Document: Statistics 232A (Fall 2011)
Professor: Beran
Latest Update: April 2, 2012
Author: Jeff Irion
http://www.math.ucdavis.edu/~jlirion

Contents

1 9-22-11

1.1 Organizational Info

- TA's Office Hours: Mondays 11-12, Wednesdays 10-11 at MSB 1117
- Office hours: Tues & Thurs 9-10 in MSB 4224
- First discussion section is on 10/4.
- No text (because nobody has written one).
- Midterm is theory.
- 7 labs ⇒ combination of theory and practice.
- Course website: `http://www.stat.ucdavis.edu/~beran/s232a.html`

1.2 Singular Value Decomposition and Moore-Penrose Pseudoinverse

Theorem 1.1. ***SVD in reduced form***

Let $\mathbf{A}$ be an $m \times n$ matrix of rank $r \leq \min(m, n)$. Then

$$\mathbf{A} = \underbrace{\mathbf{U}}_{m\times r}\underbrace{\mathbf{L}}_{r\times r}\underbrace{\mathbf{V}'}_{r\times n}$$

where $U'U = V'V = I_r$ and $L = \text{diag}\{l_i\},\ l_1 \geq l_2 \geq \ldots l_r > 0$.

Remark 1.2. ***Remark 1 about SVD***

If $U = (u_1\ u_2\ \ldots\ u_r)$, where the u_i are $m \times 1$, and $V = (v_1\ v_2\ \ldots\ v_r)$ where the v_i are $n \times 1$,

$$A = \sum_{i=1}^{r} l_i \underbrace{u_i}_{m\times 1}\underbrace{v_i'}_{1\times n}$$

This says that the SVD is not a unique representation.

Remark 1.3. ***Remark 2 about SVD***

U, V are not unique.

Remark 1.4. ***Remark 3 about SVD***

There are numerically stable algorithms for the SVD. See Matrix Computations by Golub and Van Loan.

Remark 1.5. ***Remark 4 about SVD***

We have software implementations of these stable algorithms in R, Matlab, etc.

Definition 1.6. ***Generalized Inverse***

Let A be any $m \times n$ matrix. A *generalized inverse* of A is any matrix A^- such that $AA^-A = A$.

Remark 1.7. ***Remarks about the generalized inverse***

1. A may have many generalized inverses.
2. If A is square and of full rank (i.e. invertible), then A^{-1} is the unique generalized inverse.

Theorem 1.8. ***Moore-Penrose Pseudoinverse***

Let A be any $m \times n$ matrix. Then there exists a unique matrix A^+ $(n \times m)$ satisfying the following four properties:

1. $AA^+A = A$ (generalized inverse property)
2. $A^+AA^+ = A^+$ (mirror of the generalized inverse property)
3. A^+A is symmetric
4. AA^+ is symmetric

Proof. Existence of A^+:
SVD of A:

$$A = \underbrace{U}_{m\times r}\underbrace{L}_{r\times r}\underbrace{V'}_{r\times n}.$$

Define

$$A^+ = \underbrace{V}_{n\times r}\underbrace{L^{-1}}_{r\times r}\underbrace{U'}_{r\times m}.$$

Then

$$\begin{aligned} AA^+A &= ULV' \cdot VL^{-1}U' \cdot ULV' = ULV' = A \\ A^+AA^+ &= VL^{-1}U' \cdot ULV' \cdot VL^{-1}U' = VL^{-1}U' = A^+ \\ A^+A &= VL^{-1}U' \cdot ULV' = VV' \\ AA^+ &= ULV' \cdot VL^{-1}U' = UU' \end{aligned}$$

Uniqueness:
Let B be any $m \times n$ matrix with the Moore-Penrose pseudoinverse properties. Thus, we have $ABA =$

$A,\ BAB = B,\ BA$ and AB are symmetric. We want to show $B = A^+$.

$$\begin{aligned}
A^+A &= A^+(ABA) = \underbrace{(A^+A)}_{\text{symm}}\underbrace{(BA)}_{\text{symm}} = A'(A^+)' \cdot A'B' \\
&= (AA^+A)'B' = A'B' = BA \\
AA^+ &= (ABA)A^+ = \underbrace{(AB)}_{\text{symm}}\underbrace{(AA^+)}_{\text{symm}} = B'A'(A^+)'A' = B'\underbrace{(AA^+A)'}_{A} \\
&= B'A' = AB \\
B &= B\underbrace{AB}_{=AA^+} = \underbrace{BA}_{=A^+A}A^+ = A^+AA^+ = A^+
\end{aligned}$$

Thus, A^+ is unique. □

Remark 1.9. ***Construction of*** A^+

Ideas:

1. Use the SVD and $A^+ = VL^{-1}U'$
   - You must typically threshold small l_i (set them to zero)
2. Use standard package functions.
   - In R, ginv$(\cdot)$ in library(MASS).
   - In Matlab, `pinv`

1.3 Solving Linear Equations

We are trying to solve

$$\underbrace{A}_{m\times n}\underbrace{x}_{n\times 1} = \underbrace{b}_{m\times 1}.$$

where A, b are given and x is to be found (if it exists).

Definition 1.10. ***Consistent, Solution***

The equation $Ax = b$ is *consistent* iff there exists a *solution* x_0 such that $Ax_0 = b$. Note: the solution is not necessarily unique.

Remark 1.11.

Consistency is equivalent to $b \in R(A)$ (range space of A). The range is the subspace spanned by the columns of A.

Theorem 1.12.

The equation $Ax = b$ is consistent iff $AA^+b = b$.

Proof. Suppose the equation is consistent with solution x_0. Then

$$b = Ax_0$$
$$AA^+b = AA^+Ax_0 = Ax_0 = b.$$

Conversely, suppose $AA^+b = b$. Let $x_0 = A^+b$. Then

$$Ax_0 = AA^+b = b$$

So x_0 is a solution. □

Remark 1.13.

The previous theorem is true with any pseudoinverse.

Theorem 1.14.

The solutions to the consistent equation $Ax = b$ are of the form

$$\underbrace{x(c)}_{n\times 1} = \underbrace{A^+}_{n\times m}\underbrace{b}_{m\times 1} + (I_n - A^+A)c, \qquad c \in \mathbb{R}^n \tag{1.1}$$

Proof. From the previous theorem, consistency entails that $AA^+b = b$. Then

$$Ax(c) = AA^+b + (A - \underbrace{AA^+A}_{A})c$$
$$= b.$$

i.e. all such $x(c)$ solve the equation.

Conversely, suppose x_0 is any solution: $Ax_0 = b$. Then $A^+Ax_0 = A^+b$. Hence, plugging into (1.1) we get that

$$x(x_0) = A^+b + (I_n - A^+A)x_0 = A^+b + x_0 - \underbrace{A^+Ax_0}_{A^+b}$$
$$= x_0$$

□

Theorem 1.15.

The particular solution $x(0) = A^+b$ to the consistent equation $Ax = b$ has the smallest Euclidean norm among all solutions. That is, it is the solution closest to the origin.

Definition 1.16. ***Euclidean Norm***

Let $z = (z_1, z_2, \ldots, z_n)'$. The *Euclidean norm* is

$$|z|^2 = z'z = \sum_{i=1}^{n} z_i^2$$

Proof.

$$\begin{aligned}
|x(c)|^2 &= |A^+b + (I_n - A^+A)c|^2 \\
&= (A^+b + (I_n - A^+A)c)'(A^+b + (I_n - A'A)c) \\
&= |A^+b|^2 + |(I_n - A^+A)c|^2 + (A^+b)'((I_n - A'A)c) + ((I_n - A^+A)c)'A^+b \\
&= |A^+b|^2 + |(I_n - A^+A)c|^2 + 2\underbrace{[(I_n - A^+A)c]'A^+b}_{\text{cross-product}}
\end{aligned}$$

$$\text{cross-product} = c'(I_n - A^+A)A^+b = c'(A^+ - A^+AA^+)b = 0$$

$$|x(c)|^2 \geq |A^+b|^2$$

□

1.4 General Linear Model

Definition 1.17. ***General Linear Model***

$$\underbrace{y}_{n\times 1} = \underbrace{X}_{n\times p}\underbrace{\beta}_{p\times 1} + \underbrace{e}_{n\times 1}, \qquad \text{rank}(X) = r \leq p \leq n$$

y is the observation vector $(y_1, y_2, \ldots, y_n)'$.
β is the regression coefficients $(\beta_1, \beta_2, \ldots, \beta_n)'$.
X is the design matrix $\{x_{ij}\}$.
e is the error vector $(e_1, e_2, \ldots, e_n)'$.

Remark 1.18.

Note that $y = X\beta + e$ ca be written as

$$y_i = \sum_{j=1}^{p} x_{ij}\beta_j + e_i, \qquad 1 \le i \le n.$$

Remark 1.19. ***Standard Probability Models for*** e

1. Gaussian (or Normal) error model. The $\{e_i\}$ are i.i.d. $N(0, \sigma^2)$ random variables $\Leftrightarrow e_{n\times 1} \sim N(0, \sigma^2 I_n),\ 0 < \sigma^2 < \infty$
2. Gauss-Markov model. $\mathbb{E}(e) = \mathbf{0}$, $\mathrm{Cov}(e) = \sigma^2 I_n,\ 0 < \sigma^2 < \infty$
3. Strong Gauss-Markov model. The $\{e_i\}$ are i.i.d. random variables with $\mathbb{E}(e_i) = 0,\ \mathrm{Var}(e_i) = \sigma^2,\ 0 < \sigma^2 < \infty,\ \mathbb{E}(e_i^4) < \infty$

These models support the study of statistical properties of estimators for β, σ^2.

1.5 Least Squares Estimation of β

We have our model

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

Ideas: We want to solve this equation approximately. "The" least squares estimator $\hat{\beta}$ of β minimizes $|y - X\beta|^2$ over all possible $\beta \in \mathbb{R}^p$.
Questions: existence and uniqueness of $\hat{\beta}$.

