ,  $M(t) = P_0(\exp(t)\psi)/P_0(\psi)$  and  $K(t) = \log P_0(\exp(t)\psi) - \log P_0(\psi)$ .
- $\mu = P_1(\psi)/P_0(\psi)$ ,  $\sigma^2 = P_2(\psi)/P_0(\psi) - \mu^2$ , where  $P_j(\psi)$  is defined as

$$\sum_{x=\max(0, n_{1\cdot} + n_{\cdot 1} - n)}^{\min(n_{1\cdot}, n_{\cdot 1})} \binom{n_{1\cdot}}{x} \binom{n - n_{1\cdot}}{n_{\cdot 1} - x} \psi^x x^j.$$

# Hypergeometric Distributions

- The **conditional maximum likelihood estimate** (CMLE) of  $\psi$  is denoted by  $\hat{\psi}_c$ .
- $\hat{\psi}_c$  is the solution to

$$n_{11} = P_1(\hat{\psi}_c)/P_0(\hat{\psi}_c) \text{ or } P_1(\hat{\psi}_c) = n_{11}P_0(\hat{\psi}_c). \quad (7.4)$$

- $\hat{s.e.}(\hat{\psi}_c) = \hat{\psi}_c / \sqrt{\text{var}(n_{11}|n_{1\cdot}, n_{\cdot 1}, n, \hat{\psi}_c)}$ .
- The variance of  $\hat{\psi}_c$  can be approximated by the inverse of the Fisher information matrix  $I_n(\hat{\psi}_c)$ , which is given

$$I_n(\hat{\psi}_c) = E\{[\partial_\psi \log P(n_{11}|n_{1\cdot}, n_{\cdot 1}, n, \hat{\psi}_c)]^2\} = \frac{\text{Var}(n_{11}|n_{1\cdot}, n_{\cdot 1}, n, \hat{\psi}_c)}{\hat{\psi}_c^2}.$$

# Hypergeometric Distributions

## Example 7.3

For the approximate method,  $\log(\hat{\psi}) = \log(6) = 1.792$ ,  $s.e(\log(\hat{\psi})) \approx 1.683$ . In contrast, the HG log-likelihood function is  $\ell_c(\psi) = 2 \log \psi - \log P_0(\psi)$ , where  $P_0(\psi) = 4 + 18\psi + 12\psi^2 + \psi^3$ .  $\hat{\psi}_c$  satisfies  $2 = \hat{\psi}_c P'_0(\hat{\psi}_c)/P_0(\hat{\psi}_c)$ . Thus, we can obtain  $\log(\hat{\psi}_c) = 1.493$  and its standard error is 1.492.

		Y		
		0	1	Total
X	0	2	1	3
	1	1	3	4
Total		3	4	7

# Conditional Logistic Regression

- In the logistic regression model, a **large value of the ratio of the sample size  $n$  over the number of parameters  $p$**  is required to ensure reliable inference about the unknown parameters.
- When  $n/p$  is **small**, an alternative method is needed.

# Conditional Logistic Regression

- Consider

$$P(Y_i = y_i | \mathbf{x}_i) = \frac{\exp[y_i(\alpha + \sum_{j=1}^p \beta_j x_{ij})]}{1 + \exp(\alpha + \sum_{j=1}^p \beta_j x_{ij})}, \quad (7.5)$$

- For  $n$  independent observations, we have

$$P(Y_1 = y_1, \dots, Y_n = y_n | \xi) = \frac{\exp[(\sum_i y_i)\alpha + \sum_{j=1}^p (\sum_i y_i x_{ij})\beta_j]}{\prod_i [1 + \exp(\alpha + \sum_{j=1}^p \beta_j x_{ij})]}.$$

# Conditional Logistic Regression

- The sufficient statistic for  $\beta_j$  is  $s_j = \sum_i y_i x_{ij}$ ,  $j = 1, \dots, p$  and the sufficient statistic for  $\alpha$  is  $s_0 = \sum_i y_i$ .
- We eliminate  $\alpha$  by conditioning on  $\sum_{i=1}^n y_i$ , and obtain

$$P(y_1, \dots, y_n | s_0) = \frac{\exp[\sum_{j=1}^p (\sum_i y_i x_{ij}) \beta_j]}{\sum_{S(s_0)} \exp[\sum_{j=1}^p (\sum_i y_i x_{ij}) \beta_j]}, \quad (7.6)$$

where  $S(s_0)$  denotes the conditional reference set of samples having the same value of  $s_0$ .

- Thus, we get

$$P(y_1, \dots, y_n, s_0 | \xi) = P(y_1, \dots, y_n | s_0, \beta_1, \dots, \beta_p) P(s_0 | \xi).$$

# Conditional Logistic Regression

We focus on  $\beta_p$  in the logistic regression model.

- To eliminate the other parameters  $(\alpha, \beta_1, \dots, \beta_{p-1})$ , we condition on their sufficient statistics  $s_0 = \sum_i y_i$  and  $s_j = \sum_i y_i x_{ij}$  for  $j = 1, \dots, p - 1$ .

# Conditional Logistic Regression

- We obtain

$$P(y_1, \dots, y_n | s_j, j = 0, \dots, p-1) = \frac{\exp(s_p \beta_p)}{\sum_{S(s_0, \dots, s_{p-1})} \exp(s_p^* \beta_p)},$$

where  $S(s_0, \dots, s_{p-1}) = \{(y_1^*, \dots, y_n^*) : \sum_i y_i^* = s_0, \sum_i y_i^* x_{ij} = s_j, j = 1, \dots, p-1\}$ .

- Thus, we have

$$P(y_1, \dots, y_n, s_0 | \xi) = P(y_1, \dots, y_n | s_0, \dots, s_{p-1}, \beta_p) P(s_0, \dots, s_{p-1} | \xi).$$

# Conditional Logistic Regression

- Conditional inference about  $\beta_p$  can be easily carried out using  $P(y_1, \dots, y_n | s_0, \dots, s_{p-1}, \beta_p)$  or  $P(s_p = t | s_0, \dots, s_{p-1}, \beta_p)$ .
- Let  $c(\mathbf{s}, t)$  be the number of data vectors in  $S(s_0, \dots, s_{p-1})$  for which  $s_p = t$ . Therefore,

$$P(s_p = t | s_0, \dots, s_{p-1}, \beta_p) = \frac{c(\mathbf{s}, t) \exp(t\beta_p)}{\sum_u c(\mathbf{s}, u) \exp(u\beta_p)}. \quad (7.7)$$

# Conditional Logistic Regression

- The CMLE of  $\beta_p$ , denoted by  $\hat{\beta}_{p,c}$ , satisfies the score function

$$s_{p,obs} = \frac{\sum_u c(s, u) u \exp(u\hat{\beta}_{p,c})}{\sum_u c(s, u) \exp(u\hat{\beta}_{p,c})}.$$

- The covariance matrix of  $\hat{\beta}_c$  can be approximated by

$$\partial_{\beta_p}^2 \left( \log \left[ \sum_u c(s, u) \exp(u\hat{\beta}_{p,c}) \right] \right).$$

- To test  $H_0 : \beta_p = 0$ , the exact conditional *p-value* is

$$p = \sum_{t \geq s_{p,obs}} P(s_p = t | s_0, \dots, s_{p-1}, \beta_p = 0).$$

# Conditional Logistic Regression

## Example 7.4

*Consider a small dose-response study consisting of 18 subjects. We are interested in whether mortality rates are associated with dosage of a drug. The dose dataset contains life/death outcomes for six levels of drug dosage (0 to 5) and the number of deaths for three subjects who received a specific dose of the drug.*

```
data dose;
input Dose Deaths Total @@;
datalines;
0 0 3 1 0 3 2 0 3 3 0 3 4 1 3 5 2 3 ;
proc logistic data=dose descending; model Deaths/Total=Dose;
exact Dose/estimate=both; run;
```

# Conditional Logistic Regression

Parameter DF Estimate S.E Chi-Square Pr > ChiSq

Intercept 1 -9.4745 5.5677 2.8958 0.0888

Dose 1 2.0804 1.2603 2.7249 0.0988

Odds Ratio Estimates

Effect Estimate 95% Confidence Limits

Dose 8.007 0.677 94.679

Exact Parameter Estimates

Parameter Estimate 95% Confidence Limits p-Value

Dose 1.8000 0.1157 5.8665 0.0245

Exact Odds Ratios

Parameter Estimate 95% Confidence Limits p-Value Dose 6.049 1.123 353.000

0.0245

# Binary Matched Pairs

We consider a method for **analyzing binary matched-pair data**.

- Due to the matching, the responses in the two samples are associated with each other.
- Let  $(Y_{i1}, Y_{i2})$  be the  $i$ th pair of observations,  $i = 1, \dots, n$  and

$$g[P(Y_{it} = 1)] = \alpha_i + \beta x_t \quad (7.8)$$

for  $i = 1, \dots, n$  and  $t = 1, 2$ , where  $g(\cdot)$  is the link function.

- For the logit link, we have

$$P(Y_{i1} = 1) = \frac{\exp(\alpha_i)}{1 + \exp(\alpha_i)}, \quad P(Y_{i2} = 1) = \frac{\exp(\alpha_i + \beta)}{1 + \exp(\alpha_i + \beta)}.$$

For the  $i$ th subject, the odds of  $Y_{i2} = 1$  are  $\exp(\beta)$  times the odds for  $Y_{i1}$ .

# Binary Matched Pairs

- Since the number of parameters in model (7.8) is greater than  $n$ , this causes difficulties.
- We use **conditional likelihood** is to eliminate the  $\alpha_i$ 's.
- Assuming independence of responses for different subjects and for the two observations on the same subject, the likelihood is given by

$$\prod_{i=1}^n \frac{\exp(\alpha_i y_{i1})}{1 + \exp(\alpha_i)} \frac{\exp((\alpha_i + \beta)y_{i2})}{1 + \exp(\alpha_i + \beta)}.$$

- $s_i = y_{i1} + y_{i2}$  is the **sufficient statistic** for  $\alpha_i$ ,  $i = 1, \dots, n$ .

# Binary Matched Pairs

- To estimate  $\beta$ , we eliminate the  $\alpha_i$ 's by conditioning on all the  $s_i$ 's.
- Note that  $P(Y_{i1} = Y_{i2} = 0|s_i = 0) = P(Y_{i1} = Y_{i2} = 1|s_i = 2) = 1$  and

$$P(Y_{i1}, Y_{i2}|s_i = 1) = \begin{cases} \exp(\beta)/[1 + \exp(\beta)], & y_{i1} = 0, y_{i2} = 1; \\ 1/[1 + \exp(\beta)], & y_{i1} = 1, y_{i2} = 0. \end{cases}$$

# Binary Matched Pairs

- The conditional likelihood is given by

$$\prod_{s_i=1} \left( \frac{1}{1 + \exp(\beta)} \right)^{y_{i1}} \left( \frac{\exp(\beta)}{1 + \exp(\beta)} \right)^{y_{i2}} = \frac{\exp(\beta n_{21})}{[1 + \exp(\beta)]^{n^*}},$$

where  $n_{21} = \#\{i : y_{i1} = 0, y_{i2} = 1\}$  and  $n^* = \#\{i : y_{i1} + y_{i2} = 1\}$ .

- $\hat{\beta} = \log(n_{21}/(n^* - n_{21}))$  and its standard error is given by

$$se(\hat{\beta}) = \sqrt{1/n_{21} + 1/n_{12}}.$$

## Binary Matched Pairs

- The two observations in a matched pair need not refer to the same subject. For instance, it is common to match a single control with a case according to specific criteria in a case-control study.
- $\text{logit}[P(Y_{it} = 1)] = \alpha_i + \beta_1 x_{1it} + \beta_2 x_{2it} + \dots + \beta_p x_{pit}$ , in the  $i$ th pair for  $t = 1, 2$ .
- We can eliminate  $\alpha_i$  from the likelihood and then use the **conditional ML approach** to estimate  $\beta_j$ .

# Binary Matched Pairs

$$P(Y_{i1} = 0, Y_{i2} = 1 | S_i = 1) = \exp(\mathbf{x}_{i2}^T \boldsymbol{\beta}) / [\exp(\mathbf{x}_{i1}^T \boldsymbol{\beta}) + \exp(\mathbf{x}_{i2}^T \boldsymbol{\beta})]$$

and

$$P(Y_{i1} = 1, Y_{i2} = 0 | S_i = 1) = \exp(\mathbf{x}_{i1}^T \boldsymbol{\beta}) / [\exp(\mathbf{x}_{i1}^T \boldsymbol{\beta}) + \exp(\mathbf{x}_{i2}^T \boldsymbol{\beta})],$$

where  $\mathbf{x}_{it} = (x_{1it}, \dots, x_{pit})^T$ .

# Binary Matched Pairs



$$P(Y_{i1} = 0, Y_{i2} = 1 | S_i = 1) = \exp((\mathbf{x}_{i2} - \mathbf{x}_{i1})^T \boldsymbol{\beta}) / [\exp((\mathbf{x}_{i2} - \mathbf{x}_{i1})^T \boldsymbol{\beta}) + 1],$$

which has the form of a logistic regression with no intercept and with predictor values  $\mathbf{x}_i^* = \mathbf{x}_{i2} - \mathbf{x}_{i1}$ .

- One can obtain conditional ML estimate by fitting a logistic regression model to the artificial response  $y_i^* = 1$  when  $(y_{i1} = 0, y_{i2} = 1)$ ,  $y_i^* = 0$  when  $(y_{i1} = 1, y_{i2} = 0)$ , no intercept, and  $\mathbf{x}_i^*$ .

## Example 7.5

We considered a subset of the data from the Los Angeles Study of Endometrial Cancer (Breslow and Day, 1980). This dataset consists of 63 matched pairs (a case of endometrial cancer and a control) and two prognostic factors: Gall bladder disease (yes or no) and Hypertension (yes or no). We are interested in determining the relative risk for gall bladder disease controlling for the effect of hypertension.

# Binary Matched Pairs

```
proc logistic data=Data1; strata ID;
model outcome(event='1')=Gall / clodds=Wald; run;
proc logistic data=Data1; strata ID;
model outcome(event='1')=Gall Hyper /clodds=Wald; run;
proc logistic data=Data1 exactonly; strata ID;
model outcome(event='1')=Gall; exact Gall / estimate=both; run;
proc logistic data=Data1 exactonly; strata ID;
model outcome(event='1')=Gall Hyper; exact Gall Hyper / jointonly
estimate=both; run;
```

# Conditional Likelihood

- Consider a statistical model with  $\xi = (\psi, \lambda)$ , in which  $\psi$  is the parameter of interest and  $\lambda$  is the nuisance parameter.
- The key idea of conditional likelihood is to identify a statistic  $\mathbf{s}$  such that the conditional distribution of the full data  $\mathbf{Y}$  given  $\mathbf{s}$  depends only on  $\psi$ .
- Suppose  $p(\mathbf{Y}, \mathbf{s}; \xi) = p(\mathbf{Y}|\mathbf{s}; \psi)p(\mathbf{s}; \xi)$ . Thus, the conditional likelihood method uses  $\ell_n(\psi) = \log p(\mathbf{Y}|\mathbf{s}; \psi)$  as the 'pseudo' likelihood function to carry out inference about  $\psi$ .

# Conditional Likelihood

## Example 7.6

Assume that  $y_1, \dots, y_n$  are independent and  $y_i$  follows a Poisson distribution with mean  $\exp(\lambda + \psi x_i)$ , where  $x_i$  is a covariate of interest. Suppose that  $\lambda$  is the nuisance parameter and  $\psi$  is the parameter of interest. The quantity  $s = \sum_{j=1}^n y_j$  is a **sufficient statistic** for  $\lambda$ . The log-likelihood function of the conditional distribution of  $\mathbf{Y}$  given  $s = \sum_{j=1}^n y_j$  is given by

$$\psi \sum_{i=1}^n x_i y_i - s \log \left[ \sum_{j=1}^n \exp(\psi x_j) \right],$$

which is independent of  $\lambda$ .

# Conditional Likelihood

Since  $p(\mathbf{s}; \xi)$  is discarded, we may lose some information about  $\psi$  that is contained in  $p(\mathbf{s}; \xi)$ .

