

# 5100 Methods Notes

Bookmark: Finished Ch11. Still need Ch1.

## Key LM Results

### A General Linear Model (GLM)

Suppose

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where

- $\mathbf{y} \in \mathbb{R}^n$  is the response vector,
- $\mathbf{X}$  is an  $n \times p$  matrix of known (fixed) constants,
- $\boldsymbol{\beta} \in \mathbb{R}^p$  is an unknown parameter vector, and
- $\boldsymbol{\varepsilon}$  is a vector of unobserved random errors satisfying

$$\mathbb{E}(\boldsymbol{\varepsilon}) = \mathbf{0}, \quad \text{Cov}(\boldsymbol{\varepsilon}) = \boldsymbol{\Sigma}.$$

The model is called a *linear model* because the mean of the response vector is linear in the unknown parameter vector:

$$\mathbb{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}.$$

### Ordinary Least Squares (OLS) Estimation

Assume

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \mathbb{E}(\boldsymbol{\varepsilon}) = \mathbf{0}, \quad \text{Cov}(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}.$$

Then

$$\mathbb{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta} \in \mathcal{C}(\mathbf{X}),$$

where  $\mathcal{C}(\mathbf{X})$  denotes the column space of  $\mathbf{X}$ .

To estimate  $\mathbb{E}(\mathbf{y})$ , we consider vectors of the form  $\mathbf{X}\hat{\boldsymbol{\beta}}$ .

Thus, estimating  $\mathbb{E}(\mathbf{y})$  amounts to finding the vector in  $\mathcal{C}(\mathbf{X})$  that is closest to  $\mathbf{y}$ .

Let  $\mathcal{N}(\mathbf{X}^\top)$  denote the null space of  $\mathbf{X}^\top$ .

Then  $\mathcal{C}(\mathbf{X})$  and  $\mathcal{N}(\mathbf{X}^\top)$  are orthogonal complements:

$$\mathcal{N}(\mathbf{X}^\top) \perp \mathcal{C}(\mathbf{X}).$$

The null space of a matrix  $\mathbf{A}$  is defined as

$$\mathcal{N}(\mathbf{A}) = \{\mathbf{x} : \mathbf{Ax} = \mathbf{0}\}.$$

## Least Squares Estimate (LSE)

An estimate  $\hat{\beta}$  is a *least squares estimate* (LSE) of  $\beta$  if  $\mathbf{X}\hat{\beta}$  is the vector in  $\mathcal{C}(\mathbf{X})$  that is closest to  $\mathbf{y}$ . Equivalently,

$$\hat{\beta} = \arg \min_{\beta \in \mathbb{R}^p} (\mathbf{y} - \mathbf{X}\beta)^\top (\mathbf{y} - \mathbf{X}\beta).$$

Define the error sum of squares:

$$Q(\beta) = \|\mathbf{y} - \mathbf{X}\beta\|_2^2 = (\mathbf{y} - \mathbf{X}\beta)^\top (\mathbf{y} - \mathbf{X}\beta).$$

## Identifying the LSE

There are two equivalent approaches:

- **Algebraic:** solving the normal equations
- **Geometric:** orthogonal projection of  $\mathbf{y}$  onto  $\mathcal{C}(\mathbf{X})$

## Normal Equations

Expand the objective function:

$$Q(\beta) = \mathbf{y}^\top \mathbf{y} - 2\beta^\top \mathbf{X}^\top \mathbf{y} + \beta^\top \mathbf{X}^\top \mathbf{X} \beta.$$

Taking derivatives and setting the gradient equal to zero yields

$$\nabla Q(\beta) = -2\mathbf{X}^\top \mathbf{y} + 2\mathbf{X}^\top \mathbf{X} \beta = \mathbf{0}.$$

This leads to the **normal equations**:

$$\mathbf{X}^\top \mathbf{X} \beta = \mathbf{X}^\top \mathbf{y}.$$

## Solutions to the Normal Equations

If  $\text{rank}(\mathbf{X}) = p$ , then  $\mathbf{X}^\top \mathbf{X}$  is invertible and the unique solution is

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}.$$

If  $\text{rank}(\mathbf{X}) < p$ , the normal equations have infinitely many solutions.

In this case,  $\hat{\beta}$  may not be unique, but

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$$

is unique.

## Geometric Approach

Let  $\mathbf{P}_{\mathbf{X}}$  denote the orthogonal projection matrix onto  $\mathcal{C}(\mathbf{X})$ :

$$\mathbf{P}_{\mathbf{X}} = \mathbf{X}(\mathbf{X}^{\top}\mathbf{X})^{-}\mathbf{X}^{\top},$$

where  $(\mathbf{X}^{\top}\mathbf{X})^{-}$  is any generalized inverse.

### Properties

- $\mathbf{P}_{\mathbf{X}}$  is idempotent:

$$\mathbf{P}_{\mathbf{X}}^2 = \mathbf{P}_{\mathbf{X}}.$$

- $\mathbf{P}_{\mathbf{X}}$  projects onto  $\mathcal{C}(\mathbf{X})$ .
- $\mathbf{P}_{\mathbf{X}}$  is symmetric:

$$\mathbf{P}_{\mathbf{X}}^{\top} = \mathbf{P}_{\mathbf{X}}.$$

- $\mathbf{P}_{\mathbf{X}}\mathbf{X} = \mathbf{X}$  and  $\mathbf{X}^{\top}\mathbf{P}_{\mathbf{X}} = \mathbf{X}^{\top}$ .
- $\text{rank}(\mathbf{X}) = \text{rank}(\mathbf{P}_{\mathbf{X}}) = \text{tr}(\mathbf{P}_{\mathbf{X}})$ .

## Fitted Values and Residuals

An estimate  $\hat{\beta}$  is a least squares estimate if and only if

$$\mathbf{X}\hat{\beta} = \mathbf{P}_{\mathbf{X}}\mathbf{y}.$$

The OLS estimator of  $\mathbb{E}(\mathbf{y})$  is

$$\hat{\mathbf{y}} = \mathbf{P}_{\mathbf{X}}\mathbf{y}.$$

The residual vector is

$$\hat{\mathbf{e}} = \mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{y}.$$

Note that

$$\hat{\mathbf{e}} \in \mathcal{N}(\mathbf{X}^{\top}).$$

Since  $\mathcal{C}(\mathbf{X})$  and  $\mathcal{N}(\mathbf{X}^{\top})$  are orthogonal complements, we obtain the unique decomposition

$$\mathbf{y} = \hat{\mathbf{y}} + \hat{\mathbf{e}}.$$

## ANOVA Decomposition for the Linear Model

Suppose  $\mathbf{y}$  is  $n \times 1$ ,  $\mathbf{X}$  is  $n \times p$  with  $\text{rank } r \leq p$ ,  $\beta$  is  $p \times 1$ , and  $\varepsilon$  is  $n \times 1$ . We assume the model given in (1):

$$y = X\beta + \varepsilon.$$

Then, the ANOVA table is:

Source	df	Sum of Squares
Model	$r$	$\hat{y}^\top \hat{y} = y^\top P_X y$
Residual	$n - r$	$\hat{e}^\top \hat{e} = y^\top (I - P_X) y$
Total	$n - 1$	$y^\top y = y^\top I y$

## Starting on estimability

For any  $q \times n$  matrix  $A$ ,  $AE(y)$  is a linear function of  $E(y)$ .

For any  $q \times n$  matrix  $A$ , the OLS estimator of

$$AE(y) = AX\beta$$

is

$$A[\text{OLS Estimator of } E(y)] = A\hat{y} = AP_X y = AX(X^\top X)^- X^\top y.$$

Note that

$$AE(y) = AX\beta$$

is automatically a linear function of  $\beta$  of the form

$$C\beta,$$

where

$$C = AX.$$

If  $C$  is any  $q \times p$  matrix, we say that the linear function of  $\beta$  given by  $C\beta$  is **estimable** if and only if

$$C = AX$$

for some  $q \times n$  matrix  $A$ .

The OLS estimator of an estimable linear function  $C\beta$  is

$$C(X^\top X)^- X^\top y.$$

### Uniqueness of the OLS Estimator of an Estimable $C\beta$

If  $C\beta$  is estimable, then  $C\hat{\beta}$  is the same for all solutions  $\hat{\beta}$  to the normal equations.

In particular, the unique OLS estimator of  $C\beta$  is

$$C\hat{\beta} = C(X^\top X)^- X^\top y = AX(X^\top X)^- X^\top y = AP_X y,$$

where  $C = AX$ .

Furthermore, if  $C\beta$  is estimable, then  $C\hat{\beta}$  is a **linear unbiased estimator** of  $C\beta$ .

The OLS estimator is linear because it is a linear function of  $y$ :

$$C\hat{\beta} = C(X^\top X)^- X^\top y = My,$$

where

$$M = C(X^\top X)^- X^\top.$$

The OLS estimator is unbiased because, for all  $\beta \in \mathbb{R}^p$ ,

$$\begin{aligned} E(C\hat{\beta}) &= E(C(X^\top X)^- X^\top y) \\ &= C(X^\top X)^- X^\top E(y) \\ &= AX(X^\top X)^- X^\top X\beta \\ &= AP_X X\beta \\ &= AX\beta \\ &= C\beta. \end{aligned}$$

### Gauss–Markov Model (GMM)

Suppose

$$y = X\beta + \varepsilon,$$

where

- $y \in \mathbb{R}^n$  is the response vector,
- $X$  is an  $n \times p$  matrix of known constants,
- $\beta \in \mathbb{R}^p$  is an unknown parameter vector, and
- $\varepsilon$  is a vector of random errors satisfying

$$E(\varepsilon) = 0, \quad \text{Var}(\varepsilon) = \sigma^2 I,$$

for some unknown  $\sigma^2 > 0$ .

#### Gauss–Markov Theorem.

The OLS estimator of an estimable function  $C\beta$  is the **Best Linear Unbiased Estimator (BLUE)** of  $C\beta$ , in the sense that it has the smallest variance among all linear unbiased estimators of  $C\beta$ .

## Gauss–Markov Model with Normal Errors (GMMNE)

Suppose

$$y = X\beta + \varepsilon,$$

where

- $y \in \mathbb{R}^n$ ,
- $X$  is an  $n \times p$  matrix of known constants,
- $\beta \in \mathbb{R}^p$  is unknown, and
- 

$$\varepsilon \sim \mathcal{N}(0, \sigma^2 I).$$

### Distribution of $C\hat{\beta}$ and $\hat{\sigma}^2$

In the GMMNE model, the distribution of  $C\hat{\beta}$  is

$$C\hat{\beta} \sim \mathcal{N}(C\beta, \sigma^2 C(X^\top X)^{-1}C^\top).$$

The distribution of  $\hat{\sigma}^2$  is a scaled chi-square distribution:

$$\frac{(n-r)\hat{\sigma}^2}{\sigma^2} \sim \chi_{n-r}^2,$$

equivalently,

$$\hat{\sigma}^2 \sim \frac{\sigma^2}{n-r} \chi_{n-r}^2.$$

Moreover,

$C\hat{\beta}$  and  $\hat{\sigma}^2$  are independent.

### F-Test

For  $H_0 : C\beta = d$

To test

$$H_0 : C\beta = d,$$

use the statistic

$$F = \frac{(C\hat{\beta} - d)^\top \left[ \text{Var}(C\hat{\beta}) \right]^{-1} (C\hat{\beta} - d)}{q}.$$

Since

$$\text{Var}(C\hat{\beta}) = \sigma^2 C(X^\top X)^{-1}C^\top,$$

this becomes

$$F = \frac{(C\hat{\beta} - d)^\top [C(X^\top X)^{-1}C^\top]^{-1} (C\hat{\beta} - d)/q}{\hat{\sigma}^2}.$$

Under  $H_0$ ,  $F$  follows an  $F$  distribution with

$$q \quad \text{and} \quad n - r$$

degrees of freedom.