Remark 1.20. ***Aside on Matrix Derivatives***

Suppose f is a real-valued function of some matrix $y_{m\times n} = \{y_{ij}\}$. That is, $f : \mathbb{C}^{m\times n} \to \mathbb{R}$. We define the partial derivative matrix

$$\frac{\partial f(y)}{\partial y} = \left\{\frac{\partial f(y)}{\partial y_{ij}}\right\} = \begin{pmatrix} \frac{\partial f(y)}{\partial y_{11}} & \cdots & \frac{\partial f(y)}{\partial y_{1n}} \\ \vdots & \ddots & \vdots \\ \frac{\partial f(y)}{\partial y_{m1}} & \cdots & \frac{\partial f(y)}{\partial y_{mn}} \end{pmatrix}$$

In particular, when $Y_{m\times 1} = y$ (a vector),

(a) $\frac{\partial(a'y)}{\partial y} = \frac{\partial(y'a)}{\partial y} = a$ for every $a \in \mathbb{C}^{m\times 1}$

(b) $\frac{\partial(y'y)}{\partial y} = 2y$

(c) $\frac{\partial(y'Ay)}{\partial y} = (A + A')y$, where $A \in \mathbb{C}^{m\times m}$, $= 2Ay$ if A is symmetric

Remark 1.21.

$$\frac{\partial(y'\overbrace{Az}^{})}{\partial y} = Az$$

by part (a) from above.

Remark 1.22. ***Reference Book***

H. Lütkepohl, Handbook of Matrices

Remark 1.23. ***Least Squares Criterion***

$$\begin{aligned} T(\beta) &= |y - X\beta|^2 = (y - X\beta)'(y - X\beta) \\ &= y'y + \beta' X' X\beta - 2y'X\beta \end{aligned}$$

From calculus, a necessary condition for a minimizer/maximizer of $T(\beta)$ is that

$$\frac{\partial T(\beta)}{\partial \beta} = 0$$

Using the matrix results we have, we see that

$$\begin{aligned} \frac{\partial(\beta' X' X\beta)}{\partial \beta} &= 2X'X\beta \\ \frac{\partial(y'X\beta)}{\partial \beta} &= (y'X)' = X'y \\ \frac{\partial T(\beta)}{\partial \beta} &= 2X'X\beta - 2X'\beta \end{aligned}$$

Thus, $X'X\beta = X'y$ is a necessary condition on β for minimizing $T(\beta)$.

Definition 1.24. ***Normal Equation***

The equation $X'X\beta = X'y$ is the *normal equation* for least squares estimation of β in the linear model $y = X\beta + e$.

Questions

1. Consistency of the normal equation? (Yes)
2. Solution set?

3. Do we have minimizers? (Yes)

2 9-27-11

2.1 LSE's and the Normal Equation

Proposition 2.1. ***Basic Facts & Inequalities***

$$\begin{aligned}\operatorname{rank}(AB) &\leq \min(\operatorname{rank}(A), \operatorname{rank}(B)) \\ \operatorname{tr}(AB) &= \operatorname{tr}(BA) \\ \operatorname{tr}(A+B) &= \operatorname{tr}(A) + \operatorname{tr}(B)\end{aligned}$$

A symmetric, idempotent matrix is an orthogonal projection, e.g. $AA^+,\ A^+A$.

Definition 2.2. ***Least Squares Estimator (LSE)***

Model

$$\underbrace{y}_{n\times 1} = \underbrace{X}_{n\times p}\underbrace{\beta}_{p\times 1} + \underbrace{e}_{n\times 1}, \qquad p \leq n,\ r = \operatorname{rank}(X) \leq p$$

$$\begin{aligned}T(m) &= |y - X\beta|^2 \\ \text{Least Squares Estimator} &= \arg\min_{m\in\mathbb{R}^p} T(m)\end{aligned}$$

Necessary condition: $\frac{\partial T}{\partial \beta} = 0 \leftrightarrow X'X\beta = X'y$.

Definition 2.3. ***Normal Equation***

The *normal equation* is

$$X'X\beta = X'y$$

for the LSE of β in the model $y = X\beta + e$.
Note: $\operatorname{rank}(X'X) = \operatorname{rank}(X) = r \leq p$.

Theorem 2.4.

1. $|y - X\beta|^2$ is minimized of $\beta \in \mathbb{R}^p$ by *any* solution to the normal equation.
2. $\hat{\eta} = X\hat{\beta}$ has the *same value* for every solution $\hat{\beta}$ to the normal equation.

Proof. 1. Let $\hat{\beta}$ be a solution: $X'X\hat{\beta} = X'y$. Then

$$\begin{aligned}
|y - X\beta|^2 &= (y - X\beta)'(y - X\beta) \\
&= \Big[(y - X\hat{\beta}) + X(\hat{\beta} - \beta)\Big]' \Big[(y - X\hat{\beta}) + X(\hat{\beta} - \beta)\Big] \\
&= |y - X\hat{\beta}|^2 + (\hat{\beta} - \beta)'X'X(\hat{\beta} - \beta) + 2(\hat{\beta} - \beta)' \underbrace{X'(y - X\hat{\beta})}_{=0 \text{ (Normal eq'n)}} \\
&= |y - X\hat{\beta}|^2 + |X(\hat{\beta} - \beta)|^2 \\
&\geq |y - X\hat{\beta}|^2
\end{aligned}$$

Thus, $\hat{\beta}$ minimizes $T(\beta)$.

2. Suppose $\hat{\beta}_1,\ \hat{\beta}_2$ both solve the normal equation.

$$\begin{aligned}
|X\hat{\beta}_1 - X\hat{\beta}_2|^2 &= (X\hat{\beta}_1 - X\hat{\beta}_2)'(X\hat{\beta}_1 - X\hat{\beta}_2) \\
&= (\hat{\beta}_1 - \hat{\beta}_2)'X'(X\hat{\beta}_1 - X\hat{\beta}_2) \\
&= (\hat{\beta}_1 - \hat{\beta}_2)(\underbrace{X'X\hat{\beta}_1}_{=y} - \underbrace{X'X\hat{\beta}_2}_{=y}) \\
&= 0
\end{aligned}$$

□

Remark 2.5. ***Consistency of the normal equation***

When is the normal equation consistent?

Geometrical heuristic: Let $\mathcal{R}(X)$ = range space of $X = \{\underbrace{X}_{n\times p}\ \underbrace{a}_{p\times 1} \mid a \in \mathbb{R}^p\}$ = subspace of $\mathbb{R}^p$ spanned by the columns of X.

Geometry says that $y - X\hat{\beta} \perp$ every vector in $\mathcal{R}(X)$
$\Leftrightarrow y - X\hat{\beta} \perp$ every column of X
$\Leftrightarrow X'y = X'X\hat{\beta}$.

This tells us that a solution (or solutions) exist, but it does not tell us how to find it.

Theorem 2.6. ***Algebraic Analysis of the Normal Equation***

The normal equation $X'X\beta = X'y$ is equivalent to the equation

$$X\beta = XX^+y.$$

Proof.

$$
\begin{aligned}
X'X\beta &= X'y \\
\underbrace{(X^+)'X'}_{(XX^+)'=XX^+} X\beta &= \underbrace{(X^+)'X'}_{(XX^+)'=XX^+} y \\
\underbrace{XX^+X}_{X}\beta &= XX^+y \\
X\beta &= XX^+y
\end{aligned}
$$

Conversely,

$$
\begin{aligned}
X\beta &= XX^+y \\
X'X\beta &= X'XX^+y \\
&= X'(X^+)'X'y \\
&= (XX^+X)'y \\
&= X'y
\end{aligned}
$$

□

Theorem 2.7.

The normal equation is consistent and the LSEs of β, i.e. the set of solutions of the normal equation, are

$$\hat{\beta}(c) = X^+y + (I_p - X^+X)c, \qquad c \in \mathbb{R}^p.$$

The LSE of minimum Euclidean norm is

$$\hat{\beta}(0) = X^+y.$$

$\hat{\eta} = X\hat{\beta}(c) = XX^+y \ \forall \ c \in \mathbb{R}^p$.

Proof. Consistency: $X\beta = XX^+y$ has solution $\hat{\beta}_0 = X^+y$. The solution set is

$$
\begin{aligned}
\hat{\beta}(c) &= X^+(XX^+y) + (I_p - X^+X)c, \qquad c \in \mathbb{R}^p \\
&= X^+y + (I_p - X^+X)c
\end{aligned}
$$

$\hat{\beta}(0)$ is the solution of minimum norm (see result from previous class). Finally,

$$
\begin{aligned}
\hat{\eta} &= X\hat{\beta}(c) \\
&= XX^+y + (X - \underbrace{XX^+X}_{X})c \\
&= XX^+y
\end{aligned}
$$

□

Remark 2.8.

1. The LSEs coincide with the solutions to $X\beta = y$ when the latter equation is consistent (it is usually not). That is, least squares generalizes the problem of solving consistent linear equations.
2. An alternative proof of the theorem using

$$\underbrace{X'X}_{A}\underbrace{\beta}_{X} = \underbrace{X'y}_{b}$$

This is consistent iff

$$\underbrace{X'X}_{A}\underbrace{\beta}_{X} = \underbrace{(X'X)}_{A}\underbrace{(X'X)^{+}}_{A^{+}}\underbrace{X'y}_{b}$$

which is true (use SVD of X). The solutions are

$$\begin{aligned}\hat{\beta}(c) &= (X'X)^{+}X'y + [I_p - (X'X)^{+}(X'X)]c, \qquad c \in \mathbb{R}^p \\ &= \underbrace{X^{+}y}_{\text{Lab 1}} + \underbrace{(I_p - X^{+}X)c}_{\text{use SVD}}\end{aligned}$$

Theorem 2.9. ***Uniqueness of the Normal Equation Solution***

$$X \in \mathbb{C}^{n\times p}, \qquad \operatorname{rank}(X) = r \leq p \leq n$$

The normal equation has a unique solution iff $\operatorname{rank}(X) = p$.

Proof. The solution set $\hat{\beta}(c) = X^{+}y + (I_p - X^{+}X)c$ is constant as a function of c if $\hat{\beta}(c) = \hat{\beta}(0)$ for all c. This means that

$$\begin{aligned}(I_p - X^{+}X)c &= 0 \qquad \forall\, c \in \mathbb{R}^n \\ (I_p - X^{+}X) &= 0 \\ X^{+}X &= I_p \\ \operatorname{rank}(X) &= \operatorname{rank}(X^{+}X) = \operatorname{rank}(I_p) = p\end{aligned}$$

From linear algebra, $\operatorname{rank}(X'X) = \operatorname{rank}(X) = p$, so $(X'X)^{-1}$ exists and $(X'X)^{+} = (X'X)^{-1}$. Hence, $\hat{\beta} = \hat{\beta}(0) = X^{+}y = (X'X)^{+}X'y = (X'X)^{-1}X'y$. □

2.2 Linear Parametric Functions of β

Consider

$$\psi = \underbrace{\lambda'}_{1\times p}\underbrace{\beta}_{p\times 1},$$

where λ is specified. The LSEs of of $\psi = \lambda'\beta$ are

$$\begin{aligned}\hat{\psi}(c) &= \lambda'\hat{\beta}(c), \qquad c \in \mathbb{R}^p \\ \hat{\beta}(c) &= X^{+}y + (I_p - X^{+}X)c\end{aligned}$$

For which $\lambda \in \mathbb{R}^p$ is $\hat{\psi}(c)$ uniquely defined?