## Example 7.7

Consider pairs of independent random variables  $(y_{i1}, y_{i2})$ ,  $i = 1, \dots, n$  such that both  $y_{i1}$  and  $y_{i2}$  follow a  $N(\mathbf{x}_i^T \lambda, \psi)$  distribution, where  $\mathbf{x}_i$  is a covariate vector of interest. Thus,  $s_i = y_{i1} + y_{i2}$  follows a  $N(2\mathbf{x}_i^T \lambda, 2\psi)$  distribution. Since  $y_i = y_{i1} - y_{i2}$  and  $s_i$  is a one-to-one transformation of  $(y_{i1}, y_{i2})$ , it can be shown that  $y_i$  and  $s_i$  are independent and  $y_i$  given  $s_i$  follows a  $N(0, 2\psi)$  distribution. By conditioning on  $\mathbf{s} = (s_1, \dots, s_n)$ , we lose information about  $\psi$  from the distribution of  $\mathbf{s}$ .

# Conditional Likelihood

- If the distribution of  $s$  does not contain any information about  $\psi$ ,  $s$  is called **ancillary for  $\psi$  in the presence of  $\lambda$** .
- A statistic  $s$  is said to be  **$S$ -ancillary for  $\psi$  in the presence of  $\lambda$**  if the family of  $\{p(s; \psi, \lambda) : \lambda \in \Lambda\}$  is the same for each  $\psi$ .

# Conditional Likelihood

A statistic is said to be  $P$ -ancillary for  $\psi$  in the presence of  $\lambda$  if the partial information for  $\psi$  based on the distribution of  $s$  alone is zero. That is,

$$I_\psi(\xi; s) = I_{\psi\psi} - I_{\psi\lambda} I_{\lambda\lambda}^{-1} I_{\lambda\psi} = 0, \quad (7.9)$$

where

$$I(\psi, \lambda) = \begin{pmatrix} I_{\psi\psi} & I_{\psi\lambda} \\ I_{\lambda\psi} & I_{\lambda\lambda} \end{pmatrix},$$

and  $I_{\psi\psi}$ ,  $I_{\psi\lambda}$ , and  $I_{\lambda\lambda}$  are the appropriate elements of the Fisher information matrix of  $(\psi, \lambda)$  based on  $s$ .

- $I_\psi(\xi; s) \equiv I(\psi|\lambda)$  is the Fisher information for  $\psi$  given  $\lambda$  is known.
- $I(\psi|\lambda)^{-1}$  is the Cramer-Rao lower bound for the asymptotic covariance matrix of unbiased estimators of  $\psi$  when  $\lambda$  is known.
- Also note that the upper left block of  $I(\psi, \lambda)^{-1}$  is  
$$I(\psi|\lambda)^{-1} = (I_{\psi\psi} - I_{\psi\lambda} I_{\lambda\lambda}^{-1} I_{\lambda\psi})^{-1}.$$

# Conditional Likelihood

## Example 7.8

Since  $\mathbf{s}$  follows Poisson distribution with mean  $\sum_{i=1}^n \exp(\lambda + \psi x_i)$ ,  $\mathbf{s}$  is  $P$ -ancillary for  $\psi$  in the presence of  $\lambda$ . The log-likelihood function of  $\mathbf{s}$  is given by  $\ell_n(\psi, \lambda; \mathbf{s}) = \mathbf{s} \log \sum_{i=1}^n \exp(\lambda + \psi x_i) - \sum_{i=1}^n \exp(\lambda + \psi x_i)$ . We have

$$I_{\psi\psi} = \frac{[\sum_{i=1}^n x_i \exp(g_i)]^2}{\sum_{i=1}^n \exp(g_i)}, \quad I_{\psi\lambda} = \sum_{i=1}^n x_i \exp(g_i), \quad I_{\lambda\lambda} = \sum_{i=1}^n \exp(g_i),$$

where  $g_i = \lambda + \psi x_i$ . Thus,  $I_\psi(\xi; \mathbf{s}) = 0$ .

## Example 7.9

Consider a linear regression model  $y_i = \lambda + \mathbf{x}_i^T \psi + \epsilon_i$ , where  $\epsilon_i \sim N(0, 1)$ . Suppose  $\psi \in R^p$  is the parameter of interest and  $\lambda > 0$  is the nuisance parameter. Let  $s = \sum_{i=1}^n y_i$ . Then  $s$  follows a  $N(n(\lambda + \bar{\mathbf{x}}^T \psi), n)$  distribution, where  $\bar{\mathbf{x}} = \sum_{i=1}^n \mathbf{x}_i / n$ . It can be shown that  $I_\psi(\xi; s) = 0$ . Thus,  $s$  is **P-ancillary, but it is not S-ancillary** since the set of all possible distributions of  $s$  depends on  $\psi$ .

- An important question is to identify an  $\mathbf{s}$  such that using  $p(\mathbf{Y}|\mathbf{s}; \psi)$  does not lead to loss of information on  $\psi$ .
- Assume that there exists a statistic  $\mathbf{s}_\lambda(\psi)$  such that  $\mathbf{s}_\lambda(\psi_0)$  is sufficient for  $\lambda$  and complete for each value  $\psi_0$  of  $\psi$ .
- If  $\mathbf{s}_\lambda(\psi)$  is independent of  $\psi$ , then one can use  $\ell_c(\psi) = \log p(\mathbf{Y}|\mathbf{s}_\lambda, \psi) = \log p(\mathbf{Y}|\xi) - \log p(\mathbf{s}_\lambda; \xi)$ .
- This setting includes exact inference for odds ratios and the conditional logistic regression model.

# Conditional Likelihood

- If  $s_\lambda(\psi)$  does depend on  $\psi$ , the conditional distribution of  $\mathbf{Y}$  given  $s_\lambda(\psi)$  is not well defined.
- The conditional distribution of  $\mathbf{Y}$  given  $s_\lambda(\psi_0)$ , denoted by  $p(\mathbf{Y}|s_\lambda(\psi_0), \xi)$ , also depends on  $\psi_0$ .
- For each fixed  $\psi_0$ , we may use
$$\ell_c(\xi, \psi_0) = \log p(\mathbf{Y}|s_\lambda(\psi_0), \xi) = \log p(\mathbf{Y}|\xi) - \log p(s_\lambda(\psi_0); \xi)$$
as a log-likelihood function.
- $\ell_c^*(\psi) = \ell_c(\xi, \psi_0)|_{\psi_0=\psi}$  is **not a log-likelihood function in general**.

## Example 7.10

Suppose that  $y_1, \dots, y_n$  are i.i.d  $N(\mu, \sigma^2)$ , where  $n \geq 3$ ,  $\psi = \mu$  and  $\lambda = \sigma^2$ . Since  $\ell(\xi) = -\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu)^2 - \frac{n}{2} \log \sigma^2$ ,  $s(\mu) = \sum_{i=1}^n (y_i - \mu)^2$  is a complete sufficient statistic of  $\sigma^2$  for each  $\mu$ .

# Conditional Likelihood

- First, under the assumption that  $y_i \sim N(\mu, \sigma^2)$ , the statistic  $s(\mu_0) = \sum_{i=1}^n (y_i - \mu_0)^2 = \sum_{i=1}^n (y_i - \mu + \mu - \mu_0)^2$  has a non-central  $\chi^2$  distribution with  $n$  degrees of freedom.
- Second, letting  $\delta = n(\mu - \mu_0)^2$ ,  $\ell_c(\mu, \sigma^2; \mu_0) = \log p(\mathbf{Y}|\mu, \sigma^2) - \log p(s(\mu_0); \mu, \sigma^2)$  is given by

$$\begin{aligned} & -\frac{1}{2\sigma^2} \{s(\mu) - s(\mu_0)\} - (0.5n - 1) \log s(\mu_0) \\ & + \frac{\delta}{2\sigma^2} - \log \sum_{r=0}^{\infty} \left( \frac{\delta}{2\sigma^2} \right)^r \frac{r^r}{r!} B(0.5(n-1), r+0.5). \end{aligned}$$

# Conditional Likelihood

- Consider a **conditional score statistic**  $U_\psi(\xi) = \left. \frac{\partial \ell_c(\xi, \psi_0)}{\partial \psi} \right|_{\psi_0=\psi}$ .
- $U_\psi(\xi) = \partial_\psi \log p(\mathbf{Y}|\xi) - E[\partial_\psi \log p(\mathbf{Y}|\xi)|\mathbf{s}_\lambda(\psi_0)]|_{\psi_0=\psi}$ .
- Since  $p(\mathbf{Y}|\xi) = p(\mathbf{Y}|\mathbf{s}_\lambda(\psi_0), \xi)p(\mathbf{s}_\lambda(\psi_0)|\xi)$ ,

$$\begin{aligned} & E[\partial_\psi \log p(\mathbf{Y}|\xi)|\mathbf{s}_\lambda(\psi_0)] \\ &= E[\partial_\psi \log p(\mathbf{Y}|\mathbf{s}_\lambda(\psi_0), \xi)|\mathbf{s}_\lambda(\psi_0)] + E[\partial_\psi \log p(\mathbf{s}_\lambda(\psi_0); \xi)|\mathbf{s}_\lambda(\psi_0)]. \end{aligned}$$

# Conditional Likelihood

- The first term of the above equation equals zero, whereas the second term equals  $\partial_\psi \log p(\mathbf{s}_\lambda(\psi_0); \xi)$ .
- The **conditional score statistic**  $U_\psi$  can be regarded as the residual of  $\partial_\psi \log p(\mathbf{Y}|\xi)$  under its best prediction  $\mathbf{s}_\lambda(\psi)$ .

# Conditional Likelihood

- Setting  $\mu_0 = \mu$ , we can obtain

$$\ell_c^*(\mu) = \ell_c(\mu, \sigma^2; \mu_0)|_{\mu_0=\mu} = -0.5(n-2) \log s(\mu).$$

- With some calculations, we have

$$\partial_\mu \ell_c^*(\mu) = \frac{(n-2)}{s(\mu)} \sum_{i=1}^n (y_i - \mu) \neq \partial_\mu \ell_c(\mu, \sigma^2; \mu_0) \Big|_{\mu_0=\mu} = \sigma^{-2} \sum_{i=1}^n (y_i - \mu).$$

- Let  $U_\psi(\xi) = \partial_\mu \ell_c(\mu, \sigma^2; \mu_0)|_{\mu_0=\mu}$ . We have

$$E(U_\psi(\xi)) = 0, \text{ and } \text{var}(U_\psi(\xi)) = n\sigma^{-2} = -E[\partial_\mu^2 \ell_c(\mu, \sigma^2; \mu_0)|_{\mu_0=\mu}].$$

The score statistic  $U_\psi(\xi)$  satisfies the following properties.

## Theorem 7.1

*Under some mild conditions,*

- (a)  $E_0\{U_\psi(\psi_0, \lambda)\} = 0$  for all  $\psi_0, \lambda_0$  and  $\lambda$ , where  $E_0$  denotes the expectation taken with respect to  $p(\mathbf{Y}, \xi_0)$ , in which  $\xi_0 = (\psi_0, \lambda_0)$ ;
- (b)  $E_0\{U_\psi(\xi_0)^{\otimes 2} + \partial_\psi U_\psi(\xi_0)\} = \mathbf{0}$ ;
- (c)  $U_\psi(\psi, \tilde{\lambda})$  is an unbiased estimating equation, where  $\tilde{\lambda} = \tilde{\lambda}(\psi, s_\lambda(\psi))$  is a measurable function of  $\psi$  and  $s_\lambda(\psi)$ .

Since  $U_\psi(\xi)$  may involve in  $\lambda$ , Theorem 7.1 (c) says that  $U_\psi(\psi, \tilde{\lambda})$  is **unbiased and leads to a consistent estimate** of  $\psi$  when  $\tilde{\lambda}$  is any measurable function of  $\psi$  and  $s_\lambda(\psi)$ . In practice, we can substitute  $\tilde{\lambda}$  such that  $\tilde{\lambda}(\psi) = \operatorname{argmax}_\lambda \log p(s_\lambda(\psi); \xi)$ .

- Models for Overdispersion
- Score Test Statistics for Overdispersion
- The Heteroscedastic Linear Model

# Models for Overdispersion

- Analyzing discrete data using standard generalized linear models often exhibits overdispersion.
- Overdispersion reflects the fact that the actual variability of the discrete data exceeds their nominal variances predicted by standard generalized linear models.
- Example:  $Y_i \sim B(n_i, \pi(\mathbf{x}_i))$ ,
  - The theoretical  $\text{Var}(y_i) = n_i\pi(\mathbf{x}_i)(1 - \pi(\mathbf{x}_i))$
  - The true/Empirical  $\text{Var}(y_i)$  is greater than  $n_i\pi(\mathbf{x}_i)(1 - \pi(\mathbf{x}_i))$ .
- Consequences:
  - A failure of appropriately modeling the variance-mean relationship in the exponential family
  - Underestimating the standard error of the parameter estimates in generalized linear models
  - Produce misleading results

## Statistical methods for handling overdispersion

- Assume a more general form for the variance function, which may include some additional parameter and the additional parameters may be estimated using **moment methods and quasi-likelihood methods**
- Two-level hierarchical models
  - Standard generalized linear models
  - The parameter in the standard generalized linear model is assumed to follow some distribution which contains additional parameters
  - Estimation methods include maximum likelihood

## Overdispersion Models based on Binomial GLM

- Overdispersion
  - $y_i \sim B(n_i, \pi(\mathbf{x}_i))$  and  $g(\pi(\mathbf{x}_i)) = \mathbf{x}_i^T \beta$
  - $E(y_i) = n_i \pi(\mathbf{x}_i)$  and  $Var(y_i) = n_i \pi(\mathbf{x}_i)(1 - \pi(\mathbf{x}_i))$
  - $Var(y_i) > n_i \pi(\mathbf{x}_i)(1 - \pi(\mathbf{x}_i))$
- $Var(y_i) = \phi n_i \pi(\mathbf{x}_i)(1 - \pi(\mathbf{x}_i))$ , where  $\phi = \sigma^2$  is a free parameter. If  $\phi > 1$ , then overdispersion occurs.