Under the alternative,  $F$  has a noncentral  $F$  distribution with noncentrality parameter

$$\theta = \frac{(C\beta - d)^\top [C(X^\top X)^{-1}C^\top]^{-1} (C\beta - d)}{2\sigma^2}.$$

The non-negative non-centrality parameter

$$\frac{(C\beta - d)^\top [C(X^\top X)^{-1}C^\top]^{-1} (C\beta - d)}{2\sigma^2}$$

is equal to zero if and only if  $H_0 : C\beta = d$  is true.

If  $H_0 : C\beta = d$  is true, the statistic  $F$  has a **central**  $F$ -distribution with

$$q \quad \text{and} \quad n - r$$

degrees of freedom, denoted  $F_{q,n-r}$ .

### t-Test

For  $(H_0 : c^\top \beta = d)$  for Estimable  $c^\top \beta$

Here,  $c^\top$  is a row vector and  $d$  is a scalar ( $q = 1$ ).

The test statistic is

$$t \equiv \frac{c^\top \hat{\beta} - d}{\sqrt{\widehat{\text{Var}}(c^\top \hat{\beta})}} = \frac{c^\top \hat{\beta} - d}{\sqrt{\hat{\sigma}^2 c^\top (X^\top X)^{-1} c}}.$$

The statistic  $t$  has a non-central  $t$ -distribution with non-centrality parameter

$$\frac{c^\top \beta - d}{\sqrt{\sigma^2 c^\top (X^\top X)^{-1} c}},$$

and degrees of freedom

$$n - r.$$

The non-centrality parameter

$$\frac{c^\top \beta - d}{\sqrt{\sigma^2 c^\top (X^\top X)^{-1} c}}$$

is equal to zero if and only if  $H_0 : c^\top \beta = d$  is true.

If  $H_0 : c^\top \beta = d$  is true, the statistic  $t$  has a **central**  $t$ -distribution with

$$n - r$$

degrees of freedom, denoted  $t_{n-r}$ .

### Confidence Interval

For Estimable  $c^\top \beta$ , a  $100(1 - \alpha)\%$  confidence interval for estimable  $c^\top \beta$  is given by

$$c^\top \hat{\beta} \pm t_{n-r, 1-\alpha/2} \sqrt{\hat{\sigma}^2 c^\top (X^\top X)^{-1} c}.$$

That is,

$$\text{estimate} \pm (\text{distribution quantile}) \times (\text{estimated standard error}).$$

## Reduced vs. Full

### Model and Hypotheses

Assume the Gauss–Markov model with normal errors:

$$y = X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I).$$

Suppose  $\mathcal{C}(X_0) \subset \mathcal{C}(X)$  and we wish to test

$$H_0 : E(y) \in \mathcal{C}(X_0) \quad \text{vs.} \quad H_A : E(y) \in \mathcal{C}(X) \setminus \mathcal{C}(X_0).$$

- The *reduced* model corresponds to the null hypothesis and states that

$$E(y) \in \mathcal{C}(X_0),$$

a specified subspace of  $\mathcal{C}(X)$ .

- The *full* model states that  $E(y)$  can be anywhere in  $\mathcal{C}(X)$ .

### Example

Suppose a reduced model is that every group has the same mean, and suppose the full model has every group has a unique mean.

Given:



$$X_0 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

$$X = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

### Model Interpretation:

- The **reduced model** says we assume the same mean for all observations:

$$E(y) = \mu$$

This corresponds to the design matrix  $X_0$ .

- The **full model** says there are 3 distinct means: each group (of size 2) has its own mean.  
This corresponds to the design matrix  $X$ , which is a  $4 \times 3$  matrix (though it codes for 3 groups).

### Test Statistic

For the general case, consider the test statistic

$$F = \frac{y^\top (P_X - P_{X_0})y / [\text{rank}(X) - \text{rank}(X_0)]}{y^\top (I - P_X)y / [n - \text{rank}(X)]}.$$

- When the reduced model is correct, the numerator and denominator of the  $F$  statistic are both unbiased estimators of  $\sigma^2$ , so  $F$  should be close to 1.
- When the reduced model is not correct, the numerator of the  $F$  statistic estimates something larger than  $\sigma^2$ , so  $F$  should be larger than 1. Thus, values of  $F$  much larger than 1 are not consistent with the reduced model being correct.

### Deriving the Distribution of $F$

Our main assumption about the model is

$$\varepsilon \sim \mathcal{N}(0, \sigma^2 I) \implies y \sim \mathcal{N}(X\beta, \sigma^2 I).$$

Recall the following result:

- Suppose  $\Sigma$  is an  $n \times n$  positive definite matrix.
- Suppose  $A$  is an  $n \times n$  symmetric matrix of rank  $m$  such that  $A\Sigma$  is idempotent (i.e.,  $A\Sigma A\Sigma = A\Sigma$ ).

Then, if  $y \sim \mathcal{N}(\mu, \Sigma)$ ,

$$y^\top A y \sim \chi_m^2 \left( \frac{\mu^\top A \mu}{2} \right).$$

### Distribution of the Numerator

For the numerator of our  $F$  statistic, we have

$$\mu = X\beta, \quad \Sigma = \sigma^2 I, \quad A = \frac{P_X - P_{X_0}}{\sigma^2}.$$

The rank is

$$\begin{aligned} m = \text{rank}(A) &= \text{rank} \left( \frac{P_X - P_{X_0}}{\sigma^2} \right) = \text{rank}(P_X - P_{X_0}) \\ &= \text{tr}(P_X - P_{X_0}) = \text{tr}(P_X) - \text{tr}(P_{X_0}) \\ &= \text{rank}(P_X) - \text{rank}(P_{X_0}) = \text{rank}(X) - \text{rank}(X_0). \end{aligned}$$

Therefore,

$$\frac{y^\top (P_X - P_{X_0}) y}{\sigma^2} \sim \chi_{\text{rank}(X) - \text{rank}(X_0)}^2(\theta),$$

where

$$\theta = \frac{1}{2} \beta^\top X^\top \left( \frac{P_X - P_{X_0}}{\sigma^2} \right) X \beta.$$

### Distribution of the Denominator

The denominator (mean squared error) is

$$\text{MSE} = \frac{y^\top (I - P_X) y}{n - \text{rank}(X)}.$$

Its distribution is

$$\frac{y^\top (I - P_X) y}{\sigma^2} \sim \chi_{n - \text{rank}(X)}^2.$$

This distributional result holds regardless of whether or not the reduced model is correct.

## Independence of Numerator and Denominator

We can show that

$$\frac{y^\top (P_X - P_{X_0})y}{\sigma^2} \perp \frac{y^\top (I - P_X)y}{\sigma^2},$$

because

$$\left( \frac{P_X - P_{X_0}}{\sigma^2} \right) (\sigma^2 I) \left( \frac{I - P_X}{\sigma^2} \right) = 0.$$

Indeed,

$$\begin{aligned} \frac{1}{\sigma^2} (P_X - P_X P_X - P_{X_0} + P_{X_0} P_X) &= \frac{1}{\sigma^2} (P_X - P_X - P_{X_0} + P_{X_0}) \\ &= 0. \end{aligned}$$

## Distribution of F

Thus, it follows that

$$F = \frac{y^\top (P_X - P_{X_0})y / [\text{rank}(X) - \text{rank}(X_0)]}{y^\top (I - P_X)y / [n - \text{rank}(X)]} \sim F_{\text{rank}(X) - \text{rank}(X_0), n - \text{rank}(X)}(\theta),$$

where

$$\theta = \frac{\beta^\top X^\top (P_X - P_{X_0}) X \beta}{2\sigma^2}.$$

## Noncentrality Parameter

- If  $H_0$  is true, i.e., if  $E(y) = X\beta \in \mathcal{C}(X_0)$ , then the noncentrality parameter  $\theta$  is 0 because

$$(P_X - P_{X_0})X\beta = P_X X\beta - P_{X_0} X\beta = X\beta - X\beta = 0.$$

Hence,

$$\frac{y^\top (P_X - P_{X_0})y}{\sigma^2} \sim \chi_{\text{rank}(X) - \text{rank}(X_0)}^2,$$

a central  $\chi^2$  distribution.

- If  $H_0$  is false and  $E(y) = X\beta \notin \mathcal{C}(X_0)$ , then  $(P_X - P_{X_0})X\beta \neq 0$  and  $\theta > 0$ . Hence,

$$\frac{y^\top (P_X - P_{X_0})y}{\sigma^2} \sim \chi_{\text{rank}(X) - \text{rank}(X_0)}^2(\theta).$$

In general, the noncentrality parameter quantifies how far the mean of  $y$  is from  $\mathcal{C}(X_0)$  because

$$\begin{aligned} \beta^\top X^\top (P_X - P_{X_0})X\beta &= \beta^\top X^\top (P_X - P_{X_0})^\top (P_X - P_{X_0})X\beta \\ &= \|(P_X - P_{X_0})X\beta\|^2 = \|P_X X\beta - P_{X_0} X\beta\|^2 \\ &= \|X\beta - P_{X_0} X\beta\|^2 = \|E(y) - P_{X_0} E(y)\|^2. \end{aligned}$$

If  $E(y)$  indeed lies in  $\mathcal{C}(X_0)$ , then  $P_{X_0} E(y) = E(y)$ .

## Useful Identities

Note that

$$\begin{aligned} y^\top (P_X - P_{X_0})y &= y^\top [(I - P_{X_0}) - (I - P_X)]y \\ &= y^\top (I - P_{X_0})y - y^\top (I - P_X)y \\ &= \text{SSE}_{\text{REDUCED}} - \text{SSE}_{\text{FULL}}. \end{aligned}$$

Also,

$$\begin{aligned} \text{rank}(X) - \text{rank}(X_0) &= [n - \text{rank}(X_0)] - [n - \text{rank}(X)] \\ &= \text{DFE}_{\text{REDUCED}} - \text{DFE}_{\text{FULL}}, \end{aligned}$$

where DFE denotes degrees of freedom for error.

*Result*

Thus, the  $F$  statistic has the familiar form

$$F = \frac{(\text{SSE}_{\text{REDUCED}} - \text{SSE}_{\text{FULL}})/(\text{DFE}_{\text{REDUCED}} - \text{DFE}_{\text{FULL}})}{\text{SSE}_{\text{FULL}}/\text{DFE}_{\text{FULL}}}.$$

## Equivalence of $F$ -Tests

It turns out that this reduced vs. full model  $F$ -test is equivalent to the  $F$ -test for testing

$$H_0 : C\beta = d \quad \text{vs.} \quad H_A : C\beta \neq d,$$

with an appropriately chosen  $C$  and  $d$ .

# Two-Factor Cell-Means Models

## An Example Two-Factor Experiment

Researchers were interested in studying the effects of 2 diets (low fiber, high fiber) and 3 drugs (D1, D2, D3) on weight gained by Yorkshire pigs. A total of 12 pigs were assigned to the 6 diet  $\times$  drug combinations using a balanced and completely randomized experimental design. Pigs were housed in individual pens, injected with their assigned drugs once per week, and fed their assigned diets for a 6-week period. The amount of weight gained during the 6-week period was recorded for each pig.

## Factors, Levels, Design

This experiment involves 2 factors: **Diet** and **Drug**.

- The factor **Diet** has 2 levels: low fiber and high fiber.
- The factor **Drug** has 3 levels: D1, D2, and D3.

## Treatment Design vs. Experimental Design

- A combination of one level from each factor forms a *treatment*.
- The *treatment design* used in this experiment is known as a **full-factorial treatment design** because each possible combination of one level from each factor was applied to at least one experimental unit.
- The *experimental design* is a balanced **completely randomized design (CRD)** because all possible balanced assignments of the 12 pigs to the 6 treatment groups were equally likely.

## The Cell-Means Model

For  $i = 1, 2$ ,  $j = 1, 2, 3$ , and  $k = 1, 2$ , let  $y_{ijk}$  denote the weight gain of the  $k$ th pig that received diet  $i$  and drug  $j$ , and suppose

$$y_{ijk} = \mu_{ij} + \varepsilon_{ijk}, \quad \varepsilon_{ijk} \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2).$$

Here,

$$(\mu_{11}, \mu_{12}, \mu_{13}, \mu_{21}, \mu_{22}, \mu_{23}) \in \mathbb{R} \quad \text{and} \quad \sigma^2 \in \mathbb{R}^+$$

are unknown parameters. The  $\mu_{ij}$  represent the *treatment (cell) means*.