Theorem 2.10.

The following are equivalent:

1. The LSEs $\hat{\psi}(c) = \lambda'\hat{\beta}(c),\ c \in \mathbb{R}^p$, of $\psi = \lambda'\beta$ are all equal to $\lambda' X^+ y = \hat{\psi}(0)$.
2. $X^+X\lambda = \lambda \leftrightarrow \lambda' X^+ X = \lambda'$
3. $\lambda = X'a \leftrightarrow \lambda' = a'X$ for some $a \in \mathbb{R}^n$
4. $\psi = \lambda'\beta = a'X\beta$ for all $\beta \in \mathbb{R}^p$ and some $a \in \mathbb{R}^n$. Thus, ψ is a linear function of $\eta = X\beta$.

Proof. 1 $\Leftrightarrow$ 2:

$$\begin{aligned} \hat{\psi}(c) &= \lambda' X^+ y + \lambda'(I_p - X^+X)c \qquad \forall\ c \in \mathbb{R}^p \\ &= \hat{\psi}(0) \\ \lambda'(I_p - X^+X)c &= 0 \qquad \forall\ c \in \mathbb{R}^p \\ \lambda'(I_p - X^+X) &= 0 \\ \lambda' X^+ X &= \lambda' \end{aligned}$$

2 $\Leftrightarrow$ 3:
If $\lambda' = \lambda' X^+ X$ then $\lambda' = a'X$ with $a' = \lambda' X^+$. Conversely, if $\lambda' = a'X$ for some $a \in \mathbb{R}^n$ then $\lambda' X^+ X = a' X X^+ X = a' X = \lambda$.
3 $\Rightarrow$ 4:
Obvious. □

Remark 2.11. ***Comments on previous proof***

1. Criterion 2 can be checked (approximately) by a computer.
2. If $\text{rank}(X) = p$, then $(X'X)^{-1}$ exists, $X^+ = (X'X)^{-1}X'$, and the conditions hold for every $\lambda \in \mathbb{R}^p$.
3. Later we will link this theorem to the theory of unbiased linear estimation of $\psi = \lambda'\beta$.

2.3 Polynomial Regression with One Covariate

General one-way layout model on means:

$$y_{ij} = m_i + e_{ij}, \qquad 1 \le i \le p,\ 1 \le j \le n_i$$

i labels $\{m_i \mid 1 \le i \le p\}$
j labels replications
$y = \{\{y_{ij} \mid 1 \le j \le n_i\},\ 1 \le i \le p\} =$ dictionary order

Vectorize:

$$m = \begin{pmatrix} m_1 \\ m_2 \\ \vdots \\ m_p \end{pmatrix}$$

$$y = \left\{ \left\{ y_{ij} \;\middle|\; 1 \le j \le n_i \right\},\ 1 \le i \le p \right\} = \text{ dictionary order}$$

$$n = \sum_{i=1}^{p} n_i = \text{ total sample size}$$

Let

$$\underbrace{C}_{n \times p} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ \vdots (n_1 \text{ total}) & \vdots & \cdots & 0 \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots (n_2 \text{ total}) & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & 1 \\ 0 & 0 & \cdots & \vdots (n_p \text{ total}) \\ 0 & 0 & \cdots & 1 \end{pmatrix} = \underline{\text{data-incidence matrix}}$$

The model says $\underbrace{y}_{n \times 1} = \underbrace{C}_{n \times p} \underbrace{m}_{p \times 1} + \underbrace{e}_{n \times 1}$.

$$e = \{ \{ e_{ij} \mid 1 \le j \le n_i \}, 1 \le i \le p \}$$

Note:

1. Columns of C are orthogonal, or $\text{rank}(C) = p$
2. $C'C = \text{diag}\{n_1, n_2, \ldots, n_p\}$

With no further information, this model is a *one-way layout* in which i labels the p factor levels. The LSE of m is

$$\underbrace{\hat{m}}_{p \times 1} = (C'C)^{-1} C'y$$
$$= (y_1, y_2, \ldots, y_p)'$$

where

$$y_{i\cdot} = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}.$$

<u>More generally:</u>
Suppose each m_i is a function of observed/known covariate with distinct values $x_1, x_2, \ldots, x_p$:

$$m_i = \mu(x_i), \qquad 1 \le i \le p.$$

Here the function μ may be specified (at least in part) or completely unknown.

Example 2.12. μ ***unknown***

Puts no restrictions on m_i, so previous analysis pertains:

$$\hat{m} = (y_1, \ldots, y_p).$$

2.4 Polynomial Regression

We postulate

$$\mu(x) = \sum_{k=1}^{d} \beta_k x^{k-1} = \text{ polynomial of degree } d-1 \text{ with } d \leq p.$$

The $\{\beta_k \mid 1 \leq k \leq d\}$ are unknown real values. The model is that

$$\begin{aligned} y_{ij} &= \mu(x_i) + e_{ij} \\ &= \underbrace{\sum_{k=1}^{d} \beta_k x_i^{k-1}}_{=m_i} + e_{ij} \end{aligned}$$

We vectorize:

Let

$$\underbrace{F}_{p\times d} = \begin{pmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^{d-1} \\ 1 & x_2 & x_2^2 & \cdots & x_2^{d-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_p & x_p^2 & \cdots & x_p^{d-1} \end{pmatrix}, \qquad \underbrace{\beta}_{d\times 1} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_d \end{pmatrix}$$

The polynomial model says that

$$\underbrace{y}_{n\times 1} = \underbrace{C}_{n\times p}\underbrace{F}_{p\times d}\underbrace{\beta}_{d\times 1} + \underbrace{e}_{n\times 1}$$

For $d \leq p \leq n$, $\text{rank}(F) = \text{rank}(CF) = d$.

Implication: Design matrix $X = CF$ has rank d

$$\hat{\beta} = (X'X)^{-1}X'y$$

3 9-29-11

3.1 Announcement

TA Office Hours: Monday 11-12, Wednesday 1-2

3.2 Polynomial Regression (Continued)

Remark 3.1. ***Polynomial Regression***

General Model:

$$\underbrace{y}_{n\times 1} = \underbrace{C}_{n\times p}\underbrace{m}_{m\times 1} + \underbrace{e}_{n\times 1}$$

where C = data-incidence matrix.

General One-Way Layout: $m \in \mathbb{R}^p$

Polynomial Submodels: $m_i = \sum_{k=1}^{d} \beta_k x_i^{k-1}, \quad 1 \le i \le p,\ d \le p \ \leftrightarrow\ m = F_d\beta,\ \beta = (\beta_1\ \cdots\ \beta_d)',$

$$\underbrace{F_d}_{p\times 1} = \begin{pmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^{d-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_p & x_p^2 & \cdots & x_p^{d-1} \end{pmatrix}$$

The $x_1, x_2, \ldots, x_p$ are distinct.

Theorem 3.2.

For $d \le p$, $\operatorname{rank}(F_d) = d = \operatorname{rank}(CF_d)$

Proof. Suppose that $\operatorname{rank}(F_d) < d$. Then

$$\underbrace{F_d}_{p\times d}\underbrace{c}_{d\times 1} = 0$$

for some $c \in \mathbb{R}^d,\ c = (c_1\ \cdots\ c_d)'$. Equivalently,

$$\sum_{k=1}^{d} c_k x_i^{k-1} = 0 \qquad \text{for } 1 \le i \le p$$

Thus, equation $\sum_{k=1}^{d} c_k x^{k-1} = 0$ has p distinct roots $x_1, x_2, \ldots, x_p$. $\Rightarrow\Leftarrow$ because $d - 1 < d \le p$. Hence, $\operatorname{rank}(F_d) = d$.

$$\operatorname{rank}(CF_d) \le \operatorname{rank}(F_d) = d$$
$$\operatorname{rank}(F_d) = \operatorname{rank}\left[(C'C)^{-1}C' \cdot CF_d\right] \le \operatorname{rank}(CF_d) \qquad (\operatorname{rank}(C) = p)$$

Hence, $\text{rank}(CF_d) = \text{rank}(F_d)$. $\square$

Remark 3.3.

1. F_p $(p \times p)$ has rank p. Thus, $\mathcal{R}(F_p) = \mathbb{R}^p$ ($\mathcal{R}$ = range space).
   If $x \in \mathbb{R}^p$, then $x = F_p a$ for $a = F_p^{-1}x$, so $\mathbb{R}^p \subset \mathcal{R}(F_p)$. Also $\mathcal{R}(F_p) = \mathbb{R}^p$.
   This is telling us that the polynomial submodel of degree $p-1$ is equivalent to the model where $m \in \mathbb{R}^p$.
2. More generally, $\mathcal{R}(F_1) \subset \mathcal{R}(F_2) \subset \cdots \subset \mathcal{R}(F_p) = \mathbb{R}^p$.
   The polynomial regression submodels are

$$\underbrace{m \in \mathcal{R}(F_1)}_{\text{constant}}, \ \underbrace{m \in \mathcal{R}(F_2)}_{\text{line}}, \ \cdots, \ \underbrace{m \in \mathbb{R}(F_p) = \mathbb{R}^p}_{\text{one-way layout}}$$

3. We have the model $y = CF_d\beta + e$, with $\text{rank}(CF_d) = d$, so $X = CF_d$ has full rank. So the LSE of β is uniquely $\hat{\beta} = (F_d'C'CF_d)^{-1}F_d'C'y$. This is mathematically correct but numerically unstable. The Moore-Penrose version works better: $\hat{\beta} = (CF_d)^+ y$. The SVD formula is even more accurate: $\hat{\beta} = UU'y$, where $CF_d = ULV'$. Main point: there are varying degrees of numerical instability when doing these fits.