# Models for Overdispersion

Calculate the quasi-likelihood estimate  $\hat{\beta}_P$ :

- $E(y_i) = n_i\pi(\mathbf{x}_i)$  and  $\text{Var}(y_i) = \sigma^2 n_i\pi(\mathbf{x}_i)(1 - \pi(\mathbf{x}_i))$ .



$$\sum_{i=1}^N \partial_{\beta} E(y_i)^T [\text{Var}(y_i)]^{-1} (y_i - E(y_i)) = \sum_{i=1}^N \frac{n_i \partial \pi(\mathbf{x}_i)}{\partial \beta^T} \frac{(y_i - n_i \pi(\mathbf{x}_i))}{\sigma^2 n_i \pi(\mathbf{x}_i)(1 - \pi(\mathbf{x}_i))} = \mathbf{0}.$$

- quasi-likelihood equations:

$$S_n(\beta) = \sum_{i=1}^N \frac{(y_i - n_i \pi(\mathbf{x}_i))}{\pi(\mathbf{x}_i)(1 - \pi(\mathbf{x}_i))} \frac{\partial \pi(\mathbf{x}_i)}{\partial \beta} = \mathbf{0}.$$

# Models for Overdispersion

•

$$\frac{1}{\sigma^2} \partial_{\beta} S_n(\beta) = \frac{1}{\sigma^2} \sum_{i=1}^N \partial_{\beta} (y_i - n_i \pi(\mathbf{x}_i)) \frac{\partial_{\beta} \pi(\mathbf{x}_i)^T}{\pi(\mathbf{x}_i)(1 - \pi(\mathbf{x}_i))} + \frac{1}{\sigma^2} \sum_{i=1}^N (y_i - n_i \pi(\mathbf{x}_i)) \partial_{\beta} \frac{\partial_{\beta} \pi(\mathbf{x}_i)}{\pi(\mathbf{x}_i)(1 - \pi(\mathbf{x}_i))},$$



$$-\frac{1}{\sigma^2} E[\partial_{\beta} S_n(\beta)] = \frac{1}{\sigma^2} \sum_{i=1}^N \frac{n_i}{\pi(\mathbf{x}_i)(1 - \pi(\mathbf{x}_i))} \partial_{\beta} \pi(\mathbf{x}_i)^{\otimes 2}.$$

- **Newton-Raphson algorithm:**

$$\beta^{t+1} = \beta^t + \left\{ \left[ \sum_{i=1}^N \frac{n_i}{\pi(\mathbf{x}_i)(1 - \pi(\mathbf{x}_i))} \partial_{\beta} \pi(\mathbf{x}_i)^{\otimes 2} \right]^{-1} S_n(\beta) \right\}_{\beta^t}.$$

# Models for Overdispersion

## Moment estimate

- $E \left[ \sum_{i=1}^N \frac{(y_i - n_i \pi(\mathbf{x}_i))^2}{\pi(\mathbf{x}_i)(1-\pi(\mathbf{x}_i))} \right] = n\sigma^2.$
- $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^N \frac{(y_i - n_i \hat{\pi}(\mathbf{x}_i))^2}{\hat{\pi}(\mathbf{x}_i)(1-\hat{\pi}(\mathbf{x}_i))}$ , where  $\hat{\pi}(\mathbf{x}_i) = \pi(\mathbf{x}_i^T \hat{\beta}_P)$ .
- $n^{-1} \sum_{i=1}^N E \left[ \frac{(y_i - n_i \hat{\pi}(\mathbf{x}_i))^2}{\hat{\pi}(\mathbf{x}_i)(1-\hat{\pi}(\mathbf{x}_i))} \right] \approx n^{-1} \sum_{i=1}^N \frac{n_i \sigma^2 \hat{\pi}(\mathbf{x}_i)(1-\hat{\pi}(\mathbf{x}_i))}{\hat{\pi}(\mathbf{x}_i)(1-\hat{\pi}(\mathbf{x}_i))} = \sigma^2.$

# Models for Overdispersion

Asymptotic distribution of  $\hat{\beta}_P$ :

- $\mathbf{0} = S_n(\hat{\beta}_P) = S_n(\beta_*) + \partial_\beta S_n(\beta_*)(\hat{\beta}_P - \beta_*)[1 + o_p(1)]$ , where  $\beta_*$  is the true value of  $\beta$ .
- $\hat{\beta}_P - \beta_* = [-\partial_\beta S_n(\beta_*)]^{-1} S_n(\beta_*)[1 + o_p(1)]$ .
- $-n^{-1} \partial_\beta S_n(\beta_*) \approx \sum_{i=1}^N \frac{n_i}{\hat{\pi}(\mathbf{x}_i)(1-\hat{\pi}(\mathbf{x}_i))} \partial_\beta \hat{\pi}(\mathbf{x}_i)^{\otimes 2} / n$ .
- $\frac{1}{\sigma^2} \text{Cov}(S_n(\beta_*)/\sqrt{n}) \approx \frac{1}{\sigma^2} \sum_{i=1}^N \frac{n_i}{\hat{\pi}(\mathbf{x}_i)(1-\hat{\pi}(\mathbf{x}_i))} \partial_\beta \hat{\pi}(\mathbf{x}_i)^{\otimes 2} / n = \frac{1}{n} I(\beta)$ .
- The covariance matrix of  $\hat{\beta}_P$  can be estimated by  
$$\text{Cov}(\hat{\beta}_P) \approx \hat{\sigma}^2 \left[ \sum_{i=1}^N \frac{n_i}{\hat{\pi}(\mathbf{x}_i)(1-\hat{\pi}(\mathbf{x}_i))} \partial_\beta \hat{\pi}(\mathbf{x}_i)^{\otimes 2} \right]^{-1}$$
.

## Two-level Hierarchical Model

- Level 1:  $Y_i|P_i \sim B(n_i, P_i)$
- Level 2:  $P_i$  are independent random variables with  $E(P_i) = \pi(\mathbf{x}_i)$  and  $Var(P_i) = \sigma^2\pi(\mathbf{x}_i)(1 - \pi(\mathbf{x}_i))$
- $E(Y_i) = n_i\pi(\mathbf{x}_i)$  and  $Var(Y_i) = n_i\pi(\mathbf{x}_i)(1 - \pi(\mathbf{x}_i))[1 + \sigma^2(n_i - 1)]$ .
- Note that the marginal distribution of  $Y_i$  is

$$p(y_i) = \int p(y_i|p_i)p(p_i)dp_i.$$

## beta-binomial distribution

- In level 2,  $P_i$  are iid,  $P_i \sim Beta(\alpha, \beta)$  and

$$p(P_i; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} P_i^{\alpha-1} (1 - P_i)^{\beta-1}, \quad (8.1)$$

where  $0 \leq P_i \leq 1$  and the parameters  $\alpha > 0$  and  $\beta > 0$ .

- Now make the 1-1 reparameterization  $\pi = \alpha/(\alpha + \beta)$  and  $\rho = 1/(\alpha + \beta)$ .
- Thus,  $\alpha = \frac{\pi}{\rho}$  and  $\beta = \frac{1-\pi}{\rho}$ .
- Also,  $\pi = \alpha/(\alpha + \beta) = E(P_i)$ .

# Models for Overdispersion

- The marginal distribution of  $Y_i$  is the beta-binomial distribution given by

$$P(Y_i = y_i; n_i, \alpha, \beta) = \binom{n_i}{y_i} \frac{B(\alpha + y_i, n_i + \beta - y_i)}{B(\alpha, \beta)}, \quad y_i = 0, 1, \dots, n_i,$$

where  $\alpha = \frac{\pi}{\rho}$  and  $\beta = \frac{1-\pi}{\rho}$ .

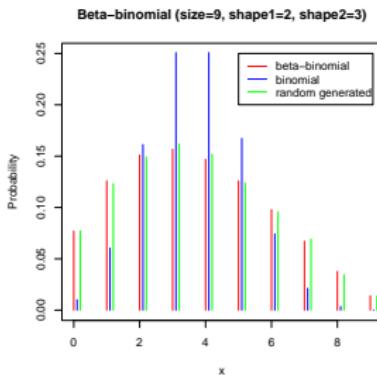
- To construct the beta-binomial regression model, formally set  $\pi \equiv \pi(\mathbf{x}_i)$ , where  $\pi(\mathbf{x}_i)$  can be based on the logit link, for example.
- Marginally,  $E(Y_i) = n_i\pi(\mathbf{x}_i)$   
and

$$\text{Var}(Y_i) = n_i\pi(\mathbf{x}_i)(1 - \pi(\mathbf{x}_i))[1 + (n_i - 1)\rho/(1 + \rho)].$$

## Example 8.1

We plot the beta-binomial probability mass function  $P(Y = y; n = 9, \alpha = 2, \beta = 3)$  and the binomial probability mass function  $P(Y = y; n = 9, 0.4)$ .

# Models for Overdispersion



**Figure:** Figure 7.1: Plots of the beta-binomial mass function  $P(Y = y; n = 9, \alpha = 2, \beta = 3)$  and the binomial probability mass function  $P(Y = y; n = 9, 0.4)$ .

# Models for Overdispersion

- We write the **beta-binomial regression** model as

$$y_i \sim BB(n_i; \pi(\mathbf{x}_i), \rho) \text{ and } \text{logit}(\pi(\mathbf{x}_i)) = \mathbf{x}_i^T \boldsymbol{\beta}. \quad (8.2)$$

- Random effects models** assume that  $\text{logit}(P_i) = \mathbf{x}_i^T \boldsymbol{\beta} + \sigma z_i$
- $z_i$  is a random variable with  $E(z_i) = 0$  and  $\text{Var}(z_i) = 1$
- Logistic-normal model** assumes  $z_i \sim N(0, 1)$
- Mixture logistic regression** assumes that  $z_i$  follows a discrete mixing distribution
- Estimation:** expectation-Maximization (EM) and stochastic approximation algorithms
- Likelihood theory**

## Example 8.2

*Consider the vegetation data. We are interested in using the quasibinomial model to model the relationship between the response  $V_2$  (or  $V_1$ ) and the climate variables  $X_1, \dots, X_5$ .*

# Models for Overdispersion

```
> outbim<-glm(V3~V4+V5+V6+V7+V8, family=quasibinomial, data=aa)
> summary(outbim)
Call: glm(formula = V3 ~ V4 + V5 + V6 + V7 + V8, family = quasibinomial, data = a1)
Deviance Residuals:
Min 1Q Median 3Q Max
-1.5162 -0.8380 -0.6208 1.2099 2.1328
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.5509809 1.0821970 1.433 0.152253
V4 0.0010865 0.0209727 0.052 0.958698
V5 -0.0011755 0.0004662 -2.521 0.011907 *
V6 -0.0026909 0.0067443 -0.399 0.690019
V7 -0.0032304 0.0027286 -1.184 0.236850
V8 -0.0010959 0.0002852 -3.843 0.000133 ***
(Dispersion parameter for quasibinomial family taken to be 0.9600969) Null deviance: 861.89 on 706 degrees of freedom Residual deviance: 793.36 on 701 degrees of freedom AIC: NA
> summary(outbim)$dispersion [1] 0.9600969
```

# Models for Overdispersion

## Example 8.3

We consider a dataset concerning the effects of chemical agents or dietary regimens on fetal development in laboratory rats. Female rats were put in iron-deficient diets and divided into 4 groups. One group of controls was given weekly injections of iron supplement to bring their iron intake to normal levels, while another group was given only placebo injections. Two other groups were given fewer iron-supplement injections than the controls. The rats were made pregnant, sacrificed 3 weeks later, and the total number of fetuses and the number of dead fetuses in each litter were counted.

# Models for Overdispersion

```
> fit2 = vglm(cbind(R,N-R)~fgrp*hb, betabin.ab(zero=2), data=lirat, trace=TRUE, subset=N>1)
> coef(fit2, matrix=TRUE)
log(shape1) log(shape2)
(Intercept) 2.36756824 -0.04226918
fgrp2 -7.41015007 0.00000000
fgrp3 -0.65608824 0.00000000
fgrp4 -9.61776494 0.00000000
hb -0.26482255 0.00000000
fgrp2:hb 0.68968413 0.00000000
fgrp3:hb -0.08016316 0.00000000
fgrp4:hb 0.61415957 0.00000000
> coef(fit2, matrix=TRUE)[,1] - coef(fit2, matrix=TRUE)[,2] logit(p)
(Intercept) fgrp2 fgrp3 fgrp4 hb fgrp2:hb
2.40983742 -7.41015007 -0.65608824 -9.61776494 -0.26482255 0.68968413
fgrp3:hb fgrp4:hb
-0.08016316 0.61415957
```

## Models for Overdispersion

For each litter, the number of dead fetuses may be considered to be  $\text{Binomial}(n, p)$  where  $n$  is the litter size and  $p$  is the probability of a fetus dying. The parameter  $p$  is expected to vary from litter to litter, even when the covariates of hemoglobin level and experimental group are accounted for.

# Models for Overdispersion

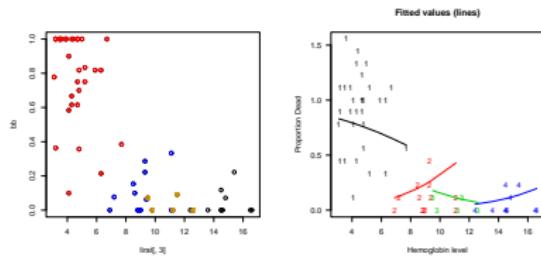


Figure: Top: plot of  $y_i/n_i$  against hemoglobin level in four groups (1: red; 2: blue; 3: orange; 4: dark); Bottom: four fitted lines.

# Models for Overdispersion

- Poisson Regression:  $y_i \sim \text{Poisson}(\mu_i)$  and  $g(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta}$ , where  $g(\cdot)$  is a given link function.
- $E(y_i) = \text{Var}(y_i) = \mu_i$
- Overdispersion:  $\text{Var}(y_i) > \mu_i = E(y_i)$
- $\text{Var}(y_i) = \sigma^2 \mu_i$ .

# Models for Overdispersion

## Quasi-likelihood Method:

- $E(y_i) = \mu_i = g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})$  and  $\text{Var}(y_i) = \sigma^2 \mu_i$ , where  $\sigma^2 > 1$  is a free parameter.
- $S_n(\boldsymbol{\beta}) = \sum_{i=1}^n \frac{\partial \mu_i}{\partial \boldsymbol{\beta}^T} \frac{(y_i - \mu_i)}{\mu_i} = \mathbf{0}$ .
- $\sum_{i=1}^n \partial_{\boldsymbol{\beta}} E(y_i)^T [\text{Var}(y_i)]^{-1} (y_i - E(y_i)) = \sum_{i=1}^n \partial_{\boldsymbol{\beta}} \mu_i^T \frac{(y_i - \mu_i)}{\sigma^2 \mu_i} = \mathbf{0}$ .

# Models for Overdispersion



$$\frac{1}{\sigma^2} \partial_{\beta} S_n(\beta) = \frac{1}{\sigma^2} \sum_{i=1}^n \partial_{\beta}(y_i - \mu_i) \frac{\partial_{\beta} \mu_i^T}{\mu_i} + \frac{1}{\sigma^2} \sum_{i=1}^n (y_i - \mu_i) \partial_{\beta} \frac{\partial_{\beta} \mu_i}{\mu_i}.$$

- We have

$$-\frac{1}{\sigma^2} E[\partial_{\beta} S_n(\beta)] = \frac{1}{\sigma^2} \sum_{i=1}^n \mu_i^{-1} \partial_{\beta} \mu_i^{\otimes 2}.$$



$$\beta^{k+1} = \beta^k + \{ [\sum_{i=1}^n \mu_i^{-1} \partial_{\beta} \mu_i^{\otimes 2}]^{-1} S_n(\beta) \}_{\beta^k},$$

for  $k = 1, \dots$ .

# Models for Overdispersion

## Moment estimate

- $E \left[ \sum_{i=1}^n \frac{(y_i - \mu_i)^2}{\mu_i} \right] = n\sigma^2$  leads to  $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}$ .
- $Cov(\hat{\beta}_P) \approx \hat{\sigma}^2 \left[ \sum_{i=1}^n \hat{\mu}_i^{-1} \partial_\beta \mu_i^{\otimes 2} \right]^{-1}$ .

# Models for Overdispersion

## Two-level hierarchical model

- Level 1:  $Y_i|\lambda_i \sim Poisson(\lambda_i)$
- Level 2:  $\lambda_i$  are independent random variables with  $E(\lambda_i) = \mu_i$  and  $Var(\lambda_i) = \sigma_i^2$ . This gives  $E(Y_i) = \mu_i$  and  $Var(Y_i) = \mu_i + \sigma_i^2$ .

# Models for Overdispersion

Negative-binomial distribution:

- $\lambda_i \sim \text{Gamma}(\mu_i, k)$ . We have  $E(\lambda_i) = \mu_i$  and  $\text{Var}(\lambda_i) = \mu_i^2/k$ .
- $Y_i \sim NB(k, \mu_i)$ .
- $E(Y_i) = \mu_i$  and  $\text{Var}(Y_i) = \mu_i + \mu_i^2/k$ .
- **Negative-binomial regression**:  $y_i \sim NB(k, \mu_i)$  and  $\log(\mu_i) = \mathbf{x}_i^T \beta$ .
- **Poisson random effects model**:  $Y_i | \lambda_i \sim \text{Poisson}(\lambda_i)$  and  $\log(\lambda_i) = \mathbf{x}_i^T \beta + \sigma z_i$ , where  $z_i$  is a random variable with  $E(z_i) = 0$  and  $\text{Var}(z_i) = 1$ .

# Models for Overdispersion

## Example 8.4

We consider a damage incidents dataset consisting of 34 observations. For each observation, the response is number of damage incidents, and covariates of interest include ship type ( $x_{i1}$ ), year of construction ( $x_{i2}$ ), period of operation ( $x_{i3}$ ), and logarithm of aggregate months service ( $x_{i4}$ ).