A cell-means table is given by

	Drug 1	Drug 2	Drug 3
Diet 1	$\mu_{11}$	$\mu_{12}$	$\mu_{13}$
Diet 2	$\mu_{21}$	$\mu_{22}$	$\mu_{23}$

## Estimability of $\beta$

For the General Linear Model, the parameter vector  $\beta$  is estimable whenever  $X$  has full column rank, i.e.,

$$\text{rank}(X) = p.$$

## Least Squares Means (LSMEANS) in SAS

SAS can be used to compute LSMEANS.

LSMEANS are simply OLS estimators of cell or marginal means.

Each LSMEAN has the form

$$c^\top \hat{\beta}$$

for an appropriate vector  $c$ .

For example, the LSMEAN for Diet 1 is  $c^\top \hat{\beta}$  with

$$c^\top = \left[ \frac{1}{3}, \frac{1}{3}, \frac{1}{3}, 0, 0, 0 \right], \quad \hat{\beta} = [\bar{y}_{11\cdot}, \bar{y}_{12\cdot}, \bar{y}_{13\cdot}, \bar{y}_{21\cdot}, \bar{y}_{22\cdot}, \bar{y}_{23\cdot}]^\top.$$

Thus, the LSMEAN for Diet 1 is

$$\frac{\bar{y}_{11\cdot} + \bar{y}_{12\cdot} + \bar{y}_{13\cdot}}{3},$$

an estimator of the marginal mean  $\mu_{1\cdot}$ .

Note that the LSMEAN for Diet 1 is simply an average of the estimated means for treatments involving Diet 1.

When data are balanced, the LSMEAN for Diet 1 is also just the average of responses for all pigs that were fed Diet 1.

When data are unbalanced, the LSMEAN for Diet 1 may not equal the average of responses for all pigs that were fed Diet 1.

## Standard Error

A *standard error* is the estimated standard deviation of a statistic.

A standard error is usually found by estimating the variance of a statistic and then taking the square root of the estimate.

Because each LSMEAN has the form  $c^\top \hat{\beta}$  for an appropriate vector  $c$ , the standard error for an LSMEAN is given by

$$\sqrt{\widehat{\text{Var}}(c^\top \hat{\beta})} = \sqrt{\hat{\sigma}^2 c^\top (X^\top X)^{-1} c}.$$

## Effects We Can Estimate

- Simple effects
- Main effects
- Interactions

## Simple Effects

A **simple effect** is the difference between cell means that differ in level for only one factor.

In our two-factor example, simple effects are differences between cell means within any row or within any column.

Consider a two-factor layout:

- **Simple effect of diet within drug 1** compares  $\mu_{11}$  (diet 1, drug 1) and  $\mu_{21}$  (diet 2, drug 1).
- Similarly, **simple effect of drug** can be examined within a fixed diet level.

For example: - Drug 1, Diet 1 vs. Diet 2  $\rightarrow$  Simple effect of Diet within Drug 1 - Diet 1, Drug 2 vs. Drug 3  $\rightarrow$  Simple effect of Drug within Diet 1

**Note:** A contrast such as  $\mu_{22} - \mu_{13}$  is **not** a simple effect, because it differs in both factors (diet and drug).

*Continued*

The simple effect of **Diet for Drug 1** is:

$$\mu_{11} - \mu_{21}$$

The simple effect of **Drug 2 vs. Drug 3 for Diet 2** is:

$$\mu_{22} - \mu_{23}$$

Where  $\mu_{ij}$  denotes the mean response for diet  $i$  and drug  $j$ .

## Main Effects

A *main effect* is the difference between marginal means associated with two levels of a factor.

In our two-factor example, the *main effect* of Diet is

$$\bar{\mu}_{1.} - \bar{\mu}_{2.}$$

In our two-factor example, the *main effects* of Drug involve the differences

$$\bar{\mu}_{.1} - \bar{\mu}_{.2}, \quad \bar{\mu}_{.1} - \bar{\mu}_{.3}, \quad \text{and} \quad \bar{\mu}_{.2} - \bar{\mu}_{.3}.$$

If

$$\bar{\mu}_{1.} = \bar{\mu}_{2.},$$

it would be customary to say, “There is no Diet main effect.”

If

$$\bar{\mu}_{.1} = \bar{\mu}_{.2} = \bar{\mu}_{.3},$$

it would be customary to say, “There are no Drug main effects.”

## Interaction Effects

The linear combination

$$\mu_{ij} - \mu_{ij'} - \mu_{i'j} + \mu_{i'j'}$$

for  $i \neq i'$  and  $j \neq j'$  is an interaction effect.

Every interaction can be expressed using this format.

For example,

$$\mu_{11} - \mu_{12} - \mu_{21} + \mu_{22} = (\mu_{11} - \mu_{12}) - (\mu_{21} - \mu_{22}) = (\mu_{11} - \mu_{21}) - (\mu_{12} - \mu_{22})$$

is an interaction effect.

These contrasts are equal if there is no interaction.

*Continued*

When all interaction effects are zero, we may say there are “no interactions” between the factors, or that the two factors do not interact.

When there are no interactions between factors, the simple effects of either factor are the same across all levels of the other factor.

For example, when there are no interactions between the factors Diet and Drug, the simple effect of Diet is the same for each level of Drug.

Likewise, any simple effect of Drug is the same for both diets.

## Testing for Non-Zero Effects

We can test whether simple effects, main effects, or interaction effects are zero versus non-zero using tests of the form

$$H_0 : C\beta = \mathbf{0} \quad \text{vs.} \quad H_A : C\beta \neq \mathbf{0}.$$

To properly set up  $C$ , look at  $\beta$  and how the parameters are arranged in  $\beta$ .

## Alternative Parametrization of Two-Factor Cell-Means Models

An alternative parameterization of the cell-means model is

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}, \quad (i = 1, 2; j = 1, 2, 3; k = 1, 2).$$

Here, -  $\mu$  is the intercept (overall mean), -  $\alpha_i$  is the effect associated with Diet, -  $\beta_j$  is the effect associated with Drug, -  $\gamma_{ij}$  is the interaction between Diet and Drug.

The parameters

$$\mu, \alpha_1, \alpha_2, \beta_1, \beta_2, \beta_3, \gamma_{11}, \gamma_{12}, \gamma_{13}, \gamma_{21}, \gamma_{22}, \gamma_{23}$$

are unknown real-valued parameters, and



$$\varepsilon_{111}, \varepsilon_{112}, \varepsilon_{121}, \varepsilon_{122}, \varepsilon_{131}, \varepsilon_{132}, \varepsilon_{211}, \varepsilon_{212}, \varepsilon_{221}, \varepsilon_{222}, \varepsilon_{231}, \varepsilon_{232}$$

are independent and identically distributed with

$$\varepsilon_{ijk} \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2),$$

for some unknown  $\sigma^2 > 0$ .

## Table of Treatments and Means

Treatment	Diet	Drug	Mean
1	1	1	$\mu + \alpha_1 + \beta_1 + \gamma_{11}$
2	1	2	$\mu + \alpha_1 + \beta_2 + \gamma_{12}$
3	1	3	$\mu + \alpha_1 + \beta_3 + \gamma_{13}$
4	2	1	$\mu + \alpha_2 + \beta_1 + \gamma_{21}$
5	2	2	$\mu + \alpha_2 + \beta_2 + \gamma_{22}$
6	2	3	$\mu + \alpha_2 + \beta_3 + \gamma_{23}$

Diet 1 = Low Fiber, Diet 2 = High Fiber

Drug 1 = D1, Drug 2 = D2, Drug 3 = D3

## Cell and Marginal Means

Any linear combination of the entries in this table is estimable.

The cell means are

$$\mu + \alpha_i + \beta_j + \gamma_{ij}.$$

The Diet marginal means are

$$\bar{\mu}_{1\cdot} = \mu + \alpha_1 + \bar{\beta}\cdot + \bar{\gamma}_{1\cdot}, \quad \bar{\mu}_{2\cdot} = \mu + \alpha_2 + \bar{\beta}\cdot + \bar{\gamma}_{2\cdot}.$$

Thus, the main effect of Diet is

$$\bar{\mu}_{1\cdot} - \bar{\mu}_{2\cdot} = \alpha_1 - \alpha_2 + \bar{\gamma}_{1\cdot} - \bar{\gamma}_{2\cdot}.$$

An example of a simple effect of Drug within Diet 1 is

$$(\mu + \alpha_1 + \beta_1 + \gamma_{11}) - (\mu + \alpha_1 + \beta_2 + \gamma_{12}) = \beta_1 - \beta_2 + \gamma_{11} - \gamma_{12}.$$

## Estimable Functions

Simple effect of Diet for Drug 1:

$$\alpha_1 - \alpha_2 + \gamma_{11} - \gamma_{21}.$$

Simple effect of Drug 1 vs. Drug 3 for Diet 2:

$$\beta_1 - \beta_3 + \gamma_{21} - \gamma_{23}.$$

Main effect of Diet:

$$\alpha_1 - \alpha_2 + \bar{\gamma}_{1\cdot} - \bar{\gamma}_{2\cdot}.$$

Interaction effect involving Diets 1 and 2 and Drugs 1 and 3:

$$\left[ (\mu + \alpha_1 + \beta_1 + \gamma_{11}) - (\mu + \alpha_2 + \beta_1 + \gamma_{21}) \right] - \left[ (\mu + \alpha_1 + \beta_3 + \gamma_{13}) - (\mu + \alpha_2 + \beta_3 + \gamma_{23}) \right]$$

which simplifies to

$$\gamma_{11} - \gamma_{13} - \gamma_{21} + \gamma_{23}.$$

## Estimation and Testing

As before, estimation or testing involves finding an appropriate matrix  $C$  to estimate  $C\beta$  or test

$$H_0 : C\beta = 0.$$

## Tests Based on Reduced vs. Full Model Comparison

Any of the tests we have discussed could alternatively be carried out using a statistic of the form

$$F = \frac{\mathbf{y}^\top (P_X - P_{X_0}) \mathbf{y} / [\text{rank}(X) - \text{rank}(X_0)]}{\mathbf{y}^\top (I - P_X) \mathbf{y} / [n - \text{rank}(X)]},$$

for an appropriate reduced model matrix  $X_0$ .

It is not always easy to specify an appropriate matrix  $X_0$ .

## Misc

### Testing for Main Effects When Factors Interact

Some statisticians argue against testing for main effects when there are interactions between factors.

Others believe that, depending on the scientific questions of interest, any contrasts of treatment means may be worth examining.

Be aware that “no main effects” does not necessarily mean “no effects.”

## Unbalanced Data and Missing Cells

Although we have focused on a balanced two-factor experiment with 2 experimental units per treatment, the techniques presented in these slides work the same way whether data are balanced or not, as long as each treatment has a response for at least one experimental unit and some treatments have more than one.

If there are no experimental units for one or more treatments, then the treatment design may not be a full-factorial treatment design, and we may have a *missing cell* or *missing cells*.

### Missing Cells

Consider the following layout with a missing cell (no data for Diet 2 with Drug 2):

	Drug 1	Drug 2	Drug 3
Diet 1	$\mu_{11}$	$\mu_{12}$	$\mu_{13}$
Diet 2	$\mu_{21}$	Missing	$\mu_{23}$

In this example, we have no data for the treatment combination Diet 2 with Drug 2.

In this case, we could fit a model with the 5 means:

$$\mu_{11}, \mu_{12}, \mu_{13}, \mu_{21}, \text{ and } \mu_{23}.$$

We could estimate any linear combination of these 5 means.

**However**, linear combinations involving  $\mu_{22}$  are **not estimable** because there is no data for that treatment combination.