3.3 Statistical Analysis under Random Error Models

Definition 3.4. ***Linear Estimability Model, Gauss-Markov Error Model***

Linear Estimability Model:

$$y = \underbrace{X}_{n\times p}\underbrace{\beta}_{p\times 1} + e, \qquad \text{rank}(X) = r \leq p \leq n$$

Gauss-Markov Error Model: e is a random vector such that $\mathbb{E}(e) = 0$, $\text{Cov}(e) = \sigma^2 I_n$, $0 < \sigma^2 < \infty$.

Definition 3.5. ***Linear Estimable, Unbiased Estimator***

A linear parametric function $\psi = \underbrace{\lambda'}_{1\times p}\underbrace{\beta}_{p\times 1}$ is *linearly estimable* if there exists $a \in \mathbb{R}^n$ such that $\tilde{\psi} = a'y$ is an *unbiased estimator* of ψ: $\mathbb{E}(\tilde{\psi}) = \lambda'\beta = \psi \ \forall \ \beta \in \mathbb{R}^p$.

Theorem 3.6.

The following are equivalent:

1. $\psi = \lambda'\beta$ is linearly estimable
2. The LSEs $\hat{\psi}(c) = \lambda'\hat{\beta}(c),\ c \in \mathbb{R}^n$, are all equal to $\hat{\psi} = \lambda' X^+ y$
3. $X^+X\lambda = \lambda \leftrightarrow \lambda' X^+ X = \lambda'$
4. $\lambda = X'a \leftrightarrow \lambda' = a'X$ for some $a \in \mathbb{R}^n$

Proof. From last time, we know that 2, 3, & 4 are equivalent. So it suffices to verify that $1 \leftrightarrow 4$.

Linear estimability gives us that there exists $a \in \mathbb{R}^n$ such that $\mathbb{E}(a'y) = \lambda'\beta\ \forall\ \beta \in \mathbb{R}^p$. Thus,

$$a' \underbrace{X\beta}_{\mathbb{E}(y)} = \lambda'\beta \qquad \forall\ \beta \in \mathbb{R}^p$$
$$a'X = \lambda'$$

□

Theorem 3.7. ***Gauss-Markov Theorem***

Suppose $\psi = \lambda'\beta$ is linearly estimable and the Gauss-Markov error model holds. The unique linear unbiased estimator of ψ with smallest variance is the LSE: $\hat{\psi} = \lambda' X^+ y$.

We have:

$$\text{linear estimators} \subset \text{unbiased estimators} \subset \text{all estimators}$$

So this theorem only looks at a small subset of estimators.

Proof. (Sketch) ψ is linearly estimable $\Leftrightarrow \lambda' = a'X$ for some $a \in \mathbb{R}^n$. So the LSE is

$$\hat{\psi} = \lambda' X^+ y = a' X X^+ y.$$

An arbitrary linear estimator $\tilde{\psi} = c'y$ is unbiased for ψ iff $c'X = \lambda'$. Hence,

$$c'X = a'X.$$

$$\begin{aligned}\mathrm{Var}(\tilde{\psi}) &= \mathrm{Var}\left[\hat{\psi} + (\tilde{\psi} - \hat{\psi})\right] \\ &= \mathrm{Var}(\hat{\psi}) + \mathrm{Var}(\tilde{\psi} - \hat{\psi}) + 2\,\mathrm{Cov}(\hat{\psi}, \tilde{\psi} - \hat{\psi})\end{aligned}$$

Note:

$$\begin{aligned}\mathrm{Var}(\tilde{\psi} - \hat{\psi}) &= \mathrm{Var}(c'y - a'XX^+y) \\ &= \mathrm{Var}\left[(c' - a'XX^+)y\right] = (c' - a'XX^+)\underbrace{\mathrm{Cov}(y)}_{\sigma^2 I_n}(c - \underbrace{XX^+}_{\text{symm}}a) \\ &= \sigma^2 |c - XX^+a|^2 \\ &> 0 \quad \text{unless } c = XX^+a \leftrightarrow \tilde{\psi} = \hat{\psi}\end{aligned}$$

$$\begin{aligned}
\operatorname{Cov}\left[\hat{\psi}, \tilde{\psi}-\hat{\psi}\right] &= \operatorname{Cov}\left[a'XX^+y, (c-XX^+a)'y\right] \\
&= a'XX^+ \underbrace{\sigma^2 I_n}_{\operatorname{Cov}(y)} (c-XX^+a) \\
&= \sigma^2\left[a'XX^+c - a'\frac{XX^+XX^+}{XX^+}a\right] \\
&= \sigma^2 a'XX^+(c-a) \\
&= \sigma^2(c'-a')XX^+a \\
&= \sigma^2 \underbrace{(c'X-a'X)}_{=0} X^+a \\
&= 0
\end{aligned}$$

□

3.4 SVD and Spectral Representation

Theorem 3.8.

Let A be an $m \times n$ matrix of rank r with singular value decomposition $A = ULV',\ U'U = V'V = I_r,\ L = \operatorname{diag}\{l_i\},\ l_1 \geq l_2 \geq \cdots \geq l_r > 0$.

Note that AA' is symmetric positive definite.

(a) A spectral representation for matrix AA' is

$$\underbrace{AA'}_{m\times m} = \begin{pmatrix} U & \overline{U} \end{pmatrix} \begin{pmatrix} L^2 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} U' \\ \overline{U}' \end{pmatrix}$$

where $(U\ \overline{U})$ is an orthogonal matrix of eigenvectors of AA' and L^2 gives the nonzero eigenvalues of AA'.

(b) A spectral representation of $A'A$ is

$$A'A = \begin{pmatrix} \underbrace{V}_{n\times r} & \underbrace{\overline{V}}_{n\times n-r} \end{pmatrix} \begin{pmatrix} L^2 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} V' \\ \overline{V}' \end{pmatrix}$$

where $(V\ \overline{V})$ is an orthogonal matrix of eigenvectors of $A'A$ and L^2 gives the nonzero eigenvalues of $A'A$.

(c) $UU' + \overline{U}\overline{U}' = I_m,\ VV' + \overline{V}\overline{V}' = I_n$

(d) $U'\overline{U} = 0,\ V'\overline{V} = 0$

Proof. (a) By the SVD,

$$AA' = UL\underbrace{V'\cdot V}_{I}LU' = UL^2U'$$

(b) By the SVD,

$$A'A = VL\underbrace{U'\cdot U}_{I}LV' = VL^2V'$$

(c)

$$I_m = \begin{pmatrix} U & \overline{U} \end{pmatrix} \begin{pmatrix} U' \\ \overline{U}' \end{pmatrix} = UU' + \overline{U}\overline{U}'$$

Similarly for $V, \overline{V}$.

(d) Eigenvectors are mutually orthogonal.

□

Remark 3.9.

Use these identities for problem (h) in HW1.

3.5 Distribution Theory

Remark 3.10. ***Canonical Representation of Least Squares Estimators***

Model: $y = X\beta + e = \eta + e,\ \eta = X\beta,\ \text{rank}(X) = r \le p \le n.$

Classical estimator of σ^2: $\hat{\sigma}^2 = \frac{1}{n-r}|y - \hat{\eta}|^2$, where $\hat{\eta} = XX^+y$ is the LSE of η.

Theorem 3.11.

Suppose X has SVD $X = ULV'$. Construct $\overline{U}$ so that $(U\ \overline{U})$ is an orthogonal matrix (e.g. previous theorem). Then

$$\hat{\eta} = UU'y, \qquad \hat{\sigma}^2 = \frac{1}{n-r}|\overline{U}y|^2.$$

Proof.

$$\begin{aligned}
\hat{\eta} &= XX^+y = ULV' \cdot VL^{-1}U'y = UU'y \\
(n-r)\hat{\sigma}^2 &= |y - \hat{\eta}|^2 = |y - UU'y|^2 = |\underbrace{(I_n - UU')}_{\overline{U}\overline{U}'} y|^2 \qquad\qquad I = UU' + \overline{U}\overline{U}' \\
&= |\underbrace{\overline{U}\overline{U}'}_{\text{symm}} y|^2 = y'\underbrace{\overline{U}\overline{U}'\overline{U}\overline{U}'}_{\text{symm}} y = y'\overline{U}\overline{U}'y \\
&= |\overline{U}'y|^2
\end{aligned}$$

□

Theorem 3.12.

Suppose the Gauss-Markov error model holds. Obviously, $\mathbb{E}(y) = \eta = X\beta,\ \mathrm{Cov}(y) = \sigma^2 I_n$.

1. $\mathbb{E}(\hat{\eta}) = \eta,\ \mathrm{Cov}(\hat{\eta}) = \sigma^2 XX^+$
2. $\mathbb{E}(\hat{\sigma}^2) = \sigma^2$

Proof. 1.

$$\begin{aligned}\mathbb{E}(\hat{\eta}) &= \mathbb{E}(XX^+y) = XX^+ \underbrace{X\beta}_{\mathbb{E}(y)} = X\beta = \eta \\ \mathrm{Cov}(\hat{\eta}) &= \mathrm{Cov}(XX^+y) = XX^+ \,\mathrm{Cov}(y) XX^+ \\ &= \sigma^2 XX^+XX^+ = \sigma^2 XX^+\end{aligned}$$

Fact:

$$\mathrm{Cov}(Ay) = A\,\mathrm{Cov}(y)A'$$

2.

$$(n-r)\hat{\sigma}^2 = |\overline{U}'y|^2 = |\overline{U}'e|^2$$

because, from the previous theorem,

$$\overline{U}'y = \overline{U}'(X\beta + e) = \overline{U}'X\beta + \overline{U}'e = \underbrace{\overline{U}'U}_{=0} LV'\beta + \overline{U}'e.$$

$$\begin{aligned}\mathbb{E}[(n-r)\hat{\sigma}^2] &= \mathbb{E}|\overline{U}'e|^2 = \mathbb{E}[e'\overline{UU}'e] = \mathbb{E}[\mathrm{tr}(e'\overline{UU}'e)] \\ &= \mathbb{E}[\mathrm{tr}(\overline{UU}'ee')] = \mathrm{tr}[\mathbb{E}(\overline{UU}'ee')] \\ &= \mathrm{tr}[\overline{UU}' \underbrace{\mathbb{E}(ee')}_{\mathrm{Cov}(e)=\sigma^2 I_n}] = \sigma^2\,\mathrm{tr}(\overline{UU}') = \sigma^2\,\mathrm{tr}(\underbrace{\overline{U}'\overline{U}}_{I_{n-r}}) \\ &= \sigma^2\,\mathrm{tr}(I_{n-r}) = (n-r)\sigma^2\end{aligned}$$

Fact:

$$\mathrm{tr}(AB) = \mathrm{tr}(BA), \qquad \mathrm{tr}(A+B) = \mathrm{tr}(A) + \mathrm{tr}(B)$$

Note: $\mathrm{Var}(\hat{\sigma}^2)$ depends on $\mathbb{E}(e_i^4)$ and more, which is not specified by the Gauss-Markov error model. □

Theorem 3.13.