We first use a “quasipoisson” model to fit the data and estimate  $\beta$ , and then use the moment method to estimate  $\sigma^2$ . Finally, we estimate the covariance matrix of  $\hat{\beta}_P$ .

# Models for Overdispersion

```
> table62<-read.table("your data", col.names=c("Ship", "Year", "Period", "Months", "Response"))
> summary(log.fit) log.fit<-glm(Response~log(Months)+factor(Period)+factor(Year)*factor(Ship),
family=quasipoisson(link=log),data=table62)
> summary(log.fit) Call: glm(formula = Response ~ log(Months) + factor(Period) + factor(Year) * factor(Ship),
family = quasipoisson(link = log), data = table62) Deviance Residuals: Min 1Q Median 3Q Max -1.891200
-0.104459 -0.000085 0.104534 2.257704
Estimate Std. Error t value Pr(>|t|)
(Intercept) -23.0736 7339.6145 -0.003 0.99754
log(Months) 0.7997 0.1961 4.079 0.00153 **
factor(Period)2 0.3553 0.1316 2.701 0.01927*
factor(Year)2 18.5364 7339.6145 0.003 0.99803 factor(Year)3 19.0895 7339.6145 0.003 0.99797
factor(Year)4 18.9458 7339.6145 0.003 0.99798
factor(Ship)2 18.2192 7339.6145 0.002 0.99806 factor(Ship)3 17.5358 7339.6145 0.002 0.99813
factor(Ship)4 -0.4765 10355.5013 -4.60e-05 0.99996 factor(Ship)5 0.7269 12735.5116 5.71e-05 0.99996
factor(Year)2:factor(Ship)2 -17.9143 7339.6145 -0.002 0.99809 factor(Year)3:factor(Ship)2 -18.5024 7339.6145
-0.003 0.99803
factor(Year)4:factor(Ship)2 -18.6498 7339.6145 -0.003 0.99801
```

# Models for Overdispersion

```
factor(Year)2:factor(Ship)3 -19.1452 7339.6146 -0.003 0.99796 factor(Year)3:factor(Ship)3 -18.1761 7339.6145  
-0.002 0.99806  
factor(Year)4:factor(Ship)3 -18.2521 7339.6146 -0.002 0.99806  
factor(Year)2:factor(Ship)4 -18.8330 12704.0254 -0.001 0.99884 factor(Year)3:factor(Ship)4 0.7643 10355.5013  
7.38e-05 0.99994  
factor(Year)4:factor(Ship)4 -0.4632 10355.5013 -4.47e-05 0.99997  
factor(Year)2:factor(Ship)5 0.4786 12735.5116 3.76e-05 0.99997 factor(Year)3:factor(Ship)5 -0.7595 12735.5116  
-5.96e-05 0.99995  
factor(Year)4:factor(Ship)5 -1.9887 12735.5117 -1.56e-04 0.99988  
(Dispersion parameter for quasipoisson family taken to be 1.219018) Null deviance: 614.539 on 33 degrees of  
freedom Residual deviance: 13.343 on 12 degrees of freedom AIC: NA  
> summary(log.fit)$dispersion
```

[1] 1.219018

# Score Test Statistics for Overdispersion

- Two-level hierarchical overdispersion model:
- $y_i|\theta_i \sim D(\theta_i, 1) \quad i = 1, \dots, n,$
- $\theta_i$  are random variables such that  $E(\theta_i) = k(\mathbf{x}_i^T \beta)$  and  $Var(\theta_i) = \tau f_i(\mathbf{x}_i^T \beta).$
- We also assume that  $E[\theta_i - k(\mathbf{x}_i^T \beta)]^r = o(\tau) \quad , r \geq 3.$
- As  $\tau \rightarrow 0$ , the overdispersion model reduces to a GLM.
- Let  $\alpha = (\beta, \tau)$ .  $\ell_n(\alpha) = \sum_{i=1}^n \log(\int p(y_i|\theta_i)p(\theta_i; \beta, \tau)d\theta_i).$

# Score Test Statistics for Overdispersion

- Homogeneity hypotheses:  $H_0 : \tau = 0$  vs.  $H_1 : \tau > 0$ .
- Score test for testing the null hypothesis:

$$S_\tau = \partial_\alpha \ell_n(\tilde{\alpha})^T \{I_{\alpha\alpha}(\tilde{\alpha})\}^{-1} \partial_\alpha \ell_n(\tilde{\alpha}) = \partial_\tau \ell_n(\alpha)^2 / \sigma_\tau^2 \mathbf{1}\{\partial_\tau \ell_n(\alpha) > 0\} \Big|_{\tilde{\alpha}},$$

where  $\sigma_\tau^2 = I_{\tau\tau} - I_{\tau\beta} I_{\beta\beta}^{-1} I_{\beta\tau}$  and  $\tilde{\alpha}$  is the estimate of  $\alpha$  under the null hypothesis  $H_0 : \tau = 0$ .

- As  $n \rightarrow \infty$ ,  $S_\tau$  converges to a  $0.5\chi_0^2 + 0.5\chi_1^2$  distribution, where  $\chi_0^2$  denotes a point mass at zero.

# Score Test Statistics for Overdispersion

- Let  $w_{1i} = \ddot{b}(.)$  and  $w_{2i} = 0.5f_i\ddot{\dot{b}}(.)$ .

- 

$$\partial_\tau \ell_n(\alpha) = \sum_{i=1}^n \frac{1}{2} f_i \{(y_i - \mu_i)^2 - \ddot{b}(.)\},$$

- 

$$I_{\beta\tau} = D_\theta(\beta)^T W_2 \mathbf{1}_n,$$

where  $D_\theta(\beta)$  is the  $n \times p$  matrix with  $(i,j)$ th element  $\frac{\partial \theta_i}{\partial \beta_j}$ .

# Score Test Statistics for Overdispersion

- $I_{\tau\tau} = \sum_{i=1}^n \frac{1}{4} f_i^2 \{2(\ddot{b}(.)^2 + b^{(4)}(.))\},$
- $I_{\beta\beta} = D_\theta(\beta)^T W_1 D_\theta(\beta),$
- $W_1 = \text{diag}(w_{11}, \dots, w_{1n})$  and  $W_2 = \text{diag}(w_{21}, \dots, w_{2n})$ ,  $D_\theta(\beta) = \partial\theta/\partial\beta$  is an  $n \times p$  matrix and  $b^{(4)}(.)$  denotes the fourth derivative of  $b(.)$  with respect to its argument.

# Score Test Statistics for Overdispersion

For the logistic random effects model,

$$I_{\tau\tau} = 0.25 \sum_{i=1}^n \{2m_i^2\pi_i^2(1-\pi_i)^2 + m_i\pi_i(1-\pi_i)(1-6\pi_i+6\pi_i^2)\}$$

$$f_i = m_i, \quad w_{1i} = m_i\pi_i(1-\pi_i), \quad w_{2i} = \frac{1}{2}m_i\pi_i(1-\pi_i)(1-2\pi_i), \quad D_\theta(\beta) = X,$$

$$\text{and } \partial_\tau \ell_n(\alpha) = \sum_{i=1}^n \{(y_i - m_i\pi_i)^2 - m_i\pi_i(1-\pi_i)\}.$$

Thus,  $\sigma_\tau^2 = I_{\tau\tau} - 0.25 \sum_{i=1}^n \sum_{j=1}^n h_{ij} w_{2i} w_{2j} / \sqrt{w_{1i} w_{1j}}$ , where

$$H = (h_{ij}) = W_1^{1/2} X (X^T W_1 X)^{-1} X^T W_1^{1/2}.$$

# Score Test Statistics for Overdispersion

For the Poisson random effects model,

$$f_i = 1, \quad I_{\tau\tau} = \frac{1}{4} \sum_{i=1}^n \{2\mu_i^2 + \mu_i\}, \quad w_{1i} = \mu_i, \quad w_{2i} = \frac{1}{2}\mu_i,$$

$$D_\theta(\beta) = X, \text{ and } \partial_\tau \ell_n(\alpha) = 0.5 \sum_{i=1}^n \{(y_i - \hat{\mu}_i)^2 - \hat{\mu}_i\}.$$

Thus, if  $X$  contains an intercept, then  $\sigma_\tau^2$  is given by  $0.5 \sum_{i=1}^n \hat{\mu}_i^2$  and

$$S_\tau = \frac{[\sum_{i=1}^n \{(y_i - \hat{\mu}_i)^2 - \hat{\mu}_i\}]^2}{[\sum_{i=1}^n \hat{\mu}_i^2]^2} \mathbf{1} \left( \sum_{i=1}^n \{(y_i - \hat{\mu}_i)^2 - \hat{\mu}_i\} > 0 \right).$$

# Chapter 9: Generalized Estimating Equations

- Quasi-likelihood
- Z-Estimators
- Generalized Estimating Equations (GEE)

- Quasi-likelihood provides an important method for making statistical inference without making parametric assumption.
- Quasi-likelihood can be applied to independent and dependent observations.

# Quasi-likelihood

- Suppose that the components of  $Y = (y_1, \dots, y_n)^T$  are independent and satisfy  $E(Y) = \mu = (\mu_1, \dots, \mu_n)^T$  and  $\text{Cov}(Y) = \sigma^2 V(\mu) = \sigma^2 \text{diag}(V_1(\mu_1), \dots, V_n(\mu_n))$ , where  $\sigma^2 = \phi^{-1}$  may be unknown,  $\mu$  is a known mean function, and  $V(\mu)$  is a known variance function. Moreover,  $\mu_i = \mu(x_i; \beta) = \mu_i(\beta)$ ,  $i = 1, \dots, n$ .
- How do we make statistical inferences about  $\beta$  given the above assumptions?

# Quasi-likelihood

- The **quasi log-likelihood** for  $\mu$  based on the data  $Y$  is given by

$$l_q(\mu, y) = \sum_{i=1}^n q_i(\mu_i, y_i), \quad q_i(\mu_i, y_i) = \int_{y_i}^{\mu_i} u_i(t) dt, \quad (9.1)$$

where  $u_i(t) = (y_i - t)/(\sigma^2 V_i(t))$ .

- The integral  $q_i(\mu_i, y_i)$  should behave like a log-likelihood function for  $\mu_i$  based on  $y_i$ .
- $U_i = (y_i - \mu_i)/(\sigma^2 V_i(\mu_i))$  has the following properties:

$$E(U_i) = 0, \quad \text{Var}(U_i) = 1/(\sigma^2 V_i(\mu_i)) = -E(\partial_{\mu_i} U_i). \quad (9.2)$$

# Quasi-likelihood

## Example 9.1

Suppose that  $y_1, \dots, y_n$  are independent random variables with mean  $\mu_i$  and variance  $\sigma^2\mu_i$ , where  $\sigma$  is unknown and  $\log(\mu_i) = \beta_1 + \beta_2x_i$ , in which  $x_i$  is the covariate of interest. We set up the quasi-likelihood function for  $(\beta_1, \beta_2)$ . Since  $V_i(\mu_i) = \mu_i$ ,  $V_i(t) = t$  if  $\mu_i = t$ . Thus,

$u_i(t) = (y_i - t)/(\sigma^2 V_i(t)) = (y_i - t)/(\sigma^2 t)$ , and therefore

$$\begin{aligned} I_q(\mu, y) &= \sum_{i=1}^n \int_{y_i}^{\mu_i} u_i(t) dt = \sum_{i=1}^n \int_{y_i}^{\mu_i} (y_i - t)/(\sigma^2 t) dt \\ &= \sum_{i=1}^n \sigma^{-2} [y_i \log(\mu_i/y_i) - (\mu_i - y_i)]. \end{aligned}$$

# Quasi-likelihood

- The maximum quasi-likelihood estimator  
 $\hat{\beta} = \text{argmax}_q l_q(\mu(\beta), y).$
- $\frac{\partial l_q}{\partial \mu} = \phi V^{-1}(\mu)(Y - \mu)$ ,  $\frac{\partial l_q}{\partial \beta} = \frac{\partial \mu}{\partial \beta}^T \frac{\partial l_q}{\partial \mu} = \phi D^T V^{-1} \mathbf{e}(\beta)$ ,  
where  $D = \partial \mu / \partial \beta^T$  and  $\mathbf{e} = \mathbf{Y} - \mu(\beta)$ .
- $S_n(\hat{\beta}) = D^T(\hat{\beta}) V^{-1}(\hat{\beta}) \mathbf{e}(\hat{\beta}) = \mathbf{0}$ ,

- The asymptotic covariance matrix for  $\hat{\beta}$  is given by  $Cov(\hat{\beta}) \approx \sigma^2(D^T V^{-1} D)^{-1}$ .
- $0 = S_n(\hat{\beta}) \approx S_n(\beta_*) + \partial_{\beta} S_n(\beta_*)(\hat{\beta} - \beta_*)$ .
- $-\partial_{\beta} S_n(\beta_*)(\hat{\beta} - \beta_*) = S_n(\beta_*)$
- $(\hat{\beta} - \beta_*) = [-\partial_{\beta} S_n(\beta_*)]^{-1} S_n(\beta_*)$ .
- $\sqrt{n}(\hat{\beta} - \beta_*) = [-n^{-1}\partial_{\beta} S_n(\beta_*)]^{-1} \frac{1}{\sqrt{n}} S_n(\beta_*)$ .
- $Cov(\sqrt{n}\hat{\beta}) \approx [-n^{-1}\partial_{\beta} S_n(\beta_*)]^{-1} Cov[\frac{1}{\sqrt{n}} S_n(\beta_*)] [-n^{-1}\partial_{\beta} S_n(\beta_*)]^{-1}$ .

# Quasi-likelihood

$$\begin{aligned} -n^{-1}\partial_{\beta}S_n(\beta_*) &= -n^{-1}\partial_{\beta}\left[\sum_{i=1}^n \partial_{\beta}\mu_i(\beta_*)V_i(\beta_*)^{-1}e_i(\beta_*)\right] \\ &= -n^{-1}\sum_{i=1}^n \partial_{\beta}[\partial_{\beta}\mu_i(\beta_*)V_i(\beta_*)^{-1}]e_i(\beta_*) \\ &\quad + n^{-1}\sum_{i=1}^n \partial_{\beta}\mu_i(\beta_*)V_i(\beta_*)^{-1}\partial_{\beta}\mu_i(\beta_*)^T \\ &\approx n^{-1}\sum_{i=1}^n \partial_{\beta}\mu_i(\beta_*)V_i(\beta_*)^{-1}\partial_{\beta}\mu_i(\beta_*)^T. \end{aligned}$$

# Quasi-likelihood

$$\begin{aligned}\text{Cov} \left[ S_n(\beta_*) / \sqrt{n} \right] &= \text{Cov} \left[ \sum_{i=1}^n \partial_\beta \mu_i(\beta_*) V_i(\beta_*)^{-1} e_i(\beta_*) / \sqrt{n} \right] \\ &= \sigma^2 \sum_{i=1}^n \partial_\beta \mu_i(\beta_*) V_i(\beta_*)^{-1} \partial_\beta \mu_i(\beta_*)^T / n = \sigma^2 D^T V^{-1} D / n,\end{aligned}$$

in which we have used the fact that  $\text{var}(e_i) = \sigma^2 V_i$ .

$$\text{Cov}(\hat{\beta}) = \text{Cov}(\sqrt{n} \hat{\beta}) / n \approx \sigma^2 (D^T V^{-1} D)^{-1}. \quad (9.3)$$

# Quasi-likelihood

- Since  $E \left[ \sum_{i=1}^n \frac{(y_i - \mu_i)^2}{V(\mu_i)} \right] = n\sigma^2$ , we can construct  
 $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}$ .