## Two Factor Additive Models

When factors do not interact, it makes sense to consider the *additive model*:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}, \quad (i = 1, 2; j = 1, 2, 3; k = 1, 2).$$

Here,  $\mu, \alpha_1, \alpha_2, \beta_1, \beta_2, \beta_3$  are unknown real-valued parameters, and -

$$\varepsilon_{111}, \varepsilon_{112}, \varepsilon_{121}, \varepsilon_{122}, \varepsilon_{131}, \varepsilon_{132}, \varepsilon_{211}, \varepsilon_{212}, \varepsilon_{221}, \varepsilon_{222}, \varepsilon_{231}, \varepsilon_{232}$$

are independent and identically distributed with

$$\varepsilon_{ijk} \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2),$$

for some unknown  $\sigma^2 > 0$ .

## Cell Means for the Additive Model

- All interactions are zero for the additive model.
- The simple effect of Diet is  $\alpha_1 - \alpha_2$  for all levels of Drug.
- The simple effect of Drug  $j$  vs. Drug  $j'$  is  $\beta_j - \beta_{j'}$ , regardless of Diet.

The table of cell means is:

	Drug 1	Drug 2	Drug 3
Diet 1	$\mu + \alpha_1 + \beta_1$	$\mu + \alpha_1 + \beta_2$	$\mu + \alpha_1 + \beta_3$
Diet 2	$\mu + \alpha_2 + \beta_1$	$\mu + \alpha_2 + \beta_2$	$\mu + \alpha_2 + \beta_3$

The marginal mean difference for Diet is

$$\bar{\mu}_{1\cdot} - \bar{\mu}_{2\cdot} = \alpha_1 - \alpha_2.$$

Estimation of  $\alpha_1$  and  $\alpha_2$  uses all available data across all three levels of Drug.

## Marginal Means for the Additive Model

Averaging over Drug yields the Diet marginal means:

$$\mu + \alpha_1 + \bar{\beta}, \quad \mu + \alpha_2 + \bar{\beta},$$

where

$$\bar{\beta} = \frac{\beta_1 + \beta_2 + \beta_3}{3}.$$

Thus, the difference between Diet marginal means is

$$(\mu + \alpha_1 + \bar{\beta}) - (\mu + \alpha_2 + \bar{\beta}) = \alpha_1 - \alpha_2.$$

Averaging over Diet yields the Drug marginal means:

$$\mu + \bar{\alpha} + \beta_j,$$

where

$$\bar{\alpha} = \frac{\alpha_1 + \alpha_2}{2}.$$

Differences such as  $\beta_1 - \beta_2$  or  $\beta_2 - \beta_3$  represent main effects of Drug.

## Tests for Main Effects in the Additive Model

No Diet main effect is equivalent to

$$\alpha_1 = \alpha_2.$$

No Drug main effects is equivalent to

$$\beta_1 = \beta_2 = \beta_3.$$

## LSMEANS for the Additive Model

LSMEANS are OLS estimators of the quantities in the margins below.

The Diet marginal LSMEANS are

$$\mu + \alpha_1 + \bar{\beta}, \quad \mu + \alpha_2 + \bar{\beta}.$$

The Drug marginal LSMEANS are

$$\mu + \bar{\alpha} + \beta_1, \quad \mu + \bar{\alpha} + \beta_2, \quad \mu + \bar{\alpha} + \beta_3.$$

For example, the LSMEAN for Diet 1 can be written as

$$c^\top \hat{\beta} = \begin{bmatrix} 1 & 1 & 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{\alpha}_1 \\ \hat{\alpha}_2 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{bmatrix} = \hat{\mu} + \hat{\alpha}_1 + \frac{\hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3}{3}.$$

Here,  $\hat{\beta}$  is any solution to the Normal Equations.

Although  $\hat{\beta}$  depends on which of infinitely many solutions is used, the quantity  $c^\top \hat{\beta}$  is the same for all solutions.

## R Full-Rank Formulation

Under the R full-rank (treatment-coded) formulation for the additive model, the table of means is:

	Drug 1	Drug 2	Drug 3	Diet Marginal
Diet 1	$\mu$	$\mu + \beta_2$	$\mu + \beta_3$	$\mu + \frac{\beta_2 + \beta_3}{3}$
Diet 2	$\mu + \alpha_2$	$\mu + \alpha_2 + \beta_2$	$\mu + \alpha_2 + \beta_3$	$\mu + \alpha_2 + \frac{\beta_2 + \beta_3}{3}$
Drug Marginal	$\mu + \frac{\alpha_2}{2}$	$\mu + \frac{\alpha_2}{2} + \beta_2$	$\mu + \frac{\alpha_2}{2} + \beta_3$	$\mu + \frac{\alpha_2}{2} + \frac{\beta_2 + \beta_3}{3}$

## Main Effects

This parameterization differs from the earlier sum-to-zero setup and instead uses a full-rank model matrix.

No Diet main effect is equivalent to

$$\alpha_2 = 0.$$

No Drug main effects is equivalent to

$$\beta_2 = \beta_3 = 0.$$

Under  $\beta_2 = \beta_3 = 0$ , all Drug marginal means collapse to the same value.

## Diet Main Effect

The null hypothesis of no Diet main effect is

$$H_0 : \alpha_2 = 0.$$

This can be written in matrix form as

$$C^\top \boldsymbol{\beta} = 0,$$

where

$$C^\top = \begin{bmatrix} 0 & 1 & 0 & 0 \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \mu \\ \alpha_2 \\ \beta_2 \\ \beta_3 \end{bmatrix}.$$

## Drug Main Effects

The null hypothesis of no Drug main effects is

$$H_0 : \beta_2 = \beta_3 = 0.$$

This can be written as

$$C\boldsymbol{\beta} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_2 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

# ANOVA

## Setup and Notation

We consider the general linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}).$$

Let

$$\mathbf{X}_1 = \mathbf{1}, \quad \mathbf{X}_m = \mathbf{X}, \quad \mathbf{X}_{m+1} = \mathbf{I}.$$

Suppose  $\mathbf{X}_2, \dots, \mathbf{X}_m$  are matrices satisfying the nested column space condition

$$\mathcal{C}(\mathbf{X}_1) \subset \mathcal{C}(\mathbf{X}_2) \subset \dots \subset \mathcal{C}(\mathbf{X}_m).$$

Let

$$\mathbf{P}_j = \mathbf{P}_{\mathbf{X}_j}, \quad r_j = \text{rank}(\mathbf{X}_j), \quad j = 1, \dots, m+1.$$

## The Total Sum of Squares

The **total sum of squares** (also called the corrected total sum of squares) is

$$\sum_{i=1}^n (y_i - \bar{y})^2.$$

In matrix form,

$$\sum_{i=1}^n (y_i - \bar{y})^2 = (\mathbf{y} - \bar{y}\mathbf{1})^\top (\mathbf{y} - \bar{y}\mathbf{1}) = (\mathbf{y} - \mathbf{P}_1 \mathbf{y})^\top (\mathbf{y} - \mathbf{P}_1 \mathbf{y}).$$

Since  $\mathbf{P}_1$  is symmetric and idempotent,

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \mathbf{y}^\top (\mathbf{I} - \mathbf{P}_1) \mathbf{y}.$$

## Partitioning the Total Sum of Squares

Recall that  $\mathbf{X}_{m+1} = \mathbf{I}$ , so  $\mathbf{P}_{m+1} = \mathbf{I}$ . Then

$$\mathbf{y}^\top (\mathbf{I} - \mathbf{P}_1) \mathbf{y} = \mathbf{y}^\top (\mathbf{P}_{m+1} - \mathbf{P}_1) \mathbf{y}.$$

Insert intermediate projections:

$$\mathbf{y}^\top (\mathbf{P}_{m+1} - \mathbf{P}_1) \mathbf{y} = \mathbf{y}^\top \left( \sum_{j=2}^{m+1} \mathbf{P}_j - \sum_{j=1}^m \mathbf{P}_j \right) \mathbf{y}.$$

Rearranging,

$$\mathbf{y}^\top (\mathbf{I} - \mathbf{P}_1) \mathbf{y} = \mathbf{y}^\top (\mathbf{P}_{m+1} - \mathbf{P}_m) \mathbf{y} + \cdots + \mathbf{y}^\top (\mathbf{P}_2 - \mathbf{P}_1) \mathbf{y}.$$

Equivalently,

$$\mathbf{y}^\top (\mathbf{I} - \mathbf{P}_1) \mathbf{y} = \sum_{j=1}^m \mathbf{y}^\top (\mathbf{P}_{j+1} - \mathbf{P}_j) \mathbf{y}.$$

### Sums of Squares Representation

The quantities in

$$\mathbf{y}^\top (\mathbf{I} - \mathbf{P}_1) \mathbf{y} = \sum_{j=1}^m \mathbf{y}^\top (\mathbf{P}_{j+1} - \mathbf{P}_j) \mathbf{y}$$

are often arranged in an ANOVA table.

We define

$$SS(j+1 \mid j) = \mathbf{y}^\top (\mathbf{P}_{j+1} - \mathbf{P}_j) \mathbf{y}.$$

In particular,

$$SSE = \mathbf{y}^\top (\mathbf{I} - \mathbf{P}_{\mathbf{X}}) \mathbf{y}.$$

### Interpretation of Sequential Sums of Squares

Note that

$$\begin{aligned} SS(j+1 \mid j) &= \mathbf{y}^\top (\mathbf{P}_{j+1} - \mathbf{P}_j) \mathbf{y} \\ &= \mathbf{y}^\top (\mathbf{I} - \mathbf{P}_j) \mathbf{y} - \mathbf{y}^\top (\mathbf{I} - \mathbf{P}_{j+1}) \mathbf{y} \\ &= SSE_j - SSE_{j+1}. \end{aligned}$$

Thus,  $SS(j+1 \mid j)$  is the **reduction in error sum of squares** when projecting  $\mathbf{y}$  onto

$$\mathcal{C}(\mathbf{X}_{j+1}) \quad \text{instead of} \quad \mathcal{C}(\mathbf{X}_j).$$

### Sequential (Type I) Sums of Squares

The quantities

$$SS(j+1 \mid j), \quad j = 1, \dots, m-1,$$

are called **Sequential Sums of Squares**.

In SAS terminology, these are known as **Type I Sums of Squares**.

Generally, the Type (I, II, III, IV) will refer to what elements of the ANOVA table are being conditioned on, particularly the first element of the table.



## Properties of the Matrices of the Quadratic Forms

The matrices of the quadratic forms in the ANOVA table have several useful properties:

- **Symmetry**  
 $A = A^T$
- **Idempotency**  
 $A^2 = A$
- **Rank relationship**  
 $\text{rank}(P_{j+1} - P_j) = r_{j+1} - r_j$
- **Zero Cross-Products**  
 $(P_{j+1} - P_j)(P_{k+1} - P_k) = 0$  for  $j \neq k$

## Distribution of Scaled ANOVA Sums of Squares

Given  $y \sim N(X\beta, \sigma^2 I)$  and an idempotent matrix  $A$ :

$$y^T A y \sim \sigma^2 \chi_{\text{rank}(A)}^2 \left( \frac{\beta^T X^T A X \beta}{2\sigma^2} \right)$$

Specifically, for the projection matrices  $P_j$  in the nested sequence:

Because

$$\frac{P_{j+1} - P_j}{\sigma^2} \cdot \sigma^2 I = P_{j+1} - P_j$$

is idempotent,

we have:

$$y^T (P_{j+1} - P_j) y \sim \sigma^2 \chi_{r_{j+1} - r_j}^2 \left( \frac{\beta^T X^T (P_{j+1} - P_j) X \beta}{2\sigma^2} \right)$$

for all  $j = 1, \dots, m$ .

## ANOVA Tables

ANOVA with Degrees of Freedom

Consider the sequential sums of squares

$$\mathbf{y}^\top (\mathbf{P}_{j+1} - \mathbf{P}_j) \mathbf{y}, \quad j = 1, \dots, m.$$

Each line in the ANOVA table corresponds to a number of degrees of freedom equal to the increase in rank when moving from  $\mathbf{X}_j$  to  $\mathbf{X}_{j+1}$ .