Suppose the Gaussian error model holds ($e \sim N_n(0, \sigma^2 I_n)$). Obviously $y \sim N_n(\eta, \sigma^2 I_n)$, where $\eta = X\beta$. Then

1. $\hat{\eta} \sim N_n(\eta, \sigma^2 XX^+)$
2. $\frac{(n-r)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-r}$
3. $\hat{\eta}$ and $\hat{\sigma}^2$ are independent random variables

Proof. 1. $\eta = XX^+y$ = linear map of $y \sim N_n(\mathbb{E}(\hat{\eta}), \mathrm{Cov}(\hat{\eta})) = N(n, \sigma^2 XX^+)$ by the Gauss-Markov calculation.

2. As in a previous proof,

$$(n-r)\hat{\sigma}^2 = |\overline{U}'e|^2 = |z|^2, \qquad \text{where } z = \underbrace{\overline{U}'}_{(n-r)\times n}\ \underbrace{e}_{n\times 1}$$

$$z \sim N_{n-r}(\mathbb{E}(z), \mathrm{Cov}(z)) = N_{n-r}(0, \sigma^2 \underbrace{\overline{U}'\overline{U}}_{I_{n-r}})$$
$$= N_{n-r}(0, \sigma^2 I_{n-r})$$

Thus,

$$\frac{(n-r)\hat{\sigma}^2}{\sigma^2} = |w|^2 = \sum_{i=1}^{n-r} w_i^2 \sim \chi^2_{n-r}$$

where $w = z/\sigma \sim N_{n-r}(0, I_{n-r})$.

3.

$$\hat{\eta} = UU'y = \text{ function of } U'y$$
$$\hat{\sigma}^2 = \frac{1}{n-r}|\overline{U}'y|^2 = \text{ function of } \overline{U}'y$$

Observe that

$$\begin{pmatrix} U'y \\ \overline{U}'y \end{pmatrix} = \begin{pmatrix} U & \overline{U}' \end{pmatrix}_{n\times n,\ \text{orthogonal}} y$$
$$\sim N_n(*, O \cdot \sigma^2 I \cdot O')$$
$$= N_n(*, \sigma^2 \underbrace{OO'}_{I_n})$$
$$= N(*, \sigma^2 I_n)$$

$U'y, \overline{U}'y$ are independent.

□

Theorem 3.14. ***Lehmann-Scheffe***

Suppose $\psi = \lambda'\beta$ is linearly estimable and the Gaussian error model holds. The unique unbiased estimator of ψ with the smallest variance is the LSE: $\hat{\psi} = \lambda' X^+ y$. The unique unbiased estimator of σ^2 with the smallest variance is $\hat{\sigma}^2 = \frac{1}{n-r}|y - \hat{\eta}|^2$.

$$\hat{\psi}, \hat{\sigma}^2 \in \text{ all unbiased estimators } \subset \text{ all estimators}$$

4 10-4-11

4.1 Comparing Least Squares Fits

Remark 4.1. ***General Model***

$$y = \underbrace{X}_{n\times p}\underbrace{\beta}_{p\times 1} + e$$

$\operatorname{rank}(X) = r \le p \le n,\ e \sim N(0, \sigma^2 I_n),\ 0 < \sigma^2 < \infty.$
Under this model, $\eta = \mathbb{E}(y) = X\beta \in \mathcal{R}(X)$.
LSE: $\hat{\eta} = XX^+y = Py$. $P = XX^+$ is symmetric and idempotent.

Remark 4.2. ***Submodel***

$$y = \underbrace{X_0}_{n\times p_0}\underbrace{\beta_0}_{p_0\times 1} + e$$

$\operatorname{rank}(X_0) = r_0 < r,\ \mathcal{R}(X_0) \subset \mathcal{R}(X)$.
Under this model, $\eta = \mathbb{E}(y) = X_0\beta_0 \in \mathcal{R}(X_0)$.
LSE: $\hat{\eta} = X_0X^+y = P_0y$. $P_0 = X_0X_0^+$ is symmetric and idempotent.

Theorem 4.3.

Let $P_1 = P - P_0$. The following hold:

1. P is symmetric and idempotent with $\operatorname{rank}(P) = \operatorname{tr}(P) = r$.
2. P_0 is symmetric and idempotent with $\operatorname{rank}(P_0) = \operatorname{tr}(P_0) = r_0$.
3. $PP_0 = P_0P = P_0$
4. P_1 is symmetric and idempotent with $\operatorname{rank}(P_1) = \operatorname{tr}(P_1) = r_1 \equiv r - r_0$.
5. $P_1P_0 = P_0P_1 = 0$ (orthogonal)

Proof.

1. $\operatorname{rank}(P) = \operatorname{rank}(XX^+) = \operatorname{tr}(XX^+) = \operatorname{rank}(X)$ (HW 1 problem 1.i)
2. Likewise
3. Because $\mathcal{R}(X_0) \subset \mathcal{R}(X)$, each column in X_0 is a linear combination of columns of X. That is,

$$\underbrace{X_0}_{n\times p_0} = \underbrace{X}_{n\times p}\underbrace{A}_{p\times p_0} \quad \text{for some } A$$

Hence, $PP_0 = XX^+X_0X_0^+ = XX^+XAX_0^+ = \underbrace{XA}_{X_0}X_0^+ = X_0X_0^+ = P_0$. Also, $P_0P = P_0'P' = (PP_0)' = P_0' = P_0$.

4. Symmetry of P_1 is obvious.

$$\begin{aligned} P_1^2 &= (P-P_0)(P-P_0) = P^2 - \underbrace{P_0P}_{P_0} - \underbrace{PP_0}_{P_0} + P_0^2 \\ &= P - P_0 - P_0 + P_0 = P - P_0 \\ &= P_1 \\ \operatorname{rank}(P_1) &= \operatorname{rank}(P-P_0) = \operatorname{tr}(P) - \operatorname{tr}(P_0) = r - r_0 \end{aligned}$$

5.

$$P_1P_0 = (P-P_0)P_0 = PP_0 - P_0^2 = P_0 - P_0 = 0$$

□

Remark 4.4.

$P = P_0 + P_1$ decomposes the orthogonal projection P as the sum of two mutually orthogonal orthogonal projections.

Theorem 4.5. ***Spectral Representations of*** P, P_0, P_1

Let

$$\underbrace{P_0}_{n\times n} = \underbrace{U_0}_{n\times r_0}\underbrace{U_0'}_{r_0\times n}, \qquad \underbrace{P_1}_{n\times n} = \underbrace{U_1}_{n\times r_1}\underbrace{U_1'}_{r_1\times n}$$

be spectral representations of P_0, P_1. Let

$$\underbrace{U}_{n\times r} = \left(\underbrace{U_0}_{n\times r_0}\ \underbrace{U_1}_{n\times r_1}\right), \qquad r = r_0 + r_1.$$

Then

1. $U_1'U_0 = U_0'U_1 = 0$
2. $U'U = I_r$
3. $P = UU'$ is a spectral representation, $= U_0U_0' + U_1U_1' = P_0 + P_1$
4. $\mathcal{R}(X) = \mathcal{R}(P) = \mathcal{R}(U)$, $\mathcal{R}(X_0) = \mathcal{R}(P_0) = \mathcal{R}(U_0)$

Proof.

1. We know $P_1P_0 = 0$ from the previous theorem. Hence,

$$\begin{aligned} U_1'U_0 &= \underbrace{U_1'U_1}_{I_{r_1}} U_1'U_0 \underbrace{U_0'U_0}_{I_{r_0}} \\ &= U_1' \underbrace{U_1U_1'}_{P_1} \underbrace{U_0U_0'}_{P_0} U_0 \\ &= 0 \\ U_0'U_1 &= (U_1'U_0)' = 0 \end{aligned}$$

2.

$$UU' = \begin{pmatrix} U_0' \\ U_1' \end{pmatrix} \begin{pmatrix} U_0 & U_1 \end{pmatrix} = \begin{pmatrix} U_0'U_0 & U_0'U_1 \\ U_1'U_0 & U_1'U_1 \end{pmatrix} = \begin{pmatrix} I_{r_0} & 0 \\ 0 & I_{r_1} \end{pmatrix} = I_r$$

3.

$$UU' = \begin{pmatrix} U_0 & U_1 \end{pmatrix} \begin{pmatrix} U_0' \\ U_1' \end{pmatrix} = U_0U_0' + U_1U_1' = P_0 + P_1 = P$$

4.

$$\begin{aligned} Xa &= XX^+Xa = XX^+(Xa) \\ XX^+b &= X(X^+b) \end{aligned} \tag{4.1}$$

Thus, $\mathcal{R}(X) = \mathcal{R}(X^+)$. $XX^+ = P = UU'$ implies that $\mathcal{R}(XX^+) = \mathcal{R}(UU')$. Then $\mathcal{R}(UU') = \mathcal{R}(U)$ by (4.1) with U instead of X. Note $U' = U^+$. Hence, $\mathcal{R}(X) = \mathcal{R}(XX^+) = \mathcal{R}(UU') = \mathcal{R}(U)$.

□

4.2 Hypothesis Testing

Remark 4.6.

Model: $y = \eta + e,\ \eta = \underbrace{X}_{n\times p}\underbrace{\beta}_{p\times 1},\ \text{rank}(X) = r \le p \le n,\ e \sim N_n(0, \sigma^2 I_n),\ 0 < \sigma^2 < \infty$. To test:

$$\begin{aligned} &H : \eta \in \mathcal{R}(X_0) \subset \mathcal{R}(X), \quad 0 < \sigma^2 < \infty \\ &K : \eta \notin \mathcal{R}(X_0),\ \eta \in \mathcal{R}(X), \quad 0 < \sigma^2 < \infty \end{aligned}$$

Remark 4.7. ***The F-test for H versus K***

Rejects H for sufficiently large values of the *F-statistic*:

$$T = \frac{(|y - \eta_0|^2 - |y - \hat{\eta}|^2)/r_1}{\hat{\sigma}^2}$$

where

$$\hat{\eta}_0 = X_0 X_0^+ y = P_0 y$$
$$\hat{\eta} = X X^+ y = P_y$$

$r_0 = \text{rank}(X_0),\ r = \text{rank}(X),\ r_1 = r - r_0,$

$$\hat{\sigma}^2 = \frac{1}{n-r}|y - \hat{\eta}|^2 = \text{ est. of } \sigma^2 \text{ under general model.}$$

Note: this can be derived as a likelihood ratio test.