## Theorem 9.1

*The quasi-likelihood estimator is Godambe efficient for  $\beta$ .*

# Quasi-likelihood

- $S_n(\mathbf{y}, \beta) = \mathbf{H}^T(\mathbf{y} - \mu(\beta)) = \sum_{i=1}^n \mathbf{h}_i(y_i - \mu_i(\beta)),$  where  $\mathbf{H} = (\mathbf{h}_1, \dots, \mathbf{h}_n)^T.$
- $0 = S_n(\mathbf{y}, \tilde{\beta}) \approx S_n(\mathbf{y}, \beta_*) + \partial_\beta S_n(\mathbf{y}, \beta_*)(\tilde{\beta} - \beta_*).$
- $\sqrt{n}(\tilde{\beta} - \beta_*) = [-n^{-1}\partial_\beta S_n(\mathbf{y}, \beta_*)]^{-1} \frac{1}{\sqrt{n}} S_n(\mathbf{y}, \beta_*).$
- $\text{Cov}(\sqrt{n}\tilde{\beta}) \approx [-n^{-1}\partial_\beta S_n(\mathbf{y}, \beta_*)]^{-1} \text{Cov} \left[ \frac{1}{\sqrt{n}} S_n(\mathbf{y}, \beta_*) \right] [-n^{-1}\partial_\beta S_n(\mathbf{y}, \beta_*)]^{-1}.$

# Quasi-likelihood

$$\begin{aligned} -n^{-1}\partial_{\beta}S_n(\mathbf{y}, \beta_*) &= -n^{-1}\partial_{\beta}\left[\sum_{i=1}^n \mathbf{h}_i(\beta_*)\mathbf{e}_i(\beta_*)\right] \\ &= -n^{-1}\sum_{i=1}^n \partial_{\beta}\mathbf{h}_i\mathbf{e}_i(\beta_*) + n^{-1}\sum_{i=1}^n \mathbf{h}_i\partial_{\beta}\mu_i(\beta_*)^T \approx n^{-1}\sum_{i=1}^n \mathbf{h}_i\partial_{\beta}\mu_i(\beta_*)^T = \mathbf{H}^TD/n. \end{aligned}$$
  
$$\begin{aligned} Cov[S_n(\mathbf{y}, \beta_*)/\sqrt{n}] &= Cov\left[\sum_{i=1}^n \mathbf{h}_i(\beta)\mathbf{e}_i/\sqrt{n}\right] = \sigma^2\sum_{i=1}^n \mathbf{h}_i(\beta_*)V_i(\beta_*)\mathbf{h}_i(\beta_*)^T/n \\ &= \sigma^2\mathbf{H}^TV\mathbf{H}/n, \end{aligned}$$
  
$$Cov(\tilde{\beta}) \approx \sigma^2(\mathbf{H}^TD)^{-1}\mathbf{H}^TV\mathbf{H}(D^T\mathbf{H})^{-1}. \quad (9.4)$$

# Quasi-likelihood

$\text{Var}(\mathbf{a}^T \tilde{\beta}) \geq \text{Var}(\mathbf{a}^T \hat{\beta})$  for any  $\mathbf{a} \in R^p$ .

$$\{\text{Cov}(\hat{\beta})\}^{-1} - \{\text{Cov}(\tilde{\beta})\}^{-1} = D^T(V^{-1} - \mathbf{H}(\mathbf{H}^T V \mathbf{H})^{-1} \mathbf{H}^T)D\sigma^{-2} \geq 0.$$

The covariance matrix of  $[D^T V^{-1} - D^T \mathbf{H}(\mathbf{H}^T V \mathbf{H})^{-1} \mathbf{H}^T]Y$  is given by

$$\begin{aligned} & [D^T V^{-1} - D^T \mathbf{H}(\mathbf{H}^T V \mathbf{H})^{-1} \mathbf{H}^T] \text{Cov}(Y) [D^T V^{-1} - D^T \mathbf{H}(\mathbf{H}^T V \mathbf{H})^{-1} \mathbf{H}^T]^T \\ &= \sigma^2 D^T(V^{-1} - \mathbf{H}(\mathbf{H}^T V \mathbf{H})^{-1} \mathbf{H}^T)D. \end{aligned}$$

Since  $\{\text{Cov}(\hat{\beta})\}^{-1} - \{\text{Cov}(\tilde{\beta})\}^{-1} \geq 0$ ,  $\text{Cov}(\hat{\beta}) - \text{Cov}(\tilde{\beta})$  is non-positive definite.

## Example 9.2

Assume that  $E(y_i) = n_i\pi_i$ ,  $\text{Var}(y_i) = \sigma^2 n_i\pi_i(1 - \pi_i)$ , and  $g(\pi_i) = \mathbf{x}_i^T \beta$  for  $i = 1, \dots, N$ . Moreover,  $n = \sum_{i=1}^N n_i$ . Since  $V_i(\mu_i) = \mu_i - \mu_i^2/(n_i)$ , we have  $V_i(t) = t - t^2/(n_i)$  and

$$\ell_q(\mu, \mathbf{y}) = \sum_{i=1}^N \int_{y_i}^{\mu_i} u_i(t) dt = \sum_{i=1}^N \int_{y_i}^{\mu_i} (y_i - t)/[\sigma^2(t - t^2/n_i)] dt.$$

$$\partial_\beta \ell_q(\mu, \mathbf{y}) = S_n(\beta)/\sigma^2 = \sum_{i=1}^N \frac{(y_i - n_i\pi_i)}{\sigma^2 n_i\pi_i(1 - \pi_i)} \frac{n_i \partial \pi_i}{\partial \beta} = \mathbf{0}.$$

- For dependent data,  $V(\mu)$  is a symmetric  $n \times n$  matrix with entries  $V_{ij}(\mu)$  and we may solve  
$$G_n(\beta) = D^T(\beta)V^{-1}(\beta)(Y - \mu(\beta)) = \mathbf{0}.$$
- There is a **conceptual difference between independent and dependent data** for quasi-likelihood.
- For dependent data,  $G_n(\beta)$  may not be the gradient vector of the quasi-likelihood function.

- $I_q(\mu, \mathbf{y}, \mathbf{t}(s)) = \sigma^{-2} \int_{\mathbf{t}(s)=\mathbf{y}}^{\mathbf{t}(s)=\mu} (\mathbf{y} - \mathbf{t})^T \{V(\mathbf{t})\}^{-1} d\mathbf{t}(s)$ , where  $\mathbf{t}(s)$  is a smooth path from  $\mathbf{y}$  to  $\mu$ .
- To make the above **integral path-independent**, the sufficient conditions are  $\frac{\partial V^{ij}}{\partial \mu_k} = \frac{\partial V^{ik}}{\partial \mu_j} = \frac{\partial V^{kj}}{\partial \mu_i}$  for all  $i, j$  and  $k$ , where we denote  $V^{-1}$  by  $(V^{ij})$ .

# Z-Estimators

- Let  $U_1, \dots, U_n$  be independently and identically distributed as  $g(U_i)$ , which is unknown in practice.
- Consider the **Z estimator**  $\hat{\xi}$  satisfying

$$S_n(\hat{\xi}) = \sum_{i=1}^n \mathbf{h}(\hat{\xi}, U_i) = \mathbf{0}, \quad (9.5)$$

- Let  $S(\xi) = \int \mathbf{h}(\xi, U)g(U)dU$  and  $S(\xi_*) = 0$ .

# Z-Estimators

---

$$\begin{array}{ccc} n^{-1} S_n(\xi) & \rightarrow & S(\xi) \\ \hat{\xi} & \rightarrow & \xi_* \\ n^{-1} S_n(\hat{\xi}) = 0 & & S(\xi_*) = 0 \\ \hat{\xi} = \operatorname{argmax}_{\xi} - [n^{-1} S_n(\xi)]^2 & & \xi_* = \operatorname{argmax}_{\xi} - [S(\xi)]^2 \end{array}$$

---

## Example 9.3

Consider a quasi-likelihood model for  $n$  i.i.d observations  $(x_1, y_1), \dots, (x_n, y_n)$ . Suppose that  $E(y_i) = \mu(x_i, \beta)$  and  $\text{Var}(y_i) = V(x_i, \beta)$ . We estimate  $\hat{\beta}$  by solving

$$D^T(\beta)V^{-1}(\beta)\mathbf{e}(\beta) = \sum_{i=1}^n \partial_\beta \mu(x_i, \beta) V(x_i, \beta)^{-1} [y_i - \mu(x_i, \beta)] = 0.$$

- a) How do we establish consistency and asymptotic normality of  $\hat{\beta}$ ?
- b) How do we construct the covariance matrix of  $\hat{\beta}$ ?
- c) How do we construct statistics to test linear hypotheses regarding  $\beta$ ?

# Z-Estimators

- Assume that  $E(y_i|x_i) = \mathbf{x}_i^T \beta = \mu_i$  without any additional information about  $\text{Var}(y_i|x_i)$ .
- $D(\beta)^T \mathbf{e}(\beta) = \sum_{i=1}^n \partial_\beta \mu(\mathbf{x}_i, \beta)[y_i - \mu(\mathbf{x}_i, \beta)] = \mathbf{0}$ .
- The same set of three questions a)-c) can be asked for the estimate  $\tilde{\beta}$ , which satisfies  $D(\tilde{\beta})^T \mathbf{e}(\tilde{\beta}) = \mathbf{0}$ .
- In this case,  $\mathbf{h}(\xi, U) = \partial_\beta \mu(\mathbf{x}_i, \beta)[y_i - \mu(\mathbf{x}_i, \beta)]$ .

- **Consistency:**  $\hat{\xi}$  converges to  $\xi_*$  (the ‘true value’) in probability (or almost surely);
- **Asymptotic Normality:**  $\sqrt{n}(\hat{\xi} - \xi_*)$  converges to a multivariate normal distribution  $N(\mathbf{0}, \Sigma(\xi_*))$ .

## Theorem 9.2

If for every  $\epsilon > 0$ ,

$$\sup_{\xi \in \Xi} |n^{-1} S_n(\xi) - S(\xi)| \rightarrow^P 0, \quad \sup_{\xi: ||\xi - \xi_*|| \geq \epsilon} ||S(\xi)|| > ||S(\xi_*)|| = 0,$$

then  $\hat{\xi}$  converges to  $\xi_*$  in probability.

## Theorem 9.3

- (a)  $S_n(\hat{\xi}) = \mathbf{0}$ ;
- (b)  $\hat{\xi}$  is a consistent estimate of  $\xi_*$ ;
- (c) for each  $\xi$  in an open subset of Euclidean space,  $\partial_\xi \mathbf{h}(\xi, U)$  is twice continuously differentiable for every  $U$  and  $\|\partial_\xi^2 \mathbf{h}(\xi, U)\| \leq f(U)$  holds for every  $\xi$  in a neighborhood of  $\xi_*$ , where  $f(U)$  is fixed integrable function;
- (d)  $E[\mathbf{h}(\xi_*, U)] = \mathbf{0}$ ,  $E\|\partial_\xi \mathbf{h}(\xi_*, U)\|^2 < \infty$ , and the matrix  $E[\partial_\xi \mathbf{h}(\xi_*, U)]$  exists and is nonsingular.

# Z-Estimators

Then we have

$$\sqrt{n}(\hat{\xi} - \xi_*) = [-E(\partial_{\xi}\mathbf{h}(\xi_*, U))]^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbf{h}(\xi_*, U_i) + o_p(1). \quad (9.6)$$

In particular,  $\sqrt{n}(\hat{\xi} - \xi_*)$  converges to a  $N(0, \Sigma(\xi_*))$  distribution as  $n \rightarrow \infty$ , where

$$\Sigma(\xi_*) = [-E(\partial_{\xi}\mathbf{h}(\xi_*, U))]^{-1} [E(\mathbf{h}(\xi_*, U)^{\otimes 2})] [-E(\partial_{\xi}\mathbf{h}(\xi_*, U))]^{-1}.$$

- Consider testing the nonlinear hypotheses:

$$H_0 : h_0(\xi) = b_0 \quad \text{vs.} \quad H_1 : h_0(\xi) \neq b_0, \quad (9.7)$$

where  $h_0(\cdot)$  is an  $r \times 1$  vector function of the  $q$ -vector  $\xi$  with  $q \geq r$  and  $b_0$  is an  $r \times 1$  specified vector.

- Wald-type test: Possible
- Score test: Possible
- Likelihood ratio test: Impossible

# Z-Estimators

The **Wald-type test statistic** is given by

$$W_n = (h_0(\hat{\xi}) - b_0)^T \left[ \text{Cov}(h_0(\hat{\xi})) \right]^{-1} (h_0(\hat{\xi}) - b_0). \quad (9.8)$$

Using the delta method, it can be shown that

$$\text{Cov}(h_0(\hat{\xi})) = H(\hat{\xi}) \text{Cov}(\hat{\xi}) H(\hat{\xi})^T,$$

where  $H(\xi) = \partial h_0(\xi) / \partial \xi$  is an  $r \times q$  matrix. It can be shown that  $W_n$  is asymptotically distributed as  $\chi^2(r)$  under the null hypothesis  $H_0$ .

- For the score test, we only consider testing the linear hypothesis  $H_0 : R\xi = b_0$ .
- Let  $R = (\mathbf{I}_r, \mathbf{0})$ ,  $R\xi = \xi(1)$  and  $\xi^T = (\xi(1)^T, \xi(2)^T)^T$ , in which  $\xi(1)$  and  $\xi(2)$  are, respectively,  $r \times 1$  and  $(q - r) \times 1$  subvectors of  $\xi$ .

•

$$H_0 : \xi(1) = b_0 \quad \text{vs.} \quad H_1 : \xi(1) \neq b_0. \quad (9.9)$$

# Z-Estimators

- Let  $\mathbf{h}(\xi, U) = (\mathbf{h}_1(\xi, U)^T, \mathbf{h}_2(\xi, U)^T)^T$ .
- $\mathbf{h}_1(\xi, U)$  corresponds to  $\xi(1)$  and  $\mathbf{h}_2(\xi, U)$  is associated with  $\xi(2)$ .
- $\hat{\xi} = (\hat{\xi}(1)^T, \hat{\xi}(2)^T)^T$  solves  $S_n(\hat{\xi}) = \mathbf{0}$ .
- Under  $H_0$ ,  $\tilde{\xi} = (b_0^T, \tilde{\xi}(2)^T)^T$  solves  $S_{n,2}(\tilde{\xi}) = \mathbf{0}$ .
- $\hat{\xi}(2)(\xi(1))$  solves  $S_{n,2}(\xi(1), \xi(2)) = 0$  for each  $\xi(1)$ . Thus,  $\tilde{\xi}(2) = \hat{\xi}(2)(b_0)$ .
- $\dot{S}_n(\xi) = \partial_\xi S_n(\xi) = \begin{pmatrix} S_{n,11} & S_{n,12} \\ S_{n,21} & S_{n,22} \end{pmatrix}$  and  $(\partial_\xi S_n(\xi))^{-1} = \begin{pmatrix} S_n^{11} & S_n^{12} \\ S_n^{21} & S_n^{22} \end{pmatrix}$ .

# Z-Estimators

$$SC_n = S_{n,1}(\tilde{\xi})^T \Sigma^{11}(\tilde{\xi}) S_{n,1}(\tilde{\xi}), \quad (9.10)$$

where

$\Sigma^{11}(\tilde{\xi}) = \{(\mathbf{I}_r, -\tilde{S}_{n,12}\tilde{S}_{n,22}^{-1})\{\sum_{i=1}^n [\mathbf{h}(\tilde{\xi}, U_i) - \bar{\mathbf{h}}(\tilde{\xi})]^{\otimes 2}\}(\mathbf{I}_r, -\tilde{S}_{n,12}\tilde{S}_{n,22}^{-1})^T\}^{-1}$ , in which  $\bar{\mathbf{h}}(\tilde{\xi}) = \sum_{i=1}^n \mathbf{h}(\tilde{\xi}, U_i)/n$ ,  $\tilde{S}_{n,12} = S_{n,12}(\tilde{\xi})$  and  $\tilde{S}_{n,22} = S_{n,22}(\tilde{\xi})$ . An advantage of using the score test statistic is that it avoids the calculation of the estimator under the alternative hypothesis. For a proper  $\Sigma^{11}(\tilde{\xi})$ , the statistic  $SC_n$  is asymptotically distributed as  $\chi^2(r)$  under the null hypothesis  $H_0$ .