**ANOVA Table with degrees of freedom is:**

Sum of Squares	Degrees of Freedom	DF
$\mathbf{y}^\top (\mathbf{P}_2 - \mathbf{P}_1) \mathbf{y}$	$\text{rank}(\mathbf{X}_2) - \text{rank}(\mathbf{X}_1)$	$r_2 - 1$
$\mathbf{y}^\top (\mathbf{P}_3 - \mathbf{P}_2) \mathbf{y}$	$\text{rank}(\mathbf{X}_3) - \text{rank}(\mathbf{X}_2)$	$r_3 - r_2$
$\vdots$	$\vdots$	$\vdots$
$\mathbf{y}^\top (\mathbf{P}_m - \mathbf{P}_{m-1}) \mathbf{y}$	$\text{rank}(\mathbf{X}_m) - \text{rank}(\mathbf{X}_{m-1})$	$r - r_{m-1}$
$\mathbf{y}^\top (\mathbf{P}_{m+1} - \mathbf{P}_m) \mathbf{y}$	$\text{rank}(\mathbf{X}_{m+1}) - \text{rank}(\mathbf{X}_m)$	$n - r$
$\mathbf{y}^\top (\mathbf{I} - \mathbf{P}_1) \mathbf{y}$	$\text{rank}(\mathbf{X}_{m+1}) - \text{rank}(\mathbf{X}_1)$	$n - 1$

### ANOVA Table with Mean Squares

Dividing each sum of squares by its corresponding degrees of freedom yields the mean squares.

Sum of Squares	Degrees of Freedom	Mean Square
SS(2   1)	$r_2 - 1$	MS(2   1)
SS(3   2)	$r_3 - r_2$	MS(3   2)
$\vdots$	$\vdots$	$\vdots$
SS( $m$   $m - 1$ )	$r - r_{m-1}$	MS( $m$   $m - 1$ )
SSE	$n - r$	MSE = $\hat{\sigma}^2$
SST <sub>0</sub>	$n - 1$	

### Independence of ANOVA Sums of Squares

Because

$$(\mathbf{P}_{j+1} - \mathbf{P}_j)(\sigma^2 \mathbf{I})(\mathbf{P}_{\ell+1} - \mathbf{P}_\ell) = \mathbf{0} \quad \text{for } j \neq \ell,$$

any two ANOVA sums of squares (not including SST<sub>0</sub>) are independent.

It is also true that the ANOVA sums of squares (not including SST<sub>0</sub>) are *mutually independent* by Cochran's Theorem, although this stronger result is not usually needed.

### ANOVA F Statistics

For  $j = 1, \dots, m - 1$ , define the ANOVA  $F$  statistic

$$F_j = \frac{\text{MS}(j+1 | j)}{\text{MSE}} = \frac{\mathbf{y}^\top (\mathbf{P}_{j+1} - \mathbf{P}_j) \mathbf{y} / (r_{j+1} - r_j)}{\mathbf{y}^\top (\mathbf{I} - \mathbf{P}_\mathbf{X}) \mathbf{y} / (n - r)}.$$

Under the general linear model,

$$F_j \sim F_{r_{j+1} - r_j, n - r}(\delta_j),$$

where the non-centrality parameter is

$$\delta_j = \frac{\boldsymbol{\beta}^\top \mathbf{X}^\top (\mathbf{P}_{j+1} - \mathbf{P}_j) \mathbf{X} \boldsymbol{\beta}}{2\sigma^2}.$$

### Relationship with Reduced vs. Full Model $F$ Statistic

The sequential ANOVA  $F_j$  statistic can be written as

$$F_j = \frac{\mathbf{y}^\top (\mathbf{P}_{j+1} - \mathbf{P}_j) \mathbf{y} / (r_{j+1} - r_j)}{\mathbf{y}^\top (\mathbf{I} - \mathbf{P}_X) \mathbf{y} / (n - r)} = \frac{\text{MS}(j+1 | j)}{\text{MSE}}.$$

This matches the reduced vs. full model  $F$  statistic

$$F = \frac{\mathbf{y}^\top (\mathbf{P}_X - \mathbf{P}_{X_0}) \mathbf{y} / (r - r_0)}{\mathbf{y}^\top (\mathbf{I} - \mathbf{P}_X) \mathbf{y} / (n - r)}.$$

### What Do ANOVA $F$ -Statistics Test?

In general, an  $F$  statistic tests

$$H_0 : \text{the non-centrality parameter is } 0 \quad \text{vs.} \quad H_A : \text{the non-centrality parameter is not } 0.$$

For the sequential ANOVA statistic  $F_j$ , the non-centrality parameter is

$$\delta_j = \frac{\boldsymbol{\beta}^\top \mathbf{X}^\top (\mathbf{P}_{j+1} - \mathbf{P}_j) \mathbf{X} \boldsymbol{\beta}}{2\sigma^2}.$$

Thus,  $F_j$  can be used to test

$$H_{0j} : \boldsymbol{\beta}^\top \mathbf{X}^\top (\mathbf{P}_{j+1} - \mathbf{P}_j) \mathbf{X} \boldsymbol{\beta} = 0 \quad \text{vs.} \quad H_{Aj} : \boldsymbol{\beta}^\top \mathbf{X}^\top (\mathbf{P}_{j+1} - \mathbf{P}_j) \mathbf{X} \boldsymbol{\beta} \neq 0.$$

### What Do ANOVA $F$ -Statistics Test Pt. II?

**Estimability does not necessarily imply testability.**

In general, for the  $j$ -th sequential test we have:

$$H_{0j} : (P_{j+1} - P_j)X\boldsymbol{\beta} = 0 \quad \text{vs.} \quad H_{Aj} : (P_{j+1} - P_j)X\boldsymbol{\beta} \neq 0$$

which, in testable form, is:

$$H_{0j} : C_j \boldsymbol{\beta} = 0 \quad \text{vs.} \quad H_{Aj} : C_j \boldsymbol{\beta} \neq 0,$$

where  $C_j$  is any matrix whose  $q = r_{j+1} - r_j$  rows form a basis for the row space of  $(P_{j+1} - P_j)X$ .

### Hypothesis Testing Interpretation

Recall that  $C\boldsymbol{\beta}$  is estimable if and only if

$$C = A\mathbf{X}$$

for some matrix  $A$ .

The hypotheses above can be written as

$$H_{0j} : (\mathbf{P}_{j+1} - \mathbf{P}_j)\mathbf{X}\boldsymbol{\beta} = 0 \quad \text{vs.} \quad H_{Aj} : (\mathbf{P}_{j+1} - \mathbf{P}_j)\mathbf{X}\boldsymbol{\beta} \neq 0.$$

This is of the form

$$H_{0j} : C_j^* \boldsymbol{\beta} = 0 \quad \text{vs.} \quad H_{Aj} : C_j^* \boldsymbol{\beta} \neq 0,$$

where

$$C_j^* = (\mathbf{P}_{j+1} - \mathbf{P}_j)\mathbf{X}.$$

As written,  $H_{0j}$  is not directly testable because  $C_j^*$  has  $n$  rows but rank

$$\text{rank}(C_j^*) = r_{j+1} - r_j < n.$$

We can rewrite  $H_{0j}$  as a testable hypothesis by replacing  $C_j^*$  with any matrix  $C_j$  whose

$$q = r_{j+1} - r_j$$

rows form a basis for the row space of  $C_j^*$ .

### Example: Multiple Regression

In multiple linear regression, we consider a sequence of nested models:

$$X_1 = 1$$

$$X_2 = [1, x_1]$$

$$X_3 = [1, x_1, x_2]$$

$$\vdots$$

$$X_m = [1, x_1, \dots, x_{m-1}]$$

Here,  $SS(j+1|j)$  is the decrease in SSE that results when the explanatory variable  $x_j$  is added to a model containing an intercept and explanatory variables  $x_1, \dots, x_{j-1}$ .

## Example: Polynomial Regression

For polynomial regression, the sequence is:

$$X_1 = 1$$

$$X_2 = [1, x]$$

$$X_3 = [1, x, x^2]$$

$$\vdots$$

$$X_m = [1, x, x^2, \dots, x^{m-1}]$$

Here,  $SS(j+1|j)$  is the decrease in SSE that results when the term  $x^j$  is added to a model containing an intercept and the lower-order terms  $x, x^2, \dots, x^{j-1}$ .

## Aside: Centering and Standardizing for Numerical Stability

It is typically best for numerical stability to center and scale a quantitative explanatory variable prior to computing higher-order terms.

In the plant density example, we could replace  $x$  by

$$\frac{x - 30}{10}$$

and work with the transformed matrices.

Because these matrices have the same column spaces as the original matrices, the ANOVA table entries are **mathematically identical** for either set of matrices.

## ANOVA Balanced Two Factor

The following assumes a balanced design, i.e., every unique combination of factors and levels occur equally often.

*Always begin by writing out the model.*

For  $i = 1, 2$ ,  $j = 1, 2, 3$ , and  $k = 1, 2$ , let  $y_{ijk}$  denote the weight gain of the  $k$ th pig that received diet  $i$  and drug  $j$ , and suppose

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}, \quad (i = 1, 2; j = 1, 2, 3; k = 1, 2),$$

where

$$\mu, \alpha_1, \alpha_2, \beta_1, \beta_2, \beta_3, \gamma_{11}, \gamma_{12}, \gamma_{13}, \gamma_{21}, \gamma_{22}, \gamma_{23}$$

are unknown real-valued parameters, and

$$\varepsilon_{111}, \varepsilon_{112}, \varepsilon_{121}, \varepsilon_{122}, \varepsilon_{131}, \varepsilon_{132}, \varepsilon_{211}, \varepsilon_{212}, \varepsilon_{221}, \varepsilon_{222}, \varepsilon_{231}, \varepsilon_{232} \stackrel{\text{i.i.d.}}{\sim} N(0, \sigma^2),$$

for some unknown  $\sigma^2 > 0$ .

## A Sequence of Models for the Mean

We could consider a sequence of progressively more complex models for the response mean that lead up to our full cell-means model:

### 1. Intercept-only model

$$\mathbb{E}(y_{ijk}) = \mu$$

### 2. Diet main effects

$$\mathbb{E}(y_{ijk}) = \mu + \alpha_i$$

### 3. Diet and drug main effects

$$\mathbb{E}(y_{ijk}) = \mu + \alpha_i + \beta_j$$

### 4. Full model with interaction

$$\mathbb{E}(y_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_{ij} \iff \mathbb{E}(y_{ijk}) = \mu_{ij}$$

## ANOVA Table for Our Two-Factor Example

Sequential sums of squares measure how much the error sum of squares decreases as terms are added to the model.

Source	Sum of Squares	DF
Diets   1	$y^\top (P_2 - P_1)y$	$2 - 1 = 1$
Drugs   1, Diets	$y^\top (P_3 - P_2)y$	$4 - 2 = 2$
Diets $\times$ Drugs   1, Diets, Drugs	$y^\top (P_4 - P_3)y$	$6 - 4 = 2$
Error	$y^\top (I - P_4)y$	$12 - 6 = 6$
C. Total	$y^\top (I - P_1)y$	$12 - 1 = 11$

Sequential SS correspond to conditioning on factors that previously entered the model.

## What Do the F-Tests in This ANOVA Table Test?

Recall that the null hypothesis for the  $j$ th  $F$ -test is true if and only if

$$\beta^\top X^\top (P_{j+1} - P_j) X \beta = 0.$$

We have the following equivalent conditions:

$$\beta^\top X^\top (P_{j+1} - P_j) X \beta = 0 \iff \beta^\top X^\top (P_{j+1} - P_j)^\top (P_{j+1} - P_j) X \beta = 0$$

$$\iff \|(P_{j+1} - P_j) X \beta\|^2 = 0 \iff (P_{j+1} - P_j) X \beta = 0 \iff C \beta = 0,$$

where  $C$  is any full-row-rank matrix with the same row space as  $(P_{j+1} - P_j) X$ .