Theorem 4.8.

1. The F-statistic has two equivalent forms:

$$T = \frac{(|y - \eta_0|^2 - |y - \hat{\eta}|^2)/r_1}{\hat{\sigma}^2} = \frac{|\hat{\eta} - \hat{\eta}_0|^2/r_1}{\hat{\sigma}^2}$$

2. Under H, the distribution of T is $F_{r,n-r}$.

Proof. By the previous theorem, $\hat{\eta} = Py = UU'y$, $\hat{\eta}_0 = P_0 y = U_0 U_0' y$.

1.

$$\begin{aligned} |y - \hat{\eta}|^2 &= |y - Py|^2 = |(I_n - P)y|^2 \\ &= y'(I_n - P)^2 y = y'(I_n - P)y \end{aligned}$$

because $I_n - P$ is symmetric and idempotent:

$$(I_n - P)^2 = I_n - 2P + \underbrace{P^2}_{=P} = I_n - P$$

Similarly,

$$\begin{aligned} |y - \hat{\eta}_0|^2 &= |y - P_0 y|^2 = \dots \\ &= y'(I_n - P_0)y \end{aligned}$$

Thus,

$$\begin{aligned} |y - \hat{\eta}_0|^2 - |y - \hat{\eta}|^2 &= y'(I_n - P_0)y - y'(I_n - P)y \\ &= y'\left[(I_n - P_0) - (I_n - P)\right] y \\ &= y'(P - P_0)y = y' P_1 y \\ &= |P_1 y|^2 \end{aligned}$$

P_1 is symmetric and idempotent.

2. Under H, $\eta \in \mathcal{R}(X_0) = \mathcal{R}(U_0) \;\leftrightarrow\; \eta = U_0 a$ for some a. Thus,

$$U_1'\eta = \underbrace{U_1'U_0}_{0} a = 0$$

$$\begin{aligned} |\hat{\eta} - \hat{\eta}_0|^2 = |P_1 y|^2 = |U_1 U_1' y|^2 = y' U_1 U_1' U_1 U_1' y \\ = |U_1' y|^2 \end{aligned}$$

$$U_1' y \sim N_{r_1}(\underbrace{U_1'\eta}_{\mathbb{E}(U_1'y)}, \underbrace{U_1'\sigma^2 I_n U_1}_{\mathrm{Cov}(U'y)}) = N_{r_1}(0, \sigma^2 I_{r_1})$$

Hence, under H,

$$\frac{|\hat{\eta} - \hat{\eta}_0|^2}{\sigma^2} \sim X^2 r_1.$$

From earlier,

$$\frac{(n-r)\hat{\sigma}^2}{\sigma^2} \sim X^2_{n-r}$$

under the general model and therefore under H. $\hat{\eta}, \hat{\sigma}^2$ are independent. Thus, under H

$$T \sim \frac{X^2_{r_1}/r_1}{X^2_{n-r}/(n-r)} \sim F_{r,n-r}$$

□

5 10-6-11

5.1 Confidence Intervals for an Estimable Linear Parametric Function

Model: $y \sim N_n(\eta, \sigma^2 I_n)$, $\eta = \underbrace{X}_{n\times p}\beta$, $\text{rank}(X) = r \le p \le n$.

The LSE of η: $\hat{\eta} = XX^+y \sim N(n, \sigma^2 XX^+)$

Theorem 5.1.

Suppose that $\psi = \underbrace{\lambda'}_{1\times p}\underbrace{\beta}_{p\times 1}$ is a linearly estimable parametric function of (i.e. $\lambda' = \lambda' X^+X$) and the model is as specified above.

1. The LSE of ψ is
$$\underbrace{\hat{\psi}}_{1\times 1} = \lambda' X^+\beta = \lambda' X^+\hat{\eta} \sim N(\psi, \sigma^2\lambda'(X'X)^+\lambda)$$
2. The pivot
$$\frac{\hat{\psi} - \psi}{\hat{\sigma}\sqrt{\lambda'(X'X)^+\lambda}} \sim t_{n-r}$$

Proof. (sketch)

$$\begin{aligned}
\hat{\psi} &= \lambda' X^+ y = \lambda' X^+ \underbrace{XX^+y}_{\hat{\eta}} = \lambda' X^+\eta \\
\hat{\eta} &\sim N_n(n, \sigma^2 XX^+) \\
\hat{\psi} &\sim N(\mathbb{E}(\hat{\psi}), \text{Cov}(\hat{\psi})) \\
\text{Cov}(\hat{\psi}) &= (\lambda' X^+)\,\text{Cov}(\hat{\eta})(\lambda' X^+)' \\
&= \lambda' X^+ \cdot \sigma^2 XX^+(X^+)'\lambda \\
&= \sigma^2\lambda' X^+(X^+)'\lambda = \sigma^2\lambda'(X'X)^+\lambda
\end{aligned}$$

□

Notes:

1. Use this result to get confidence intervals for ψ
2. Invert the confidence intervals to test (for example)
$$\begin{aligned}
H &: \psi = \psi_0,\ 0 < \sigma^2 < \infty \\
K &: \psi \neq \psi_0,\ 0\sigma^2 < \infty
\end{aligned}$$

5.2 Risk and Estimated Risk of a Submodel Fit

Remark 5.2.

As before for the F-test, we have our general model:

$$y = \eta + e, \quad \eta = \underbrace{X}_{n\times p}\underbrace{\beta}_{p\times 1}, \quad \operatorname{rank}(X) = r \le p \le n$$

The LSE of η is $\hat{\eta} = Py$ for $P = XX^+$.

Submodel:

$$y = \eta_0 + e, \quad \eta_0 = \underbrace{X_0}_{n\times p_0}\underbrace{\beta_0}_{p_0\times 1}, \quad \operatorname{rank}(X_0) = r_0 < r, \quad \mathcal{R}(X_0) \subset \mathcal{R}(X)$$

$e \sim N_n(0, \sigma^2 I_n)$. The LSE of η_0 is $\hat{\eta}_0 = P_0 y$, where $P_0 = X_0 X_0^+$.

Remark 5.3. ***Estimation Approach to Comparing Fits***

1. The general model is taken to be true (unlike in testing).
2. We assess an estimator $\tilde{\eta}$ of η through its (quadratic) *risk*

$$r^{-1} \underbrace{\mathbb{E}|\tilde{\eta} - \eta|^2}_{\text{(under general model)}} := R(\tilde{\eta}, \eta, \sigma^2)$$

3. Ideally, minimize risk by choice of $\tilde{\eta}$.

Theorem 5.4. ***Mallows (1973)***

1. $R(\hat{\eta}, \eta, \sigma^2) = \sigma^2$
2. $R(\hat{\eta}_0, \eta, \sigma^2) = r^{-1}\left[r_0\sigma^2 + |\eta - P_0\eta|^2\right] = r^{-1}\left[r_0\sigma^2 + \operatorname{tr}\{(I_n - P_0)\eta\eta'\}\right]$

Proof. 1.

$$\begin{aligned} rR(\hat{\eta}, \eta, \sigma^2) &= \mathbb{E}|\hat{\eta} - \eta|^2 = \mathbb{E}\operatorname{tr}[(\hat{\eta} - \eta)(\hat{\eta} - \eta)'] \\ &= \operatorname{tr}(\operatorname{Cov}(\hat{\eta})) = \operatorname{tr}[\sigma^2 P] = \sigma^2 \operatorname{tr}(P) \\ &= r\sigma^2 \end{aligned}$$

2.

$$rR(\hat{\eta}_0, \eta, \sigma^2) = \mathbb{E}|\underbrace{\hat{\eta}_0}_{=Py} - \eta|^2 = \mathbb{E}\,|(P_0 y - P_0\eta) - (I_n - P_0)\eta|^2$$

$$= \mathbb{E}|P_0(y-\eta)|^2 + |(I_n - P_0)\eta|^2 - \mathbb{E}\left[(\eta - P_0\eta)'\underbrace{P_0(y-\eta)}_{\mathbb{E}P_0(y-\eta)=0}\right]$$

$$\mathbb{E}|P_0(y-\eta)|^2 = \mathbb{E}\,\mathrm{tr}\left[P_0(y-\eta)(y-\eta)'P_0\right] \qquad P_0 \text{ is symmetric \& idempotent}$$
$$= \mathrm{tr}\left[P_0\,\mathrm{Cov}(y)P_0\right] = \sigma^2\,\mathrm{tr}(P_0)$$
$$= \sigma^2 r_0$$

Notice:

$$|(I_n - P_0)\eta|^2 = \mathrm{tr}\left[(I_n - P_0)\eta\eta'(I_n - P_0)\right]$$
$$= \mathrm{tr}\left[(I_n - P_0)^2\eta\eta'\right]$$
$$= \mathrm{tr}[(I_n - P_0)\eta\eta']$$

□

Note: $R(\hat{\eta}_0, \eta, \sigma^2)$ depends on σ^2 and $\eta\eta'$ which are unknown. We therefore estimate σ^2 by $\hat{\sigma}^2$ and $\eta\eta'$ by $yy' - \hat{\sigma}^2 I_n$.

Remark 5.5. ***Mallows' idea (1973)***

Justify their estimation through $\mathbb{E}(\hat{\sigma}^2) = \sigma^2$ (recall: $\hat{\sigma}^2 = \frac{1}{n-r}|y - \hat{\eta}|^2$) and $\mathbb{E}[yy' - \hat{\sigma}^2 I_n] = \eta\eta'$ because

$$\mathbb{E}(yy') = \mathbb{E}(\eta + e)(\eta + e)' = \mathbb{E}[\eta\eta' + e\eta' + \eta e' + ee']$$
$$= \eta\eta' + 0 + 0 + \sigma^2 I_n = \eta\eta' + \sigma^2 I_n$$

Theorem 5.6.