# Z-Estimators

- $\tilde{\xi}(2) - \xi(2)_* = -S_{n,22}(\xi_*)^{-1} S_{n,2}(\xi_*) + O_p(n^{-1})$ .
- $S_{n,1}(\tilde{\xi}) = S_{n,1}(\xi_*) + S_{n,12}(\xi_*)(\tilde{\xi}(2) - \xi(2)_*)[1 + o_p(1)] = (\mathbf{I}_r, -S_{n,12}S_{n,22}^{-1})S_n(\xi_*)[1 + o_p(1)]$ .
- $n^{-1}S_{n,12} \rightarrow S_{12}$  and  $n^{-1}S_{n,22} \rightarrow S_{22}$ .
- Using the CLT, we have  $n^{-1/2}S_n(\xi_*) \rightarrow N(0, \text{Var}[\mathbf{h}(\xi_*, U)])$ . Thus,  
$$n^{-1/2}(\mathbf{I}_r, -S_{n,12}S_{n,22}^{-1})S_n(\xi_*) \xrightarrow{L} N(\mathbf{0}, (\mathbf{I}_r, -S_{12}S_{22}^{-1})\text{Var}[\mathbf{h}(\xi_*, U)](\mathbf{I}_r, -S_{12}S_{22}^{-1})^T)$$
.

# Z-Estimators

- $\hat{\xi} = \hat{\beta} \rightarrow^p \xi_* = \beta_*$
- $\xi_*$  is the solution of  $S(\xi) = E\{\partial_\beta \mu(\mathbf{x}, \beta) V(\mu(\beta))^{-1} [y - \mu(\mathbf{x}, \beta)]\} = 0$ .
- $\sqrt{n}(\hat{\beta} - \beta_*) \rightarrow^L N(0, \Sigma(\beta_*))$ .
- $\mathbf{h}(\beta, U) = \partial_\beta \mu(\mathbf{x}, \beta) V(\mu(\beta))^{-1} [y - \mu(\mathbf{x}, \beta)]$
- $E[\mathbf{h}(\beta, U)^{\otimes 2}] = \sigma^2 [\partial_\beta \mu(\mathbf{x}, \beta) V(\mu(\beta))^{-1} \partial_\beta \mu(\mathbf{x}, \beta)^T]$
- $E[\partial_\beta \mathbf{h}(\beta, U)] = \partial_\beta \mu(\mathbf{x}, \beta) V(\mu(\beta))^{-1} \partial_\beta \mu(\mathbf{x}, \beta)^T$

# Generalized Estimating Equations

- **Correlated data** arise from many studies, such as longitudinal studies on brain development, quality of life studies and measures of immune response in cancer studies, crossover studies about drug comparisons, and multivariate measures in a psychometric study.
- **Longitudinal designs** can be very powerful, because they enable one to study changes within individuals over time or under a variety of differing conditions.
- **An important consideration** in the analysis of correlated data is to account for the correlated measurements.
- Ignoring correlations can overestimate the standard error for the between-subject effects, which leads to inefficient estimation.

# Generalized Estimating Equations

- For continuous outcomes, it is often appropriate to use the **linear mixed model for normally distributed errors**, which requires the assumption of a covariance matrix.
- When the data have **time-dependent covariates, missing data, or non-normality**, then the linear mixed model can be inappropriate.
- GEEs are particularly ideal for discrete response data, such as binary and polytomous data.

# Generalized Estimating Equations

- The key idea of GEEs is to use the **same link functions and linear predictor set-up as for GLMs in the independence case.**
- The difference between the GLM and GEE methods is that the **GEE method accounts for the correlation structure of the covariance matrix of the responses in the estimation process.**
- GEE accounts for the misspecification of the correlation structure.
- For the  $j$ -th time point from the  $i$ -th subject, we observe a scalar response  $y_{ij}$ , and a  $p \times 1$  vector of covariates of interest, denoted by  $\mathbf{x}_{ij} = (x_{ij1}, \dots, x_{ijp})^T$ , for  $i = 1, \dots, n$  and  $j = 1, \dots, T_i$ .

# Generalized Estimating Equations

## Example 9.4

We consider a respiratory dataset (Stokes, Davis, and Koch, 1995). Patients in each of two centers were randomly assigned to groups receiving the active treatment or a placebo. Respiratory status (0=poor, 1=good) was recorded for each of four visits for each patient. Covariates of interest include center, treatment, gender, age, and baseline respiratory status. We are interested in comparing two treatments for the respiratory disorder. Due to the binary response, we can use the variance function for the binomial distribution

$V(\mu_{ij}) = \mu_{ij}(1 - \mu_{ij})$  and the logit link function

$g(\mu_{ij}) = \log(\mu_{ij}/(1 - \mu_{ij})) = \mathbf{x}_{ij}^T \boldsymbol{\beta}$ .

# Generalized Estimating Equations

- $y_{ij} \sim D(\theta_{ij}, \phi)$ . Thus,  $E(y_{ij}) = \dot{b}(\theta_{ij}) = \mu_{ij}$  and  $\text{Var}(y_{ij}) = \phi^{-1} \ddot{b}(\theta_{ij})$ .
- $g(\mu_{ij}) = \mathbf{x}_{ij}^T \beta$ .
- GEEs assume a **working covariance matrix** given by  $\phi^{-1} V_i = \phi^{-1} B_i^{1/2} R_i(\alpha) B_i^{1/2}$ , where  $B_i = \text{diag}(\ddot{b}(\theta_{i1}), \dots, \ddot{b}(\theta_{iT_i}))$  and  $R_i(\alpha)$  is a working correlation matrix for  $\mathbf{y}_i$ .
- $G_n(\alpha, \beta) = \sum_{i=1}^n D_i^T V_i^{-1} [\mathbf{y}_i - \mu_i(\beta)] = \mathbf{0}$ , where  $D_i = \partial \mu_i / \partial \beta$ .
- The **GEE estimator**  $\hat{\beta}$  is the solution of these GEEs.

# Generalized Estimating Equations

- The **working correlation matrix**  $R_i(\alpha)$  is fully specified by a vector of parameters  $\alpha$ .
- The  $\alpha$  is usually estimated using the Pearson residual  
$$\hat{e}_{ij} = (y_{ij} - \hat{\mu}_{ij}) / \sqrt{\hat{\phi}^{-1} \ddot{b}(\hat{\theta}_{ij})}.$$
- **Independent GEE:**  $R(\alpha)$  is an identity matrix.
- **Exchangeable:**  $\text{Corr}(y_{ij}, y_{ik}) = \rho$  for  $j \neq k$
- **Autoregressive:**  $\text{Corr}(y_{ij}, y_{ik}) = \rho^{|j-k|}$  for  $j \neq k$
- **$m$ -dependent:**  $\text{Corr}(y_{ij}, y_{ik}) = \rho_{|j-k|}$  for  $|j - k| \leq m$
- **Unstructured:**  $\text{Corr}(y_{ij}, y_{ik}) = \rho_{jk}$  for  $j \neq k$ .

# Generalized Estimating Equations

- Compute an initial estimate  $\beta^{(0)}$  based on an independent working correlation matrix.
- Compute the working correlation matrix  $R_i(\alpha)$  based on the Pearson residuals and the current  $\beta^{(r)}$ .
- Compute an estimate of the covariance  $V_i = \phi B_i^{1/2} R_i(\hat{\alpha}) B_i^{1/2}$ .
- Update  $\beta$  according to

$$\beta^{(r+1)} = \beta^{(r)} + \left[ \sum_{i=1}^n D_i^T V_i^{-1} D_i \right]^{-1} \sum_{i=1}^n D_i^T V_i^{-1} (\mathbf{y}_i - \mu_i) \Bigg|_{\beta^{(r)}}.$$

- Repeat the previous steps until convergence.

# Generalized Estimating Equations

- $\hat{\beta} \rightarrow \beta_*$
- $\sqrt{n}(\hat{\beta} - \beta_*) \rightarrow N(0, \Sigma)$ .
- $\hat{\Sigma} = K_0^{-1}[n^{-1} \sum_{i=1}^n D_i^T V_i^{-1} \text{Cov}(Y_i) V_i^{-1} D_i] K_0^{-1}$ , in which  $K_0 = n^{-1} \sum_{i=1}^n D_i^T V_i^{-1} D_i$  and  $\text{Cov}(Y_i)$  can be replaced by  $(\mathbf{y}_i - \mu_i(\hat{\beta}))^{\otimes 2}$ .

# Generalized Estimating Equations

- The parameters  $\alpha$  and  $\phi$  can be estimated based on an assumed working correlation form.
- For the **exchangeable working correlation matrix**, we have

$$\hat{\alpha} = \frac{1}{(N^* - p)\phi} \sum_{i=1}^n \sum_{j < k} \hat{e}_{ij} \hat{e}_{ik},$$

where  $N^* = 0.5 \sum_{i=1}^n T_i(T_i - 1)$ .

- $\hat{\phi} = \frac{1}{(N-p)} \sum_{i=1}^n \sum_{j=1}^{T_i} \hat{e}_{ij}^2$ , where  $N = \sum_{i=1}^n T_i$  is the total number of measurements.

# Generalized Estimating Equations

- Another way of estimating  $\alpha$  in the working correlation matrix is to simultaneously solve  $G_n(\xi) = \mathbf{0}$  and

$$\sum_{i=1}^n \left( \frac{\partial \eta_i}{\partial \alpha} \right)^T H_i^{-1} (W_i - \eta_i) = \mathbf{0}, \quad (9.11)$$

where  $\xi = (\alpha, \beta)$ ,  $W_i = (r_{i1} r_{i2}, r_{i1} r_{i3}, \dots, r_{iT_i-1} r_{iT_i}, r_{i1}^2, \dots, r_{iT_i}^2)^T$  and  $\eta_i = E(W_i; \alpha, \beta)$ , in which  $r_{ij} = (y_{ij} - \mu_{ij}) / \sqrt{\phi^{-1} \ddot{b}(\theta_{ij})}$  (Prentice, 1988).

- If  $H_i$  is set as  $\text{Cov}(W_i)$ , then this  $H_i$  is optimal. However,  $H_i$  involves assuming higher moments of the  $y_{ij}$ s.

# Generalized Estimating Equations

GEE2 assumes

$$\sum_{i=1}^n \tilde{D}_i^T \tilde{V}_i^{-1} \begin{pmatrix} \mathbf{y}_i - \mu_i(\beta) \\ \mathbf{s}_i - \sigma_i \end{pmatrix} = \mathbf{0},$$

where  $\sigma_i = E(\mathbf{s}_i; \alpha, \beta)$ ,  $\mathbf{s}_i = (e_{i1} e_{i2}, e_{i1} e_{i3}, \dots, e_{iT_i-1} e_{iT_i}, e_{i1}^2, \dots, e_{iT_i}^2)$ ,

$$\tilde{D}_i = \begin{pmatrix} \partial_\beta \mu_i & \mathbf{0} \\ \partial_\beta \sigma_i & \partial_\alpha \sigma_i \end{pmatrix} \quad \text{and} \quad \tilde{V}_i = \begin{pmatrix} \text{Var}(\mathbf{y}_i) & \text{Cov}(\mathbf{y}_i, \mathbf{s}_i) \\ \text{Cov}(\mathbf{s}_i, \mathbf{y}_i) & \text{Cov}(\mathbf{s}_i) \end{pmatrix}.$$

This estimating equation for estimating both  $\alpha$  and  $\beta$  is the score equation under a **quadratic exponential model**. However, **GEE2 is unable to model the Gaussian-type AR(1), MA(1), and exchangeable correlation structures appropriately**.

# Generalized Estimating Equations

- For longitudinal data, the GEE estimator  $\hat{\beta}$  can be as efficient as the maximum likelihood estimator.
- The GEE estimator  $\hat{\beta}$  is robust in the sense that it is a consistent estimator even if the working correlation matrix is misspecified.
- Valid standard errors of  $\hat{\beta}$  can be obtained using the sandwich estimator of  $\text{Cov}(\hat{\beta})$ .
- The GEE method can be applied to continuous responses, because GEE just requires specifying the mean and covariance among the continuous responses.

# Generalized Estimating Equations

- Why do we care about correctly modeling the within subject association?
- Correctly modeling the within subject association can improve the efficiency or precision with which  $\beta$  can be estimated.
- The 'sandwich' estimator of  $\text{Cov}(\hat{\beta})$  is a large sample (or asymptotic) estimator. Such an estimator is quite reliable when the number of subjects  $n$  is large and the number of longitudinal measures is relatively small in balanced longitudinal designs.
- For GEE, the Wald test is too liberal, whereas the score test is too conservative for small samples.

- Data may come from a true complicated process.
- Finding a 'right'/'fitted' model to interpret a dataset and to approximate the true complicated process.
- Fitted Model  $\neq$  True Process
- Discrepancy = Fitted Model  $\ominus$  True Process
- How do we use statistical tools to detect such discrepancies?

# What is a Discrepancy?

Detect two types of discrepancies:

- Discrepancy exists between isolated observations (e.g., influential points and outliers) and the rest of the observations
  - residuals
  - leverages
  - case-deletion measures
  - local influence measures
- Any systematic discrepancies between the data and the fitted values obtained from statistical models
  - graphical procedures of residuals, such as partial residual and added variable plots
  - goodness-of-fit test statistics and test procedures for testing specific alternatives

# Diagnostic Measures

Consider  $\mathbf{y} = X\beta + \epsilon$ , where  $\epsilon = (\epsilon_1, \dots, \epsilon_n)^T \sim N(\mathbf{0}, \sigma^2 I_n)$ .

- The quantity  $H = X(X^T X)^{-1}X^T$  is called the hat matrix, where  $H = (h_{ij})$ .
- The raw residuals,  $\hat{\mathbf{e}} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - H\mathbf{y} = Q\mathbf{y}$ ,  $Q = I_n - H$ , provide important information about the fitted model, such as model misspecification, outliers, and influential points.
- $\hat{e}_i = y_i - \hat{y}_i = (1 - h_{ii})\epsilon_i - \sum_{j \neq i} h_{ij}\epsilon_j$ .
- $\hat{\mathbf{e}}$  is an approximation of the random error vector  $\epsilon$  and can be used to measure the **goodness-of-fit of the fitted model to the data**.

# Diagnostic Measures

- The  $\hat{e}_i$  can be used to check a **particular data point**.
- Because  $\text{Cov}(\hat{\mathbf{e}}) = \sigma^2(I_n - H)$ , all components of  $\hat{\mathbf{e}}$  may have different variances and are correlated with each other.
- $r_i = \frac{\hat{e}_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$ , where  $\hat{\sigma}$  is an estimate of  $\sigma$ .
- Any observation with  $|r_i|$  greater than 2.0 should be checked further.

# Diagnostic Measures

- The diagonal elements  $h_{ii}$  of the hat matrix  $H$  are called **leverages** and can be used for **assessing each  $\mathbf{x}_i$** .
- Let  $\mathbf{x}_i = (1, x_{i2}, \dots, x_{ip})^T$  and  $\tilde{X} = [X_2, \dots, X_p]$ . We have  $H = n^{-1}\mathbf{1}\mathbf{1}^T + X_c(X_c^T X_c)^{-1}X_c^T$ , where  $X_c = (I_n - n^{-1}\mathbf{1}\mathbf{1}^T)\tilde{X}$  is the centered  $\tilde{X}$ , and the  $(i,j)$ th component of  $\tilde{X}$ , denoted by  $x_{cij}$ , is given by  $x_{cij} = x_{ij} - \bar{x}_{\cdot j}$ , where  $\bar{x}_{\cdot j} = \sum_{i=1}^n x_{ij}/n$ .
- $h_{ii} = \mathbf{x}_i^T (X^T X)^{-1} \mathbf{x}_i = n^{-1} + \mathbf{x}_{ci}^T (X_c^T X_c)^{-1} \mathbf{x}_{ci}$ .

# Diagnostic Measures

- $h_{ii} \in (0, 1)$
- If  $h_{ii} = 1$ , then  $h_{ij} = 0$  for all  $j \neq i$ .
- $h_{ii} = \sum_{j=1}^n h_{ij}^2 = h_{ii}^2 + \sum_{j \neq i} h_{ij}^2$  and  $h_{ii}(1 - h_{ii}) \geq 0$ .
- $\sum_{i=1}^n h_{ii} = p$  and  $\max_i h_{ii} \geq p/n$
- $1/a_i \geq h_{ii} \geq n^{-1}$ , because  $h_{ii} = \sum_{j=1}^n h_{ij}^2 \geq a_i h_{ii}^2$ , where  $a_i = \#\{j : \mathbf{x}_j = \mathbf{x}_i\}$
- Any  $\mathbf{x}_i$  having  $h_{ii} > 2p/n$  should be looked at more closely. Large values of  $h_{ii}$  correspond to extreme points in the covariate space.