### Diet Test

$$(P_2 - P_1) X \beta = 0 \iff C \beta = 0,$$

where

$$C \beta = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & -\frac{1}{3} & -\frac{1}{3} & -\frac{1}{3} \end{bmatrix} \begin{bmatrix} \mu_{11} \\ \mu_{12} \\ \mu_{13} \\ \mu_{21} \\ \mu_{22} \\ \mu_{23} \end{bmatrix} = \bar{\mu}_{1\cdot} - \bar{\mu}_{2\cdot}.$$

### Drug Test

$$(P_3 - P_2) X \beta = 0 \iff C \beta = 0,$$

where

$$C \beta = \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} & 0 & \frac{1}{2} & -\frac{1}{2} & 0 \\ \frac{1}{2} & 0 & -\frac{1}{2} & \frac{1}{2} & 0 & -\frac{1}{2} \end{bmatrix} \begin{bmatrix} \mu_{11} \\ \mu_{12} \\ \mu_{13} \\ \mu_{21} \\ \mu_{22} \\ \mu_{23} \end{bmatrix} = \begin{bmatrix} \bar{\mu}_{\cdot 1} - \bar{\mu}_{\cdot 2} \\ \bar{\mu}_{\cdot 1} - \bar{\mu}_{\cdot 3} \end{bmatrix}.$$

### Diet-X-Drug Interaction Test

$$(P_4 - P_3) X \beta = 0 \iff C \beta = 0,$$

where

$$C \beta = \begin{bmatrix} 1 & -1 & 0 & -1 & 1 & 0 \\ 1 & 0 & -1 & -1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu_{11} \\ \mu_{12} \\ \mu_{13} \\ \mu_{21} \\ \mu_{22} \\ \mu_{23} \end{bmatrix} = \begin{bmatrix} \mu_{11} - \mu_{12} - \mu_{21} + \mu_{22} \\ \mu_{11} - \mu_{13} - \mu_{21} + \mu_{23} \end{bmatrix}.$$

## ANOVA for Balanced Two-Factor Experiments

The diet–drug experiment is **balanced** in the sense that every treatment (defined by a diet–drug combination) has the same number of experimental units.

Each experimental unit provides a single response measurement (weight gain), so the resulting dataset is balanced in the sense that each treatment has the same number of independent, constant-variance observations.

Due to this balance, the tests for diets, drugs, and diets  $\times$  drugs in the ANOVA table are exactly the same as the tests for diet main effects, drug main effects, and diet  $\times$  drug interactions previously expressed as tests of  $C\beta = 0$ .

## ANOVA Unbalanced Two Factor

When data are unbalanced, the Type I ANOVA test for two-way interactions is the same as the test for two-way interactions discussed previously.

However, the Type I ANOVA tests for individual factors are not the tests for main effects discussed previously.

Furthermore, the Type I results for individual factors depend on the order that the factors appear in the Type I ANOVA table.

### Example

An experiment was conducted to study the effect of storage time and storage temperature on the amount of active ingredient in a drug lost during storage. A total of 16 vials of the drug, each containing approximately 30 mg/mL of active ingredient, were assigned (using a completely randomized design) to the following treatments:

1. Storage for 3 months at 20° C
2. Storage for 3 months at 30° C
3. Storage for 6 months at 20° C
4. Storage for 6 months at 30° C

#### *Data*

6 of the 16 vials were damaged during shipment to the lab where the active ingredient was measured. The amount of active ingredient was measured only for the 10 undamaged vials. The table below shows the amount of active ingredient lost during storage (in tenths of mg/mL) for each of the undamaged vials.

Storage Time	Storage Temperature 20° C	30° C
3 months	3, 5	11, 13, 15
6 months	5, 6, 6, 7	16

as temperature  $\uparrow$ ,  $y \uparrow$  as length  $\uparrow$ ,  $y \uparrow$



## A Cell Means Model for the Data

Let  $y_{ijk}$  denote the amount of active ingredient lost from the  $k$ th vial treated with the  $i$ th storage time and  $j$ th temperature.

Let  $n_{ij}$  denote the number of vials measured for the  $i$ th storage time and  $j$ th temperature.

Suppose

$$y_{ijk} = \mu_{ij} + \varepsilon_{ijk}, \quad (i = 1, 2; j = 1, 2; k = 1, \dots, n_{ij}),$$

where  $\mu_{11}, \mu_{12}, \mu_{21}, \mu_{22}$  are unknown real-valued parameters and the  $\varepsilon_{ijk}$  terms are i.i.d. normal random variables with mean 0 and some unknown variance  $\sigma^2 > 0$ .

We could consider a sequence of progressively more complex models for the response mean that lead up to our full cell-means model:

### 1. Intercept-only model

$$\mathbb{E}(y_{ijk}) = \mu$$

### 2. Add storage time

$$\mathbb{E}(y_{ijk}) = \mu + \alpha_i$$

### 3. Add storage time and temperature

$$\mathbb{E}(y_{ijk}) = \mu + \alpha_i + \beta_j$$

### 4. Full model with interaction

$$\mathbb{E}(y_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_{ij} \iff \mathbb{E}(y_{ijk}) = \mu_{ij}$$

## ANOVA Table

Source	Sum of Squares	DF
Time   1	$y^\top (P_2 - P_1)y$	$2 - 1 = 1$
Temp   1, Time	$y^\top (P_3 - P_2)y$	$3 - 2 = 1$
Time $\times$ Temp   1, Time, Temp	$y^\top (P_4 - P_3)y$	$4 - 3 = 1$
Error	$y^\top (I - P_4)y$	$10 - 4 = 6$
C. Total	$y^\top (I - P_1)y$	$10 - 1 = 9$

### What Do the $F$ -Tests in This ANOVA Table Test?

Recall the null hypothesis for  $F_j$  is true if and only if

$$\beta^\top X^\top (P_{j+1} - P_j) X \beta = 0.$$

We have the following equivalent conditions:

$$\beta^\top X^\top (P_{j+1} - P_j) X \beta = 0 \iff \beta^\top X^\top (P_{j+1} - P_j)^\top (P_{j+1} - P_j) X \beta = 0$$

$$\iff \|(P_{j+1} - P_j) X \beta\|^2 = 0 \iff (P_{j+1} - P_j) X \beta = 0 \iff C \beta = 0,$$

where  $C$  is any full-row-rank matrix with the same row space as  $(P_{j+1} - P_j) X$ .

### Interpreting $(P_{j+1} - P_j) X$

Let's take a look at  $(P_{j+1} - P_j) X$  for each test in the ANOVA table.

When computing  $(P_{j+1} - P_j) X$ , we can use any model matrix  $X$  that specifies one unrestricted treatment mean for each of the four treatments.

The entries in any rows of  $(P_{j+1} - P_j) X$  are coefficients defining linear combinations of the elements of the parameter vector  $\beta$  corresponding to the chosen model matrix  $X$ .

### Time | 1 ANOVA Test

$$(P_2 - P_1) X \beta = 0 \iff C \beta = 0,$$

where

$$\begin{aligned} C \beta &= \begin{bmatrix} \frac{2}{5} & \frac{3}{5} & -\frac{4}{5} & -\frac{1}{5} \end{bmatrix} \begin{bmatrix} \mu_{11} \\ \mu_{12} \\ \mu_{21} \\ \mu_{22} \end{bmatrix} \\ &= \left( \frac{2}{5} \mu_{11} + \frac{3}{5} \mu_{12} \right) - \left( \frac{4}{5} \mu_{21} + \frac{1}{5} \mu_{22} \right). \end{aligned}$$

This weighted average reflects the sample sizes in the cells.

### Time | 1 ANOVA Test $\neq$ Time Main Effect Test

Null for Time | 1 ANOVA test:

$$\frac{2}{5} \mu_{11} + \frac{3}{5} \mu_{12} = \frac{4}{5} \mu_{21} + \frac{1}{5} \mu_{22}.$$

Null for Time main effect test:

$$\frac{1}{2} \mu_{11} + \frac{1}{2} \mu_{12} = \frac{1}{2} \mu_{21} + \frac{1}{2} \mu_{22},$$

i.e.,

$$\bar{\mu}_{1\cdot} = \bar{\mu}_{2\cdot}$$

## Two Factors: Different Types of Sums of Squares

Source	Type I	Type II	Type III
<i>A</i>	$SS(A \mid 1)$	$SS(A \mid 1, B)$	$SS(A \mid 1, B, AB)$
<i>B</i>	$SS(B \mid 1, A)$	$SS(B \mid 1, A)$	$SS(B \mid 1, A, AB)$
<i>AB</i>	$SS(AB \mid 1, A, B)$	$SS(AB \mid 1, A, B)$	$SS(AB \mid 1, A, B)$
Error	$SSE$	$SSE$	$SSE$
C. Total	$SS_{Total}$	?	?

- **Type I (Sequential):** reduction in  $SSE$  due to a factor, given the terms that entered the model previously.
- **Type II:** accounts for all terms that do *not* involve the factor under consideration.
- **Type III:** reduction in  $SSE$  due to a factor *given all other terms in the model*.

## Three Factors: Different Types of Sums of Squares

Source	Type I	Type II	Type III
<i>A</i>	$SS(A \mid 1)$	$SS(A \mid 1, B, C, BC)$	$SS(A \mid 1, B, C, AB, AC, BC, ABC)$
<i>B</i>	$SS(B \mid 1, A)$	$SS(B \mid 1, A, C, AC)$	$SS(B \mid 1, A, C, AB, AC, BC, ABC)$
<i>C</i>	$SS(C \mid 1, A, B)$	$SS(C \mid 1, A, B, AB)$	$SS(C \mid 1, A, B, AB, AC, BC, ABC)$
<i>AB</i>	$SS(AB \mid 1, A, B, C)$	$SS(AB \mid 1, A, B, C, AC, BC)$	$SS(AB \mid 1, A, B, C, AC, BC, ABC)$
<i>AC</i>	$SS(AC \mid 1, A, B, C, AB)$	$SS(AC \mid 1, A, B, C, AB, BC)$	$SS(AC \mid 1, A, B, C, AB, BC, ABC)$
<i>BC</i>	$SS(BC \mid 1, A, B, C, AB, AC)$	$SS(BC \mid 1, A, B, C, AB, AC)$	$SS(BC \mid 1, A, B, C, AB, AC, ABC)$
<i>ABC</i>	$SS(ABC \mid 1, A, B, C, AB, AC, BC)$	$SS(ABC \mid 1, A, B, C, AB, AC, BC)$	$SS(ABC \mid 1, A, B, C, AB, AC, BC)$
Error	$SSE$	$SSE$	$SSE$

Notes:

- No interaction involving the factor under consideration is included.
- Unlike Type III sums of squares, we do **not** account for the  $ABC$  interaction when testing lower-order terms.

## Sums of Squares for Balanced Data

For **balanced data**, the three types of sums of squares are identical:

$$\text{Type I} = \text{Type II} = \text{Type III}.$$

This equality is not obvious (at least to most normal humans), but it is true. We will not attempt to prove this in 510.

The ANOVA  $F$ -tests in the ANOVA table can be used to test for factor main effects and interactions.

## Sums of Squares for Unbalanced Data

For **unbalanced data**, the types of sums of squares differ.

- Type I sums of squares always add to the total sum of squares, even when data are unbalanced.
- Type II and Type III sums of squares do not add to anything special when data are unbalanced.
- The ANOVA  $F$ -tests in the **Type III** ANOVA table can be used to test for factor main effects and interactions.
- Type I and Type II ANOVA  $F$ -tests do **not**, in general, test for factor main effects or interactions (except for the  $F$ -test for the highest-order interaction, which is the same for all three types).

## Type IV Sums of Squares

In addition to computing Type I, II, and III sums of squares, SAS can compute Type IV sums of squares.

Type IV sums of squares are only relevant for factorial designs with missing cells.

When cells are missing, I recommend determining the linear combinations of the estimable cell means that are of scientific interest, and then conducting the corresponding tests as tests of  $H_0 : C\beta = d$ .

## Calculation of Type I, II, and III Sums of Squares

Every Type I, II, or III sum of squares is the error sums of squares for a reduced model minus the error sum of squares for a model that adds one term to the reduced model;

$$y^T(I - P_{X_{\text{reduced}}})y - y^T(I - P_{X_{\text{reduced}} + \text{term}})y = y^T(P_{X_{\text{reduced}} + \text{term}} - P_{X_{\text{reduced}}})y$$

where  $C(X_{\text{reduced}}) \subset C(X_{\text{reduced}} + \text{term}) \subseteq C(X)$ .