1. The *estimated risk* of $\hat{\eta}_0$ is
$$\hat{R}(\hat{\eta}_0) = r^{-1}\left[\hat{\sigma}^2 r + \mathrm{tr}[(I_n - P_0)(yy' - \hat{\sigma}^2 I_n)]\right]$$
and it satisfies $\mathbb{E}[\hat{R}(\hat{\eta}_0)] = R(\hat{\eta}_0, \eta, \sigma^2)$, i.e. $\hat{R}(\hat{\eta}_0)$ is an unbiased risk estimator.
2. $\hat{R}(\hat{\eta}_0) = r^{-1}[|y - \hat{\eta}_0|^2 + (2r_0 - n)\hat{\sigma}^2]$

Proof. 1. By substituting $\hat{\sigma}^2$, $yy' - \hat{\sigma}^2 I_n$ for σ^2, $\eta\eta'$ in $R(\hat{\eta}_0, \eta, \sigma^2)$.

2.

$$r\hat{R}(\hat{\eta}_0) = r_0\hat{\sigma}^2 - \mathrm{tr}(I_n - P_0)\hat{\sigma}^2 + \mathrm{tr}[(I_n - P_0)yy']$$
$$-\mathrm{tr}(I_n - P_0)\hat{\sigma}^2 = -(n - r_0)\hat{\sigma}^2 \qquad \text{because } \mathrm{tr}(P_0) = r_0$$
$$\mathrm{tr}[(I_n - P_0)yy'] = y'(I_n - P_0)y = y'(I_n - P_0)'(I_n - P_0)y = |y - P_0 y|^2 = |y - \hat{\eta}_0|^2$$

□

Notes:

1. Mallows C_p criterion (1973) is

$$C_p = |y - \hat{\eta}_0|^2 + 2r_0\hat{\sigma}^2 = \text{one-to-one transform of } \hat{R}(\hat{\eta}_0)$$

2. There exists asymptotic theory under which $\hat{R}(\hat{\eta}_0) \xrightarrow{p} R(\hat{\eta}_0, \eta, \sigma^2)$.

5.3 Specifying Submodels for Means

Remark 5.7.

Model: $\underbrace{y}_{n\times 1} = \underbrace{C}_{n\times p}\underbrace{m}_{p\times 1} + \underbrace{e}_{n\times 1}$, $e \sim N_n(0, \sigma^2 I_n)$, where C is the data-incidence matrix (see polynomial regression example) which records the replication pattern.

$$C'C = \text{diag}\{n_i\},$$

where n_i is the number of observations of m_i.

Submodel: Add the restriction that $m \in \mathcal{F}$ is a subspace of $\mathbb{R}^p$, with $\dim(\mathcal{F}) = r < p$. For now, assume that $\text{rank}(C) = p$, i.e. each n_i is nonzero (we have at least one observation for each mean; this is called *complete design*).

Theorem 5.8.

Suppose that $\mathcal{F} = \text{range}(\underbrace{F}_{p\times t})$, where $\text{rank}(\mathcal{F}) = r$. Let $\underbrace{Q}_{p\times p} = FF^+$ have spectral representation $\underbrace{Q}_{p\times p} = \underbrace{V}_{p\times r}\underbrace{V'}_{r\times p}$, where $V'V = I_n$. The following are equivalent:

1. $m \in \mathcal{F}$
2. $m = \underbrace{F}_{p\times t}\underbrace{\alpha}_{t\times 1}$ for some $\alpha \in \mathbb{R}^t$
3. $m = Qm$
4. $m = Q\underbrace{\beta}_{p\times 1}$ for some $\beta \in \mathbb{R}^p$
5. $m = V\gamma$ for some $\gamma \in \mathbb{R}^r$

Proof. 1 ⇔ 2: By definition.
2 ⇔ 3: The relation $m = F\alpha$ is consistent iff $FF^+m = m \Leftrightarrow Qm = m$.
3 ⇒ 4: for $\beta = m$
4 ⇒ 3: $m = Q\beta = QQ\beta = Qm$
4 ⇔ 5: The relation $m = \gamma j$ is consistent iff $VV^+m = m \Leftrightarrow UU'm = m$ (because $U' = U^+$) $\Leftrightarrow Qm = m$. □

Notes: Submodels on m have multiple equivalent expressions which lead to multiple expressions for least squares estimators under the submodels.

Theorem 5.9.

Under the submodel $m \in \mathcal{F}$, the LSE of $\eta_{\mathcal{F}} = C_m$ is

$$\hat{\eta}_{\mathcal{F}} = CQ(CQ)^+y = C(CQ)^+y = CU(CU)^+y = CF(CF)^+y$$

Proof.
$m = Q\beta \;\Rightarrow\; y = \underbrace{CQ}_{X}\beta + e$. The LSE of $CQ\beta$ is $XX^+y = CQ(CQ)^+y = C(CQ)^+y$ (by HW 1 Problem 1.m). $m = V\gamma \;\Rightarrow\; y = \underbrace{CV}_{X}\gamma + e$. The LSE of $CV\gamma$ is now $CV(CV)^+y$. $m = F\alpha \;\Rightarrow\; y = CF\alpha + e$. The LSE of $CF\alpha$ is $CF(CF)^+y$.

Note that $(CQ)^+ = (C\underbrace{VV'}_{Q})^+ = V(CU)^+$ (by HW1 Problem 1.l).

Note: we assume that $\operatorname{rank}(\underbrace{C}_{n\times p}) = p$. Then

1. The LSE $\hat{m}_{\mathcal{F}}$ of $m_{\mathcal{F}}$ (the submodel restricted to m) is
$$\begin{aligned}\hat{m}_{\mathcal{F}} &= C^+\hat{\eta}_{\mathcal{F}} = (C'C)^{-1}C'\eta_{\mathcal{F}} \\ &= Q(CQ)^+y = (Ca)^+y = V(CV)^+y = F(CF)^+y\end{aligned}$$
2. By HW 1 Problem 1.g, we get that
$$\begin{aligned}\hat{\eta}_{\mathcal{F}} &= CV(CV)^+y = CV(V'C'CV)^+V'C' \\ &= CV(V'C'CV)^{-1}V'C' \qquad \textbf{because } \operatorname{rank}(CV) = \operatorname{rank}(V) = r,\ \operatorname{rank}(C) = p\end{aligned}$$

and $\operatorname{rank}[(CV)'CV] = \operatorname{rank}(CV)$. This is a Moore-Penrose-free expression! □

5.4 Projection Form of One-Way ANOVA

Remark 5.10. ***One-Way Layout of Means***

Model: $y_{ij} = m_i + e_{ij},\ 1 \le i \le p,\ 1 \le j \le n_i$. The e_{ij} are i.i.d. $\sim N(0, \sigma^2)$.
Vectorize:

$$\begin{aligned} y &= Cm + e \\ y &= \begin{pmatrix} y_1 & y_2 & \cdots & y_n \end{pmatrix}' \\ m_i &= \begin{pmatrix} m_1 & m_2 & \cdots & m_p \end{pmatrix}' \\ C &= \text{data incidence matrix} \\ n &= \sum_{i=1}^{p} n_i \end{aligned}$$

Remark 5.11. ***Basic ANOVA Decomposition***

Classically, $m_i = \mu + \alpha_i,\ 1 \le i \le p,\ \alpha_0 = 0.$

$$\alpha_\cdot = \frac{1}{p}\sum_{i=1}^{p} \alpha_i = \text{average over dotted subscript}$$

$$\begin{pmatrix} m_1 \\ m_2 \\ \vdots \\ m_p \end{pmatrix} = \begin{pmatrix} m_\cdot \\ m_\cdot \\ \vdots \\ m_\cdot \end{pmatrix} + \begin{pmatrix} m_1 - m_\cdot \\ m_2 - m_\cdot \\ \vdots \\ m_p - m_\cdot \end{pmatrix}$$

This is the *vector form of ANOVA decomposition*.

Theorem 5.12. ***Projection Form***

Let $\underbrace{u}_{p\times 1} = p^{-1/2}(1,\dots,1)$. $U'U = \mathbf{1}$. Let $\underbrace{P_0}_{p\times p} = UU'$, $P_1 = I_p - P_0$. Then

1. P_0, P_1 are each symmetric and idempotent.
2. $P_0P_1 = P_1P_0 = 0$
3. $I_p = P_0 + P_1$

Proof. In particular, $m = P_0m + P_1m$ is the ANOVA decomposition because

$$P_0m = U(U'm) = \frac{1}{p}e(e'm)$$
$$= \frac{1}{p}\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}\sum_{i=1}^{p} m_i = \begin{pmatrix} m_\cdot \\ \vdots \\ m_\cdot \end{pmatrix}$$
$$P_1m = m - P_0m = \begin{pmatrix} m_1 - m_\cdot \\ m_2 - m_\cdot \\ m_p - m_\cdot \end{pmatrix}$$

The submodels $\{m_i\}$ are all equal $\Leftrightarrow$ $\{\alpha_i\}$ are all $0 \Leftrightarrow P_1m = 0 \Leftrightarrow P_0m = m \Leftrightarrow m_i = P_0\beta$ for some $\beta \in \mathbb{R}^p$.

Here the submodel is $y = \eta_0 + e,\ \eta_0 = CP_0\beta$. The LSE $\hat{\eta}_0 = CP_0(CP_0)^+y = C(CP_0)^+y$.

General model is $y = \eta + e,\ \eta = Cm,\ m \in \mathbb{R}^p$. The LSE $\hat{\eta} = C^+y = C(C'C)^{-1}C'y,\ \hat{m} = (C'C)^{-1}C'y.$ □

Next time we will address the relation to classical simple formulas.