# Diagnostic Measures

## Cook's distance (Cook, 1977)

- Cook's distance measures the distance between  $\hat{\beta}$  and the estimate of  $\beta$  without the  $i$ -th observation, denoted by  $\hat{\beta}_{(i)}$ .
- A **deletion model with  $(y_i, \mathbf{x}_i)$  deleted** can be defined as  
 $y_j = \mathbf{x}_j^T \beta + \epsilon_j, \text{ for } j = 1, \dots, i-1, i+1, \dots, n.$
- $\hat{\beta}_{(i)} = \hat{\beta} - \frac{(X^T X)^{-1} \mathbf{x}_i \hat{\epsilon}_i}{1-h_{ii}}$  and  $\hat{\sigma}_{(i)}^2 = \frac{n-p-r_i^2}{n-p-1} \hat{\sigma}^2$ .
- $C_i = \frac{(\hat{\beta} - \hat{\beta}_{(i)})^T X^T X (\hat{\beta} - \hat{\beta}_{(i)})}{p \hat{\sigma}^2} = \frac{1}{p} r_i^2 \frac{h_{ii}}{1-h_{ii}}.$
- The **larger the difference** between  $\hat{\beta}$  and  $\hat{\beta}_{(i)}$ , the **more influential** the  $i$ -th data point is.

# Diagnostic Measures

- The **covariance ratio statistic** measures the influence of  $\mathbf{x}_i$  using the determinant of the covariance matrix of  $\hat{\beta}$ .
- $CR_i = \frac{|\text{cov}(\hat{\beta}_{[i]})|}{|\text{cov}(\hat{\beta})|} = \frac{|X^T X|}{|X_{(i)}^T X_{(i)}|}$ , where  $X_{(i)}$ , an  $(n - 1) \times p$  matrix, is obtained from  $X$  with the  $i$ -th row deleted.
- $CR_i = (1 - h_{ii})^{-1}$  is greater than 1.
- The larger the  $CR_i$ , the more influential the  $\mathbf{x}_i$ .

# Diagnostic Measures

- The AP statistic measures the influence of  $(y_i, \mathbf{x}_i)$
- $AP_i = \frac{|X^{*T} X^*|}{|X_{(i)}^{*T} X_{(i)}^*|} = (1 - h_{ii}^*)^{-1} = \left(1 - h_{ii} - \frac{\hat{e}_i^2}{\hat{\mathbf{e}}^T \hat{\mathbf{e}}}\right)^{-1}$ , where  $h_{ii}^*$  is the  $(i, i)$ th diagonal element of the hat matrix  $H^*$  of  $X^* = (X, \mathbf{y})$  and  $X_{(i)}^*$  is obtained from  $X^*$  with the  $i$ th row deleted.
- The larger the  $AP_i$ , the more influential the  $(y_i, \mathbf{x}_i)$ .

# Diagnostic Measures

## Example 10.1

An important question is whether the proposed linear regression model provides an adequate fit to the steam data. We examine the diagnostic measures based on the fitted linear model to the steam data. Residuals reveal that the 6th point is influential, whereas Cook's distance indicates that the 6th and 10th points are influential. The index plot of the leverages identifies the 4th and 10th points as leverage points. We refitted the linear regression model by removing only the 6th point and removing both the 4th and 10th points. We observed that removing both 4th and 10th points strongly influences the estimates and the general fit of the model.

# Diagnostic Measures

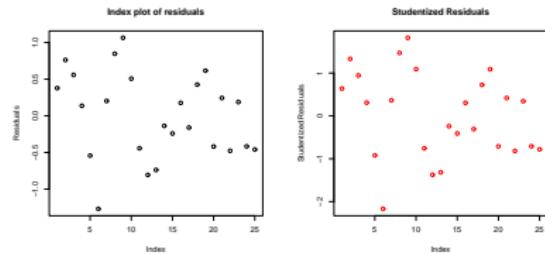
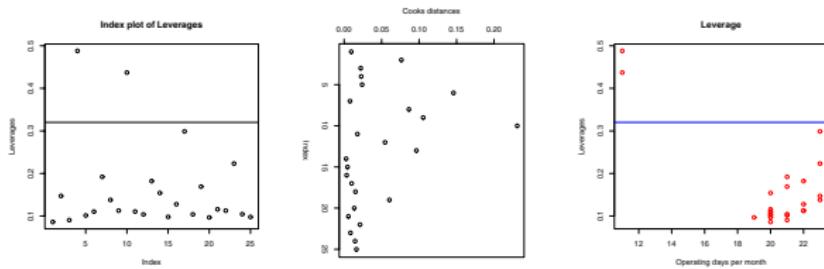


Figure: Index plots of residuals and studentized residuals for the Steam data.

# Diagnostic Measures



**Figure:** Index plots of leverages (top) and Cook's distance (middle), and the plot of operating days per month against leverage (bottom) based on the linear regression model for the Steam data.

# Diagnostic Measures

- Generalized linear models are closely related with the classical linear model.
- Consider two models as follows:

$$(A) \quad y_i \sim D(\theta_i, \phi) \text{ and } g(\mu_i) = \mathbf{x}_i^T \beta, \text{ for } i = 1, \dots, n;$$

$$(B) \quad \hat{V}^{-\frac{1}{2}} \hat{Z} = \hat{V}^{-\frac{1}{2}} \hat{D} \beta + \epsilon, \quad \epsilon \sim (\mathbf{0}, \phi I_n),$$

where  $\hat{D} = D(\hat{\beta})$ ,  $\hat{\mathbf{e}} = \mathbf{y} - \mu(\hat{\beta})$ , and  $\hat{Z} = \hat{D} \hat{\beta} + \hat{\mathbf{e}}$ .

- Recall that  $V(\hat{\beta}) = \text{diag}(v_1(\hat{\beta}), \dots, v_n(\hat{\beta}))$ ,  $\mathbf{e} = \mathbf{y} - \mu(\beta)$ , and  $D(\beta) = \partial \mu(\beta) / \partial \beta$ .

# Diagnostic Measures

Two common features exist between generalized linear models (A) and linear models (B):

- The  $i$ -th case of both models (A) and (B) corresponds to the observation  $(\mathbf{x}_i, y_i)$  (note that  $V$  is a diagonal matrix).
- The maximum likelihood estimator of  $\beta$  in model (A) equals the least squares estimator of  $\beta$  in model (B).
- We use diagnostic measures based on model (B) as diagnostic measures for model (A).

# Diagnostic Measures

- (i) Pearson residuals  $\mathbf{r}_p = (\hat{V}^{-\frac{1}{2}}Z - \hat{V}^{-\frac{1}{2}}\hat{D}\beta)_{\hat{\beta}} = \hat{V}^{-\frac{1}{2}}\hat{\mathbf{e}}$  and studentized residuals  $r_i = \frac{r_{pi}}{\hat{\sigma}\sqrt{1-h_{ii}}}$ , where  $r_{pi}$  is the  $i$ th component of  $\mathbf{r}_p$  and  $h_{ii}$  is an element of  $H$  defined below.
- (ii) Hat matrix  $H = (h_{ij}) = \{V^{-\frac{1}{2}}D(D^T V^{-1}D)^{-1}D^T V^{-\frac{1}{2}}\}_{\hat{\beta}}$ . Since  $(D^T V^{-1}\mathbf{e})_{\hat{\beta}} = \mathbf{0}$ , we have  $H\mathbf{r}_p = \mathbf{0}$  and  $(I_n - H)\mathbf{r}_p = \mathbf{r}_p$ .
- (iii) Cook's distance is
- $$C_i = \frac{(\hat{\beta} - \hat{\beta}_{(i)})^T (\hat{D}^T \hat{V}^{-1} \hat{D})(\hat{\beta} - \hat{\beta}_{(i)})}{p\hat{\sigma}^2} \approx C_i^I = \frac{h_{ii}}{1-h_{ii}} \frac{r_i^2}{p}.$$

## Theorem 10.1

Let  $h_{ii}^*$  be the diagonal element of the projection matrix  $H^*$  obtained from  $X^* = (\hat{V}^{-\frac{1}{2}}\hat{D}, \hat{V}^{-\frac{1}{2}}\hat{Z})$ . Then we have

$$AP_i = \left\{ \frac{|X_{(i)}^{*T} X_{(i)}^*|}{|X^{*T} X^*|} \right\}_{\hat{\beta}} = (1 - h_{ii}^*)^{-1} = \left( 1 - h_{ii} - r_{pi}^2 / (\mathbf{r}_p^T \mathbf{r}_p) \right)^{-1}, \quad (10.1)$$

where  $X_{(i)}^*$  is obtained from  $X^*$  with the  $i$ -th row deleted.

# Diagnostic Measures

## Example 10.2

Consider the logistic regression model where  $y_i \sim B(m_i, p_i)$  and  $\text{logit}(p_i) = \mathbf{x}_i^T \boldsymbol{\beta}$  for  $i = 1, \dots, n$ . Recall the results:  $\mu_i = m_i p_i$ ,  $v_i = m_i p_i (1 - p_i)$ , and  $\partial_{\boldsymbol{\beta}} \mu_i = m_i p_i (1 - p_i) \mathbf{x}_i$ . Therefore,  $V = \text{diag}(m_1 p_1 (1 - p_1), \dots, m_n p_n (1 - p_n))$ ,  $\mathbf{e} = \mathbf{y} - \boldsymbol{\mu}$ , and  $D = V\mathbf{X}$ . Thus,  $\mathbf{r}_p = \hat{V}^{-1/2}(\mathbf{y} - \hat{\boldsymbol{\mu}})$  and its elements are the Pearson residuals given by  $r_{pi} = \frac{y_i - m_i \hat{p}_i}{\sqrt{m_i \hat{p}_i (1 - \hat{p}_i)}}$ . The hat matrix  $H = (h_{ii})$  equals  $\hat{V}^{1/2} \mathbf{X} (\mathbf{X}^T \hat{V} \mathbf{X})^{-1} \mathbf{X}^T \hat{V}^{1/2}$ . For the  $i$ th observation with leverage  $h_{ii}$ , the standardized residual is

$$r_i = \frac{y_i - m_i \hat{p}_i}{\sqrt{m_i \hat{p}_i (1 - \hat{p}_i)(1 - h_{ii})}}.$$

# Diagnostic Measures

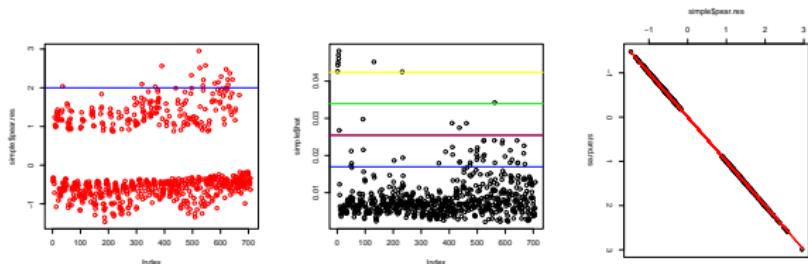


Figure: Index plots of Pearson residuals and leverages, and plot of Pearson residuals against studentized residuals.

# Diagnostic Measures

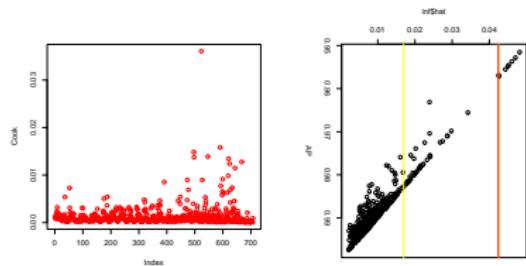


Figure: Index plots of  $C_i^I$  and plot of leverages against  $AP_i$ .

## Example 10.3

*In Example 5.9, we used the logistic regression model to fit the vegetation data in Example 3.4 in order to model the relationship between the distribution of V2 and the climate variables  $X_1, \dots, X_5$ . Here, we examined the diagnostic measures to see whether the proposed logistic regression model provides an adequate fit to the vegetation data.*

# Diagnostic Measures

- A fundamental approach of influence diagnostics is to compare  $\hat{\xi} = (\hat{\beta}, \hat{\phi})$  with  $\hat{\xi}_{(i)} = (\hat{\beta}_{(i)}, \hat{\phi}_{(i)})$  under the *i-th case deletion model* defined as  $y_j \sim D(\theta_j, \phi)$  and  $g(\mu_j) = g(b(\theta_j)) = \mathbf{x}_j^T \beta$ , for  $j = 1, \dots, n$ ,  $j \neq i$ .
- Run the Newton-Raphson algorithm with the *i-th* observation deleted for each  $i = 1, \dots, n$ . This could be **computationally intensive** for large  $n$ .
- Use the first order approximations  $(\hat{\beta}_{(i)}^I, \hat{\phi}_{(i)}^I)$  of  $(\hat{\beta}_{(i)}, \hat{\phi}_{(i)})$ . Such approximations may be **sufficient for diagnostic purposes**.

## Lemma 10.1

The *first order approximation*  $\hat{\beta}_{(i)}$  of  $\beta$  in a Case Deletion Measure (CDM) can be expressed as

$$\hat{\beta}'_{(i)} = \hat{\beta} - \left\{ \frac{(D^T V^{-1} D)^{-1} v_i^{-\frac{1}{2}} d_i r_{pi}}{1 - h_{ii}} \right\}_{\hat{\beta}}, \quad (10.2)$$

where  $d_i^T$  is the  $i$ -th row of  $D(\beta)$ .

# Diagnostic Measures

- To find the influential points, we can compute a certain “**distance**” between  $\hat{\xi}$  and  $\hat{\xi}_{(i)}$ .
- Construct a **generalized Cook's distance** for  $\beta$ :  
$$C_i \stackrel{\Delta}{=} \|\hat{\beta} - \hat{\beta}_{(i)}\|_M^2 = (\hat{\beta} - \hat{\beta}_{(i)})^T M (\hat{\beta} - \hat{\beta}_{(i)}).$$
- $C_i = \frac{(\hat{\beta} - \hat{\beta}_{(i)})^T (D^T V^{-1} D)_{\hat{\beta}} (\hat{\beta} - \hat{\beta}_{(i)})}{p \hat{\sigma}^2}.$
- $C_i^I = \frac{h_{ii}}{1-h_{ii}} \frac{r_i^2}{p}.$

## Likelihood Displacement (Distance)

- $LD_i(\xi) = 2\{\ell_n(\hat{\xi}) - \ell_n(\hat{\xi}_{(i)})\}.$
- $LD_i = LD_i(\hat{\xi}) \approx (\hat{\xi} - \hat{\xi}_{(i)})^T \{-\partial_{\xi}^2 \ell_n(\hat{\xi})\} (\hat{\xi} - \hat{\xi}_{(i)}).$
- $LD_i(\xi_1 | \xi_2) = 2\{\ell_n(\hat{\xi}_1, \hat{\xi}_2) - \ell_n(\hat{\xi}_{1(i)}, \tilde{\xi}_2(\hat{\xi}_{1(i)}))\},$  where  $\hat{\xi}_{(i)} = (\hat{\xi}_{1(i)}^T, \hat{\xi}_{2(i)}^T)^T$  and  $\tilde{\xi}_2(\hat{\xi}_1)$  is the maximum likelihood estimator of  $\xi_2$  for fixed  $\xi_1.$
- $LD_i(\xi_1 | \xi_2) \approx (\hat{\xi}_1 - \hat{\xi}_{1(i)})^T \{L_{11} - L_{12}L_{22}^{-1}L_{21}\} (\hat{\xi}_1 - \hat{\xi}_{1(i)}),$  where  $L_{ij}, (i = 1, 2; j = 1, 2)$  are four partitions of  $-\partial_{\xi}^2 \ell_n(\hat{\xi})$  (or  $E\{-\partial_{\xi}^2 \ell_n(\hat{\xi})\}$ ) corresponding to  $(\xi_1, \xi_2).$

# Diagnostic Measures

- (i) **Deviance residuals:**  $rD_i = \{\text{sign}(y_i - \hat{\mu}_i)\} Dv_i(y_i, \mu_i(\hat{\beta}))$ .
- (ii) **Difference of deviances:**  $\triangle_i Dv = Dv(\hat{\beta}) - Dv_{(i)}(\hat{\beta}_{(i)})$ , where  $Dv$  is the deviance for all  $n$  observations and  $Dv_{(i)}$  is the deviance for all  $n$  observations except the  $i$ th observation.
- (iii) For classical linear regression,  $Dv(\hat{\beta}) = RSS$  (the residual sum of squares), and thus

$$\begin{aligned}\triangle_i Dv &= RSS - RSS_{(i)} = n\hat{\sigma}^2 - (n-1)\hat{\sigma}_{(i)}^2 \\ &= (n-1)(\hat{\sigma}^2 - \hat{\sigma}_{(i)}^2) + \hat{\sigma}^2.\end{aligned}$$

## Theorem 10.2

*For generalized linear models, the first order approximation of  $\Delta_i Dv$  is given by*

$$\Delta_i Dv^I = Dv_i(y_i, \hat{\mu}_i) + h_{ii} r_i^2. \quad (10.3)$$

## Example 10.4

In logistic regression, since

$Dv_i = 2\{y_i \log(y_i/\hat{p}_i) + (1 - y_i) \log[(1 - y_i)/(1 - \hat{p}_i)]\}$ , we can calculate  $rD_i = \{\text{sign}(y_i - \hat{p}_i)\}Dv_i$  and  $\Delta_i Dv^I$  for  $i = 1, \dots, n$ .