As usual,  $X$  represents the model matrix for the most complex model under consideration (a.k.a., the full model).

For all Type III sums of squares, the reduced+term model is the full model.

## Alternative Computation of Sums of Squares

Let  $SS = y^T(P_{X_{\text{reduced}} + \text{term}} - P_{X_{\text{reduced}}})y$  represent any Type I, II, or III sum of squares.

Let  $q = \text{rank}(X_{\text{reduced}} + \text{term}) - \text{rank}(X_{\text{reduced}})$  be the degrees of freedom associated with  $SS$ .

Let  $C$  be any  $q \times p$  matrix whose rows are a basis for the row space of  $(P_{X_{\text{reduced}} + \text{term}} - P_{X_{\text{reduced}}})X$ .

$$y^T(P_{j+1} - P_j)y = (C\hat{\beta})^T \{C(X^T X)^{-1}C^T\}^{-1}C\hat{\beta}$$

Then the ANOVA  $F$  statistic

$$\frac{SS/q}{MSE} = \frac{\hat{\beta}^T C^T [C(X^T X)^{-1}C^T]^{-1}C\hat{\beta}/q}{\hat{\sigma}^2}.$$

Thus, any  $SS$  can be computed as  $\hat{\beta}^T C^T [C(X^T X)^{-1}C^T]^{-1}C\hat{\beta}$  for an appropriate matrix  $C$ .

# Orthogonal Contrasts

## Introduction

Orthogonal contrasts are:

- designed to be independent of one another (within the same model)
- useful because they allow testing of multiple hypotheses simultaneously without inflating the probability of a Type I error.
- constructed such that they do not overlap in terms of the information they provide about the data.

## Orthogonal Linear Combinations

Under the model

$$y = X\beta + \epsilon, \epsilon \sim N(0, \sigma^2 I),$$

two estimable linear combinations  $c_1^T \beta$  and  $c_2^T \beta$  are orthogonal if and only if their best linear unbiased estimators  $c_1^T \hat{\beta}$  and  $c_2^T \hat{\beta}$  are uncorrelated.

**Blue's Theorem:**

Recall  $c_k^T \beta$  is estimable if and only if there exists  $a_k$  such that

$$c_k^T = a_k^T X.$$

$$\text{Cov}(c_1^T \hat{\beta}, c_2^T \hat{\beta})$$

$$= \text{Cov}(a_1^T X \hat{\beta}, a_2^T X \hat{\beta}) = \text{Cov}(a_1^T P_X y, a_2^T P_X y)$$

$$= a_1^T P_X \text{Cov}(y, y) P_X^T a_2 = a_1^T P_X \text{Var}(y) P_X^T a_2$$

$$= a_1^T P_X (\sigma^2 I) P_X^T a_2 = \sigma^2 a_1^T P_X P_X^T a_2$$

$$= \sigma^2 a_1^T P_X a_2 = \sigma^2 a_1^T X (X^T X)^{-} X^T a_2 = \sigma^2 c_1^T (X^T X)^{-} c_2.$$

Thus, estimable linear combinations  $c_1^T \beta$  and  $c_2^T \beta$  are orthogonal if and only if  $c_1^T (X^T X)^{-} c_2 = 0$ .

## Orthogonal Contrasts Definition

A linear combination  $c^T \beta$  is a contrast if and only if  $c^T 1 = 0$ .

Two estimable contrasts  $c_1^T \beta$  and  $c_2^T \beta$  that are orthogonal are called **orthogonal contrasts**.

That is,

1.  $c_1^T \beta$  and  $c_2^T \beta$  are orthogonal:

2.

$$\text{Cov}(c_1^T \hat{\beta}, c_2^T \hat{\beta}) = 0$$

3. Contrast coefficients add to zero:  $c_1^T 1 = c_2^T 1 = 0$ .

## Connection to the ANOVA Table

Source	Sum of Squares	DF
Diets	$y^T (P_2 - P_1) y$	$2 - 1 = 1$
Drugs	$y^T (P_3 - P_2) y$	$4 - 2 = 2$
Diets $\times$ Drugs	$y^T (P_4 - P_3) y$	$6 - 4 = 2$
Error	$y^T (I - P_4) y$	$12 - 6 = 6$
C. Total	$y^T (I - P_1) y$	$12 - 1 = 11$

## Connection to Orthogonal Contrasts

**Drug main effect**

**Diet main effect**

SS Formulas:

$$y^T (P_2 - P_1) y = \frac{(c_1^T \hat{\beta})^2}{c_1^T (X^T X)^{-1} c_1}$$

$$y^T (P_3 - P_2) y = \frac{(c_2^T \hat{\beta})^2}{c_2^T (X^T X)^{-1} c_2} + \frac{(c_3^T \hat{\beta})^2}{c_3^T (X^T X)^{-1} c_3}$$

$$y^T (P_4 - P_3) y = \frac{(c_4^T \hat{\beta})^2}{c_4^T (X^T X)^{-1} c_4} + \frac{(c_5^T \hat{\beta})^2}{c_5^T (X^T X)^{-1} c_5}$$

$$y^T (I - P_4) y$$

$$y^T (I - P_1) y$$

Total DF = 11

## Thoughts and Commentary

### Additional Partitioning of ANOVA Sums of Squares

The previous example shows how the Drug and Diet  $\times$  Drug sums of squares can each be partitioned into two single-degree of freedom sums of squares corresponding to estimable orthogonal contrasts.

More generally, any ANOVA sum of squares with  $q$  degrees of freedom can be partitioned into  $q$  single-degree-of-freedom sums of squares corresponding to  $q$  estimable orthogonal linear combinations  $c_1^T \beta, \dots, c_q^T \beta$ .

## ANOVA Partitioning is Not Always Necessary

Just because we can partition ANOVA sums of squares does not mean we need to partition ANOVA sums of squares.

The goals of an analysis typically involve constructing estimates or conducting tests of scientific interest.

The tests of scientific interest do not necessarily involve orthogonal linear combinations.

For example, suppose the goal of the researchers who conducted the diet-drug study is to determine which of the three drugs is best for enhancing weight gain of pigs on each diet.

## Comments on the Analysis

Note that the main analysis focuses on pairwise comparisons of drugs within each diet.

This involves a set of six contrasts, but the contrasts are not pairwise orthogonal within either diet.

The sums of squares for these contrasts do not add up to any ANOVA sums of squares, but they are the contrasts that best address the researchers' questions.

If we want to control the probability of one or more type  $I$  errors, we could use Bonferroni's method. In this case, the adjustment for multiple testing would not change the conclusions.

## Cell Means vs. Additive Model

We used the cell means model for analysis even though the interactions were not significant at the 0.05 level.

I tend to prefer the cell means model in experiments with a full-factorial treatment design even if interactions are not significant.

The cell means model is less restrictive than an additive model.

The cell means model estimator of error variance  $\sigma^2$  is not inflated by incorrectly specifying an additive mean structure when the additive mean structure is too restrictive.

Using the cell means model honors the treatment structure.

Using the cell means model avoids problems with using the data once to select a model and a second time to perform inference.

Some other statisticians may favor a different strategy, especially in experiments with many factors or few degrees of freedom for error.

## Brief Aside on Some Relevant Linear Algebra

### Orthogonal and Orthonormal Vectors

The  $m \times 1$  vectors  $p_1, \dots, p_n$  are said to be *orthogonal* if and only if

$$p_i^\top p_j = 0 \text{ for all } i \neq j.$$

The  $m \times 1$  vectors  $p_1, \dots, p_n$  are said to be *orthonormal* if and only if

$$p_i^\top p_j = \begin{cases} 0 & \text{if } i \neq j \\ 1 & \text{if } i = j \end{cases}$$

## Orthogonal Matrices

A square matrix  $P$  is said to be *orthogonal* if and only if

$$P^\top P = I.$$

Note that because  $P$  is square,

$$P^\top P = I$$

implies that

$$(P^\top)^{-1} = P$$

and

$$P^{-1} = P^\top.$$

Thus,

$$P^\top P = PP^\top = I.$$

It follows that a square matrix  $P$  is orthogonal if and only if the rows of  $P$  are orthonormal vectors and the columns of  $P$  are orthonormal vectors.

## The Spectral Decomposition Theorem

An  $n \times n$  symmetric matrix  $H$  may be decomposed as

$$H = P\Lambda P^T = \sum_{i=1}^n \lambda_i p_i p_i^T,$$

where

- $P = [p_1, \dots, p_n]$  is an  $n \times n$  orthogonal matrix whose columns  $p_1, \dots, p_n$  are the orthonormal eigenvectors of  $H$ , and
- $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_n)$  is a diagonal matrix whose diagonal entries  $\lambda_1, \dots, \lambda_n \in \mathbb{R}$  are the eigenvalues of  $H$  (with  $\lambda_i$  corresponding to  $p_i$  for  $i = 1, \dots, n$ ).

## Aitken Model

An alternative decomposition is based on the Cholesky decomposition.

Let

$$y = X\beta + \varepsilon, \quad E(\varepsilon) = 0, \quad \text{Var}(\varepsilon) = \sigma^2 V.$$



This model is identical to the Gauss–Markov linear model except that

$$\text{Var}(\varepsilon) = \sigma^2 V \quad \text{instead of} \quad \sigma^2 I.$$

- $V$  is assumed to be a known positive definite variance matrix.
- $\sigma^2$  is an unknown positive variance parameter.

We need a transformation of our model that results in a new model fulfilling the Gauss–Markov assumptions.

## A Transformation of the Aitken Model

Let  $V^{1/2}$  be the symmetric square root of  $V$ .

Note that  $V$  positive definite implies  $V^{1/2}$  is positive definite and therefore nonsingular.

Using  $V^{-1/2}$  to denote  $(V^{1/2})^{-1}$ , we have

$$V^{-1/2}y = V^{-1/2}X\beta + V^{-1/2}\varepsilon.$$

Define

$$z = V^{-1/2}y, \quad W = V^{-1/2}X, \quad \delta = V^{-1/2}\varepsilon.$$

Then

$$z = W\beta + \delta, \quad E(\delta) = 0, \quad \text{Var}(\delta) = \sigma^2 I,$$

because

$$\begin{aligned} \text{Var}(\delta) &= \text{Var}(V^{-1/2}\varepsilon) \\ &= V^{-1/2} \text{Var}(\varepsilon) V^{-1/2} \\ &= V^{-1/2}(\sigma^2 V) V^{-1/2} \\ &= \sigma^2 I. \end{aligned}$$

Thus, after transformation, we are back to the Gauss–Markov model we are familiar with.

We can apply all the results we have established previously for the Gauss–Markov model.

## Estimation of $E(y)$ under the Aitken Model

Note that

$$E(y) = E(V^{1/2}V^{-1/2}y) = V^{1/2}E(V^{-1/2}y) = V^{1/2}E(z).$$

Because the Gauss–Markov model holds for  $z$ , the best estimator of  $E(z)$  is

$$\hat{z} = P_W z = W(W^\top W)^{-1}W^\top z,$$

where  $W = V^{-1/2}X$ .

Substituting,

$$\begin{aligned}\hat{z} &= V^{-1/2}X((V^{-1/2}X)^\top(V^{-1/2}X))^{-1}(V^{-1/2}X)^\top V^{-1/2}y \\ &= V^{-1/2}X(X^\top V^{-1}X)^{-1}X^\top V^{-1}y.\end{aligned}$$

Thus, to estimate

$$E(y) = V^{1/2}E(z),$$

we use

$$\hat{y} = V^{1/2}\hat{z} = X(X^\top V^{-1}X)^{-1}X^\top V^{-1}y.$$

## Estimation of Linear Functions under the Aitken Model

Likewise, if  $C\beta$  is estimable, the BLUE is the ordinary least squares estimator

$$C(W^\top W)^{-1}W^\top z,$$

which can be expressed as

$$\begin{aligned}C(W^\top W)^{-1}W^\top z &= C(X^\top V^{-1/2}V^{-1/2}X)^{-1}X^\top V^{-1/2}V^{-1/2}y \\ &= C(X^\top V^{-1}X)^{-1}X^\top V^{-1}y.\end{aligned}$$

The estimator

$$C(X^\top V^{-1}X)^{-1}X^\top V^{-1}y = C\hat{\beta}_V$$

is called a **Generalized Least Squares (GLS)** estimator.