Following example 10.3, we calculated  $rD_i$  and  $\Delta_i Dv^I$  for the BC data (Fig. 23).

# Diagnostic Measures

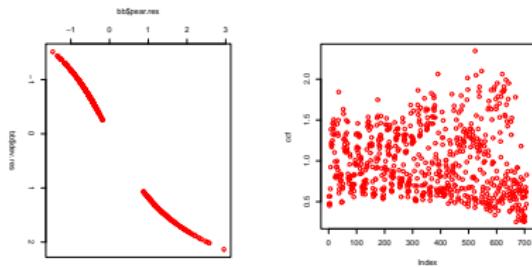


Figure: Top: Pearson residuals and deviance residuals; Bottom: Index plots of difference of deviance.

# Generalized Leverage

Leverage can be used for assessing the importance of individual covariates of interest  $\mathbf{x}_i$ . For the linear regression model, we have

- $Var(\hat{y}_i) = \sigma^2 h_{ii}$ ;
- $Var(\hat{e}_i) = \sigma^2(1 - h_{ii})$ , where  $\hat{e}_i = y_i - \mathbf{x}_i^T \hat{\beta}$ ;
- $\rho(y_i, \hat{y}_i)^2 = h_{ii}$ , the squared correlation coefficient of  $y_i$  and  $\hat{y}_i$ ;
- $h_{ij} = \partial \hat{y}_i / \partial y_j$ , where  $\hat{y}_i = \mathbf{x}_i^T \hat{\beta} = \sum_{j=1}^n h_{ij} y_j$ ;
- $h_{ii} = \partial \hat{y}_i / \partial y_i$  reflects directly the influence of the observation  $y_i$  on the fit
- These properties no longer hold for  $H = V^{-0.5} D (D^T V^{-1} D)^{-1} D^T V^{-0.5} = (h_{ij})$  in most generalized linear models. We need to find a more adequate definition of leverage for generalized linear models.

# Generalized Leverage

## Definition 10.1

Let  $Y = (y_1, \dots, y_n)^T$  be an  $n$ -vector of responses with probability density function  $p(y; \xi)$  and  $E(Y) = \mu = \mu(\xi)$ , where  $\xi$  is an unknown parameter. An estimator of  $\xi$  is denoted by  $\tilde{\xi} = \tilde{\xi}(Y)$  and  $\tilde{Y} = \mu(\tilde{\xi})$  may be regarded as the fitted values. Then

$$GL(\tilde{\xi}) = \partial \tilde{Y} / \partial Y^T = (\partial \tilde{y}_i / \partial y_j) \quad (10.4)$$

is defined as the **generalized leverage** of  $\tilde{\xi}$ .

## Lemma 10.2

Let  $\ell_n(\xi; y)$  be the log-likelihood of  $Y$  and  $\hat{\xi} = \hat{\xi}(Y)$  be the unique maximum likelihood estimator of  $\xi$ . If  $\ell_n(\xi; y)$  has second order continuous derivatives with respect to  $\xi$  and  $y$ , then we have

$$GL(\hat{\xi}) = \{(D_\xi)[-\partial_\xi^2 \ell_n(\xi)]^{-1} \partial_{\xi Y}^2 \ell_n(\xi)\}_{\hat{\xi}}, \quad (10.5)$$

where  $D_\xi = \partial_\xi \mu(\xi)$  and  $\partial_{\xi Y}^2 \ell_n(\xi) = \partial^2 \ell_n(\xi) / \partial \xi \partial Y^T$ .

# Generalized Leverage

## Theorem 10.3

*For generalized linear models, the generalized leverage of the maximum likelihood estimator  $\hat{\xi} = (\hat{\beta}^T, \hat{\phi})^T$  is given by*

$$GL(\hat{\xi}) = D(D^T V^{-1} D + \sum_{i=1}^n \hat{e}_i \ddot{\theta}_i)^{-1} D^T V^{-1}. \quad (10.6)$$

# Generalized Leverage

- The **generalized leverage** is closely connected with the “**effective residual curvature matrix**”  $\sum_{i=1}^n \hat{e}_i \ddot{\theta}_i$ .
- The leverage of  $\hat{\beta}$  with known  $\sigma^2$  is given by  
$$GL(\hat{\xi}) = GL(\hat{\beta}) = \left\{ \frac{\partial \mu}{\partial \beta^T} [-\partial_{\beta\beta}^2 \ell_n(\beta)]^{-1} \partial_{\beta Y}^2 \ell_n(\beta) \right\}_{\beta=\hat{\beta}},$$
- $GL(\hat{\xi}) \approx D(D^T V^{-1} D)^{-1} D^T V^{-1}$  and the **tangent plane leverage**  $H = V^{-\frac{1}{2}} D(D^T V^{-1} D)^{-1} D^T V^{-\frac{1}{2}}$  have the same diagonal terms.

# Generalized Leverage

## Theorem 10.4

Consider  $\tilde{\xi}(Y)$  as the minimizer of  $Q(\xi; Y)$ , which is defined by

$$Q(\xi; Y) = \sum_{i=1}^n \rho_i(f(\mathbf{x}_i, \beta); \phi),$$

where  $\rho_i(\cdot, \cdot)$  is a known function,  $\xi = (\beta, \phi)$ ,  $E(y_i) = \mu_i = f(\mathbf{x}_i, \beta)$ , and  $\phi$  may denote the nuisance parameter. Suppose that  $Q(\cdot, \cdot)$  has second-order continuous derivatives with respect to  $\xi$  and  $\mathbf{y}$  and  $\tilde{\xi}(Y)$  is unique. Then

$$GL(\tilde{\xi}) = \{(D_\xi)(-\partial_\xi^2 Q(\xi))^{-1}(\partial_{\xi Y}^2 Q(\xi))\}_{\tilde{\xi}}. \quad (10.7)$$

- It is true that any statistical model for a given dataset is wrong, although some models are useful.
- How do we interpret results from a misspecified model?
- Can we make statistical inferences based on the misspecified model?
- Are there any methods that can detect model misspecification?

## Assumptions for GLMs

- Independence Assumption
- Distributional Assumption
- Structural Assumption
- Homoscedasticity/Overdispersion
- We should be aware regarding the possible violation of these four key assumptions of generalized linear models.
- We may ask whether we can carry out statistical inference, such as estimation and hypothesis testing under misspecified models.

## Example 10.5

Suppose that we generate a dataset with 100 observations  $y_i$ , in which  $y_i \sim \chi^2(4)$ . Assume that we fit a simple linear model  $y_i = \beta + \epsilon_i$ , where  $\epsilon_i \sim N(0, \sigma^2)$ . In this case, the fitted model  $p(y, \xi)$  and the true model  $g(y)$  are, respectively, given by  $\log p(y, \xi) = -0.5 \log \sigma^2 - (y - \beta)^2 / 2\sigma^2$  and  $\log g(y) = -0.5y + \log y + \text{const}$ . The estimate  $\hat{\xi} = (\sum_{i=1}^{100} y_i / 100, \sum_{i=1}^{100} (y_i - \bar{y})^2 / 100)^T$  maximize  $\ell_n(\xi)$ . The covariance matrix of  $\sqrt{n}\hat{\xi}$  can be approximated by  $\text{diag}(\hat{\sigma}^2, 2\hat{\sigma}^4) = \text{diag}(8, 128)$ .

# Goodness-of-fit Statistics

- Does  $\hat{\xi}$  converges to a pseudo true value  $\xi_*$ ?
- Does  $\sqrt{n}(\hat{\xi} - \xi_*) \rightarrow^d N(0, \Sigma)$ ?
- What is a valid  $\Sigma$ ?

# Goodness-of-fit Statistics

- **Consistency:**  $\hat{\xi}$  converges to  $\xi_*$  (the ‘pseudo-true value’) in probability (or almost surely);
- **Asymptotic Normality:**  $\sqrt{n}(\hat{\xi} - \xi_*)$  converges to a  $N(\mathbf{0}, \Sigma(\xi_*)^{-1})$  distribution.

## Theorem 10.5

If for every  $\epsilon > 0$ ,

$$\sup_{\xi \in \xi} |G_n(\xi) - G(\xi)| \rightarrow^P 0, \quad \sup_{\xi: ||\xi - \xi_*|| \geq \epsilon} G(\xi) < G(\xi_*),$$

then  $\hat{\xi}$  converges to  $\xi_*$  in probability.

## Theorem 10.6

Suppose that the following conditions hold: We have

$$\sqrt{n}(\hat{\xi} - \xi_*) = [-E(\partial_{\xi}^2 \log p(U, \xi_*))]^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n \log p(U_i, \xi_*) + o_p(1). \quad (10.8)$$

In particular,  $\sqrt{n}(\hat{\xi} - \xi_*)$  converges to a  $N(0, \Sigma(\xi_*))$  distribution as  $n \rightarrow \infty$ , where

$$\Sigma(\xi_*) = [-E(\partial_{\xi}^2 \log p(U, \xi_*))]^{-1} [E(\partial_{\xi} \log p(U, \xi_*))^{\otimes 2}] [-E(\partial_{\xi}^2 \log p(U, \xi_*))]^{-1}.$$

# Goodness-of-fit Statistics

- $H_0 : E(y_i) = g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta}_0)$  is true for all  $i$ .
- Under the null hypothesis  $H_0$ , a plot of  $\hat{e}_i = y_i - \hat{\mu}_i$  against  $z_i$  should **oscillate around 0**, where  $z_i$  can be either  $x_i$ , or a function of  $x_i$ .
- This motivates us to combine the pseudo residuals with  $z_i$  to construct several stochastic processes of  $z \in [-\infty, \infty]$  as follows:

$$R(z; \hat{\xi}) = n^{-1/2} \sum_{i=1}^n \hat{e}_i \mathbf{1}(z_i \leq z), \quad (10.9)$$

where  $\mathbf{1}(A)$  denotes an indicator function of an event  $A$ .

- We can construct the Kolmogorov-Smirnov (KS) test statistic  $R = \sup_z |R(z; \hat{\xi})|$ . We reject the null hypothesis  $H_0$  when  $R$  is greater than a threshold at a given significance level.

## Theorem 10.7

*Under some conditions,  $R(z; \hat{\xi})$  converge weakly to  $G(z)$  in Skorokhod space  $D[-\infty, +\infty]$ , where  $G(z)$  is a Gaussian process with zero mean and covariance function  $\Sigma(z_1, z_2)$ .*

## Calculation of $R(z; \hat{\xi})$

- Generate  $\{v_{i,m} : i = 1, \dots, n\}$  from the standard normal distribution  $N(0, 1)$ .
- 

$$\begin{aligned} R^m(z) &= n^{-1/2} \sum_{i=1}^n \{y_i - g^{-1}(\mathbf{x}_i^T \hat{\beta})\} v_{i,m} \mathbf{1}(z_i \leq z) \\ &- n^{-1} \sum_{i=1}^n \partial g^{-1}(\mathbf{x}_i^T \beta_*) \mathbf{x}_i^T \mathbf{1}(z_i \leq z) [\mathbf{I}_p, \mathbf{0}] \\ &- \left\{ \sum_{i=1}^n \partial_{\xi}^2 \log p(y_i, \hat{\xi}) \right\}^{-1} \sum_{i=1}^n v_{i,m} \partial_{\xi} \log p(y_i, \hat{\xi}), \end{aligned}$$

- The third step is to calculate the likelihood ratio  $R^m = \sup_z |R^m(z)|$ .
- We repeat the above three steps  $J$  times and obtain  $J$  realizations:  $\{R^m : m = 1, \dots, J\}$ .

## Example 10.6

*The steam dataset: We want to assess whether the proposed linear regression model provides an adequate fit to the steam data.* We applied the residual processes with  $x_{i2}$  and  $x_{i3}$ , respectively, as  $z_i$  to check the goodness-of-fit of the model. Our results do not reveal the inadequacy of the proposed model. In SAS, the procedure 'proc genmod' is mainly used to fit generalized linear models and generalized estimating equations. The 'model' statement is used to define the response variable and covariates of interest. The option 'assess' computes goodness-of-fit statistics and the variable  $F2$  is used as  $z_i$  in the indicator function. The 'resample' command is the number of iterations  $J$  and 'seed' is the random seed.

# Goodness-of-fit Statistics

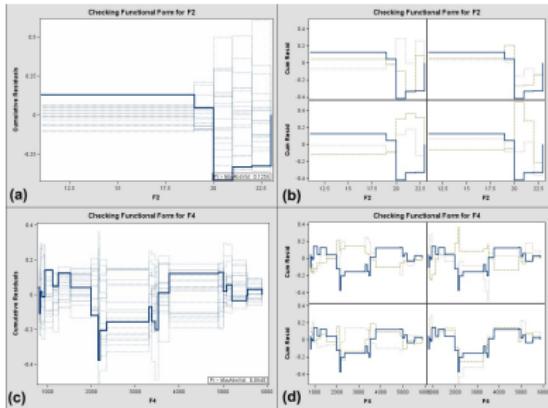


Figure: Residual processes based on  $x_{i2}$  and  $x_{i4}$  for the Steam data.

## Example 10.7

*In Example 5.9, we used the logistic regression model to fit the vegetation data in Example 3.4 in order to model the relationship between the distribution of V2 and the climate variables  $X_1, \dots, X_5$ . Here, we applied the residual processes with  $X_1$  and  $X_2$ , respectively, as  $z_i$  to check the goodness-of-fit of the model. Our results revealed the inadequacy of the proposed model.*

# Goodness-of-fit Statistics

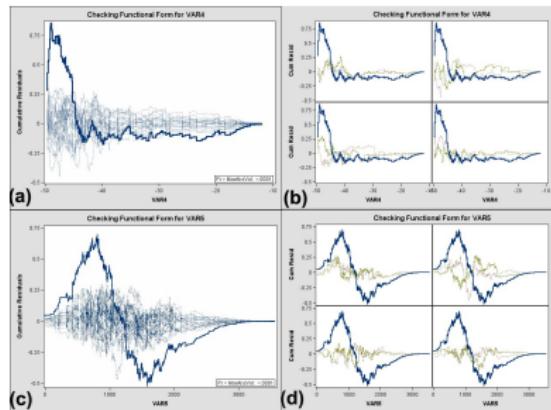


Figure: Residual processes based on  $x_{i2}$  and  $x_{i4}$  for the Steam data.