This estimator is the BLUE of any estimable  $C\beta$  under the Aitken Model.

## Aitken Equations

The GLS estimator

$$\hat{\beta}_V = (X^\top V^{-1}X)^{-1}X^\top V^{-1}y$$

is a solution to the Aitken equations:

$$X^\top V^{-1}Xb = X^\top V^{-1}y.$$

These follow from the normal equations

$$W^\top Wb = W^\top z,$$

since

$$\begin{aligned}
W^\top Wb &= W^\top z \\
\iff X^\top V^{-1/2} V^{-1/2} Xb &= X^\top V^{-1/2} V^{-1/2} y \\
\iff X^\top V^{-1} Xb &= X^\top V^{-1} y.
\end{aligned}$$

Thus,

$$\hat{\beta}_V = (X^\top V^{-1} X)^{-1} X^\top V^{-1} y$$

is a solution to the generalized least squares problem.

### Weighted Least Squares

When  $V$  is diagonal, the term **Weighted Least Squares (WLS)** is often used instead of GLS.

If

$$V = \text{diag}(v_{11}, \dots, v_{nn}),$$

the least squares problem becomes

Find  $b$  to minimize

$$(y - Xb)^\top V^{-1} (y - Xb) = \sum_{i=1}^n \frac{1}{v_{ii}} (y_i - x_{(i)}^\top b)^2,$$

where  $x_{(i)}^\top$  is the  $i$ th row of  $X$ .

### Inference Under the Aitken Model with Normal Errors

- The Aitken Model with Normal errors:

$$y = X\beta + \epsilon, \epsilon \sim \mathcal{N}(0, \sigma^2 V).$$

- Under the Aitken Model with Normal errors, we can back transform to convert known formulas in terms of  $z$  and  $W$  to formulas in terms of  $y$  and  $X$  to allow inference about estimable  $C\beta$  under the Aitken Model with Normal errors.

### An Example

Researchers were interested in comparing the dry weight of maize seedlings from two different genotypes. For each genotype, nine seeds were planted in each of four trays. The eight trays in total were randomly positioned in a growth chamber. Three weeks after the emergence of the first seedling, emerged seedlings were harvested from each tray and weighed together after drying to obtain one weight for each tray. Although nine seeds were planted in each tray, fewer than nine seedlings emerged in many of the trays. Thus, weights were recorded on a per seedling basis, and the number of seedlings that emerged in each tray was also recorded.

*A Model for the Data*

Let  $n_{ij}$  be the number of seedlings for the  $j$ th tray of genotype  $i$  ( $i = 1, 2; j = 1, 2, 3, 4$ ).

Let  $y_{ijk}$  be the dry weight of the  $k$ th seedling in the  $j$ th tray of genotype  $i$  ( $i = 1, 2; j = 1, 2, 3, 4; k = 1, \dots, n_{ij}$ ).

Suppose all seedling weights are independent and normally distributed with common variance  $\sigma^2$  and genotype-specific means,  $\mu_1$  for genotype 1 and  $\mu_2$  for genotype 2:

$$y_{ijk} \stackrel{ind}{\sim} N(\mu_i, \sigma^2).$$

Now let  $y_{ij} = \bar{y}_{ij}$

$$y_{ij} = \bar{y}_{ij}$$

It follows that

$$y_{ij} \sim N(\mu_i, \sigma^2/n_{ij})$$

or, equivalently,

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

where

$$\epsilon_{ij} \sim N(0, \sigma^2/n_{ij}).$$

### Model in Matrix and Vector Form

$$\begin{bmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{14} \\ y_{21} \\ y_{22} \\ y_{23} \\ y_{24} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} + \begin{bmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{13} \\ \epsilon_{14} \\ \epsilon_{21} \\ \epsilon_{22} \\ \epsilon_{23} \\ \epsilon_{24} \end{bmatrix}$$

$$y = X\beta + \epsilon, \quad \epsilon \sim N(0, \sigma^2 V)$$

## Linear Mixed Effects

### Motivation

In a linear model we distinguish between two types of effects:

**fixed effects** vs. **random effects**

Which effect to choose depends on

1. the context of the data,
2. the research questions of interest, and
3. how the data are collected.

## Fixed Effects

Data have been gathered from all the levels of the factor that are of interest. Number of levels is typically small.

## Random Effects

Factor variable has many possible levels and levels typically represent a larger population of interest. However, observing all levels is not feasible and we only have a random sample of levels in the data.

Examples:

1. Assessing effectiveness of a new curriculum after statewide implementation: school effect
2. Operator or machine effect in Gauge R&R studies
3. Hospital effect

We are interested in whether the factor has a significant effect in explaining the response, but only in a general way.

Some general remarks:

- Data analysis differs depending on type of effect (fixed or random); hence misspecification of the type of effect can lead to incorrect conclusions.
- Random factor analysis is usually used if there is reason to believe that the levels observed in the experiment could reasonably be a random sample of all levels.
- An interaction term involving both a fixed and a random factor should be considered a random factor.
- A factor that is nested in a random factor should be considered random.

If a model contains both fixed and random effects, we call it a **mixed effects model**.

## The Linear Mixed-Effects Model

$$y = X\beta + Zu + e$$

- $X$  is an  $n \times p$  matrix of known constants
- $\beta \in \mathbb{R}^p$  is an unknown parameter vector
- $Z$  is an  $n \times q$  matrix of known constants
- $u$  is a  $q \times 1$  random vector — We model its Variance
- $e$  is an  $n \times 1$  vector of random errors

Also:

- The elements of  $\beta$  are considered to be non-random and are called *fixed effects*.
- The elements of  $u$  are random variables and are called *random effects*.
- The elements of the error vector  $e$  are always considered to be random variables.

Because the model includes both fixed and random effects (in addition to the random errors), it is called a *mixed-effects* model, or more simply, a *mixed model*.

The model is called a *linear mixed-effects* model because (as we will soon see)

$$E(y | u) = X\beta + Zu,$$

which is a linear function of fixed and random effects.

## Model Assumptions

We assume that

$$E(e) = 0, \quad \text{Var}(e) = R,$$

and

$$E(u) = 0, \quad \text{Var}(u) = G,$$

with

$$\text{Cov}(e, u) = 0.$$

The random effects do not affect the mean structure.

## Mean and Variance of $y$

The marginal mean of  $y$  is

$$\begin{aligned} E(y) &= E(X\beta + Zu + e) \\ &= X\beta + ZE(u) + E(e) \\ &= X\beta. \end{aligned}$$

The marginal variance of  $y$  is

$$\begin{aligned} \text{Var}(y) &= \text{Var}(X\beta + Zu + e) \\ &= \text{Var}(Zu + e) \\ &= \text{Var}(Zu) + \text{Var}(e) \\ &= Z \text{Var}(u) Z^\top + R \\ &= ZGZ^\top + R \equiv \Sigma. \end{aligned}$$

## Normality Assumption

We usually consider the special case in which

$$\begin{bmatrix} u \\ e \end{bmatrix} \sim \mathcal{N} \left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} G & 0 \\ 0 & R \end{bmatrix} \right).$$

This implies

$$y \sim \mathcal{N}(X\beta, ZGZ^\top + R).$$

The conditional mean and variance, given the random effects,

$$E(y \mid u) = X\beta + Zu, \quad \text{Var}(y \mid u) = R.$$

## Example

Suppose an experiment was conducted to compare the height of plants grown at two soil moisture levels (labeled 1 and 2).

The soil moisture levels were randomly assigned to 4 pots, with 2 pots per moisture level. For each moisture level, 3 seeds were planted in one pot and 2 seeds were planted in the other.

After a four-week growing period, the height of each seedling was measured.

Let  $y_{ijk}$  denote the height for soil moisture level  $i$ , pot  $j$ , and seedling  $k$ .

### Covariance Structure

Note that

$$\text{Var}(y_{ijk}) = \sigma_p^2 + \sigma_e^2, \quad \forall i, j, k.$$

For plants within the same pot,

$$\text{Cov}(y_{ijk}, y_{ijk^*}) = \sigma_p^2, \quad \forall i, j, \text{ and } k \neq k^*.$$

For plants from different pots,

$$\text{Cov}(y_{ijk}, y_{i^*j^*k^*}) = 0, \quad \text{if } i \neq i^* \text{ or } j \neq j^*.$$

Thus:

- Any two observations from the same pot have covariance  $\sigma_p^2$ .
- Any two observations from different pots are uncorrelated.

### Estimation with Known Variance Components

If  $\sigma_p^2/\sigma_e^2$  were known, we would use GLS to estimate any estimable  $C\beta$  by

$$C\hat{\beta}_V = C(X^\top V^{-1}X)^{-1}X^\top V^{-1}y.$$

However, we seldom know  $\sigma_p^2/\sigma_e^2$  or, more generally,  $\Sigma$  or  $V$ .

### Estimation with Unknown Variance Components

For the general problem where

$$\text{Var}(y) = \Sigma$$

is an unknown positive definite matrix, we can rewrite

$$\Sigma = \sigma^2 V,$$

where  $\sigma^2$  is an unknown positive variance and  $V$  is an unknown positive definite matrix.

As in our simple example, each entry of  $V$  is usually assumed to be a known function of a small number of unknown parameters.

Thus, our strategy for estimating an estimable  $C\beta$  involves estimating the unknown parameters in  $V$  to obtain

$$C\hat{\beta}_{\hat{V}} = C(X^\top \hat{V}^{-1}X)^{-1}X^\top \hat{V}^{-1}y.$$

In general,

$$C\hat{\beta}_{\hat{V}} = C(X^\top \hat{V}^{-1}X)^{-1}X^\top \hat{V}^{-1}y$$

is a nonlinear estimator that is an approximation to

$$C\hat{\beta}_V = C(X^\top V^{-1}X)^{-1}X^\top V^{-1}y,$$

which would be the BLUE of  $C\beta$  if  $V$  were known.

## Experimental Design Terminology

**Experiment** – An investigation in which the investigator applies some treatments to experimental units and then observes the effect of the treatments on the experimental units by measuring one or more response variables.

**Treatment** – A condition or set of conditions applied to experimental units in an experiment.

**Experimental Unit** – The physical entity to which a treatment is randomly assigned and independently applied.

**Response Variable** – A characteristic of an experimental unit that is measured after treatment and analyzed to assess the effects of treatments on experimental units.

**Observational Unit** – The unit on which a response variable is measured.

There is often a one-to-one correspondence between experimental units and observational units, but that is not always true.

### Example: Plant Heights and Soil Moisture

In our example involving plant heights and soil moisture levels, pots were the experimental units because soil moisture levels were randomly assigned to pots.

Seedlings were the observational units because the response was measured separately for each seedling.

Whenever there is more than one observational unit for an experimental unit or whenever the response is measured multiple times for an experimental unit, we say we have **multiple observations per experimental unit**.

This scenario is also referred to as **subsampling** or **pseudo-replication**.



## Importance of Random Effects for Multiple Observations

Whenever an experiment involves multiple observations per experimental unit, it is important to include a random effect for each experimental unit.

Without a random effect for each experimental unit, a one-to-one correspondence between observations and experimental units is assumed.

Including random effects in a model is one way to account for a lack of independence among observations that might be expected based on the design of an experiment.

## Experimental Design Types

**Completely Randomized Design (CRD)** – Experimental design in which, for given number of experimental units per treatment, all possible assignments of treatments to experimental units are equally likely.

**Block** – A group of experimental units that, prior to treatment, are expected to be more like one another (with respect to one or more response variables) than experimental units in general.

**Randomized Complete Block Design (RCBD)** – Experimental design in which separate and completely randomized treatment assignments are made for each of multiple blocks in such a way that all treatments have at least one experimental unit in each block.

## ANOVA Mixed

## Cochran Satterthwaite

## Split Plots