6 10-11-11

6.1 Models for Means

Remark 6.1. ***Submodels***

General Model: $y = Cm + e$
Submodels restrict m, e.g.

$$m = \underbrace{Q}_{p\times p}\beta, \qquad \beta \in \mathbb{R}^p, \qquad Q \text{ symmetric \& idempotent}$$

with $\text{rank}(Q) = \text{tr}(Q) = r$. Equivalently,

$$m = \underbrace{V}_{n\times r}\underbrace{\gamma}_{r\times 1}, \qquad \gamma \in \mathbb{R}^p, \qquad Q = VV' = \text{ spectral representation}$$

6.1.1 One-Way Layout of Means

Remark 6.2. ***One-Way Layout***

General Model (classical form)

$$y_{ij} = m_i + e_{ij}, \quad 1 \le i \le p, \quad 1 \le j \le n_i \tag{6.1}$$

where $\{e_{ij}\}$ are i.i.d. $N(0, \sigma^2)$, $\sigma^2 > 0$. Vectorize:

$$\underbrace{m}_{p\times 1} = (m_1, \ldots, m_p)'$$

$$\underbrace{y}_{n\times 1} = \left\{\{y_{ij} \mid 1 \le j \le n_i\}, 1 \le i \le p\right\}, \qquad n = \sum_{i=1}^{p} n_i$$

$$\underbrace{C}_{n\times p} = \text{ data incidence matrix}$$

$$C'C = \begin{pmatrix} n_1 & 0 & \cdots & 0 \\ 0 & n_2 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & n_p \end{pmatrix}$$

Thus, the general model (6.1) can be written as: $y = Cm + e$. The LSE of m is:

$$\hat{m} = C^+ y = (C'C)^{-1}C'y = (y_{1.}, y_{2.}, \cdots, y_{p.})'$$

$$y_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$$

The LSE of $\eta = Cm$ is

$$\hat{\eta} = C\hat{m}.$$

Remark 6.3. ***Projection Form of Basic ANOVA***

Classical form: $m_i = \mu + \alpha_i,\ 1 \le i \le p$, where $\alpha_\cdot = 0$.
This is a one-to-one map of $\{m_i\}$ to $\mu,\ \{\alpha_i\}$ in which $\mu = m_\cdot,\ \alpha_i = m_i - m_\cdot$.

Projection Form
Let $P_0 = uu',\ \underbrace{u}_{p\times 1} = p^{-1/2}(1, 1, \ldots, 1)'$.

$$P_1 = I_p - P_0 = I - uu'$$

Note:

1. P_0, P_1 are symmetric & idempotent
2. $P_0P_1 = P_1P_0 = 0$
3. $I_p = P_0 + P_1$

The *ANOVA decomposition* is

$$m = P_0m + P_1m$$

which is equivalent to $m_i = \mu + \alpha_i,\ \alpha_\cdot = 0$, because

$$\begin{aligned} uu'm &= u(u'm) \\ &= (m_\cdot, m_\cdot, \cdots, m_\cdot)' \\ P_1m = m - P_0m &= \begin{pmatrix} m_1 - m_\cdot \\ m_2 - m_\cdot \\ \vdots \\ m_p - m_\cdot \end{pmatrix} \end{aligned}$$

Remark 6.4. ***Classical Submodel***

$\{m_i\}$ equal $\Leftrightarrow$ $\{\alpha_\cdot\}$ all $0 \Leftrightarrow\ m = P_0\beta,\ \beta \in \mathbb{R}^p\ (m = P_0m)$.
So $y = CP_0\beta + e$ is the submodel, with $e \sim N_n(0, \sigma^2 I_n)$.

The LSE of $\eta_0 = CP_0\beta$ is

$$\hat{\eta}_0 = CP_0(CP_0)^+y = C(CP_0)^+y.$$

the LSE of $m_0 = P_0\beta$ is

$$\hat{m}_0 = (CP_0)^+y.$$

We have assumed that $n_i > 0,\ 1 \le i \le p$, so $\text{rank}(C) = p$.

Remark 6.5. ***Reduction to Elementary Form***

$$\hat{m}_0 = (CP_0)^+ y = (C \underbrace{uu'}_{P_0})^+ y = u(Cu)^+ y \qquad \text{by lab 1, problem 1.l } (S = I,\ T = u)$$

$$(Cu)^+ = (\underbrace{u'}_{1\times p}\ \underbrace{C'}_{p\times p}\ \underbrace{C}_{p\times p}\ \underbrace{u}_{p\times 1})^+ u'C' = \left(\frac{n}{p}\right)^+ u'C' \qquad \text{by lab 1, problem 1.g}$$

$$\hat{m}_0 = \left(\frac{p}{n}\right) uu'C'y = \begin{pmatrix} y_{..} \\ y_{..} \\ \vdots \\ y_{..} \end{pmatrix}$$

Remark 6.6. ***Ranks of General Model and Submodel Deisgn Matrices***

Both have the form $QC\beta$.

$$Q = I_p \qquad \text{(General Model)}$$

with $r = \text{rank}(C) = p$. So C is of full rank.

$$Q = P_0 \qquad \text{(Submodel)}$$

with $r_0 = \text{rank}(CP_0) = \text{rank}(P_0) = \text{tr}(P_0) = \text{tr}(uu') = \text{tr}(u'u) = \text{tr}(1) = 1$.

Consequences

1. F-test for H: submodel $m = P_0\beta$ holds
   vs. K: not so, general model holds
   refers
   $$T = \frac{|\hat{\eta} - \hat{\eta}_0|^2/(p-1)}{\hat{\sigma}^2} \quad \text{with} \quad \hat{\sigma}^2 = \frac{1}{n-p}|y - \hat{\eta}|^2$$
   to $F_{p-1,n-p}$.
2. Estimated risks of $\hat{\eta}, \hat{\eta}_0$ as competing estimators for $\eta = Cm$ under the general model are:
   $$\hat{R}(\hat{\eta}) = \hat{\sigma}^2$$
   $$\hat{R}(\hat{\eta}_0) = \underbrace{p^{-1}}_{r^{-1}} \left[|y - \hat{\eta}_0|^2 + (\underbrace{2}_{2r_0} - n)\hat{\sigma}^2 \right]$$
   because $r = p,\ r_0 = 1$.

6.1.2 Two-Way Layout of Means

(*complete layout* = at least 1 observation for every mean)

General Model (classical form):

$$y_{ijk} = m_{ij} + e_{ijk}, \qquad \underbrace{1 \le i \le p_1, \quad 1 \le j \le p_2}_{\text{label levels of factors}}, \qquad \underbrace{1 \le k \le n_{ij}}_{\text{labels replications}}$$

where $\{e_{ijk}\}$ are i.i.d. $N(0,\sigma^2),\ \sigma^2>0$.
Vectorize:
Set $p=p_1p_2$. Let

$$\begin{aligned}\underbrace{m}_{p\times 1} &= \{\{m_{ij} \mid 1\le i\le p_1\},\ 1\le j\le p_2\}\\ &= \text{ mirror dictionary order}\\ &= \text{ stack columns of matrix } \{m_{ij}\}\\ &= \text{ special case of array order}\\ \underbrace{y}_{n\times 1} &= \{\{\{y_{ijk} \mid 1\le k\le n_{ij}\}1\le i\le p_1\},\ 1\le j\le p_2\}\end{aligned}$$

$$n=\sum_{i=1}^{p_1}\sum_{j=1}^{p_2} n_{ij}$$

e is similar.

$$\underbrace{C}_{n\times p}=\begin{pmatrix} 1 & 0 & \cdots & 0\\ \vdots_{n_{11}\text{ total}} & \vdots & \ddots & \vdots\\ 1 & 0 & \cdots & 0\\ 0 & 1 & \cdots & 0\\ 0 & \vdots_{n_{21}\text{ total}} & \ddots & \vdots\\ 0 & 1 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & 1\\ \vdots & \vdots & \ddots & \vdots_{n_{p_1p_2}\text{ total}}\\ 0 & 0 & \cdots & 1\end{pmatrix}$$

Remark 6.7. ***General Model***

$y=Cm+e,\ e\sim N_n(0,\sigma^2 I_n)$. LSE of m is

$$\hat{m}=C^+y=(C'C)^{-1}C'y=\{\{y_{ij} \mid 1\le i\le p_1\},\ 1\le j\le p_2\}$$

$$y_{ij.}=\frac{1}{n_{ij}}\sum_{k=1}^{n_{ij}} y_{ijk}$$

The LSE of η is

$$\hat{\eta}=C\hat{m}.$$

Remark 6.8. ***Classical ANOVA Decomposition***

$$\begin{aligned} m_{ij} &= \mu+\alpha_i+\beta_j+\gamma_{ij}\\ \alpha_\cdot=\beta_\cdot=\gamma_{i\cdot}=\gamma_{\cdot j} &= 0\\ \updownarrow &\text{ one-to-one}\\ \mu=m_{\cdot\cdot},\quad \alpha_i=m_{i\cdot}-m_{\cdot\cdot},\quad \beta_j &= m_{\cdot j}-m_{\cdot\cdot},\quad \gamma_{ij}=m_{ij}-m_{i\cdot}-m_{\cdot j}+m_{\cdot\cdot}\end{aligned}$$

7 10-13-11

7.1 The Kronecker Product and vec

Definition 7.1. ***Kronecker Product***

The *Kronecker product* of A $(m \times n)$ and B $(r \times s)$ is

$$A \otimes B = \begin{pmatrix} a_{11}B & a_{12}B & \cdots & a_{1n}B \\ a_{21}B & a_{22}B & \cdots & a_{2n}B \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}B & a_{m2}B & \cdots & a_{mn}B \end{pmatrix}$$

Definition 7.2. ***vec***

If $\underbrace{X}_{n \times p} = \begin{pmatrix} x_{(1)} & x_{(2)} & \cdots & x_{(p)} \end{pmatrix} = \{x_{ij}\}$, where each $x_{(i)}$ is $n \times 1$, then

$$\operatorname{vec}(X) = \begin{pmatrix} x_{(1)} \\ x_{(2)} \\ \vdots \\ x_{(p)} \end{pmatrix}$$

Thus, it takes: stacked columns of $X \leftrightarrow \{\{x_{ij} \mid 1 \le i \le n\},\ 1 \le j \le p\}$.

Note: we can reverse $\operatorname{vec}(X)$ given n, p, "unvec."

Basic Properties (of Mardia, Kent, Bibby)

1. For scalar c, $c(A \otimes B) = (CA) \otimes B = A \otimes (cB)$. We can write $cA \otimes B$.
2. $A \otimes (B \otimes C) = (A \otimes B) \otimes C$. We can write $A \otimes B \otimes C$.
3. $(A \otimes B)' = A' \otimes B'$.
4. $(A \otimes B)(F \otimes G) = (AF) \otimes (BG)$
5. $(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$ for A, B nonsingular.
6. $(A + B) \otimes C = (A \otimes C) + (B \otimes C)$
7. $A \otimes (B + C) = (A \otimes B) + (A \otimes C)$
8. $(A_1 + A_2) \otimes (B_1 + B_2) = (A_1 \otimes B_1) + (A_1 \otimes B_2) + (A_2 \otimes B_1) + (A_2 \otimes B_2)$
9. $\operatorname{vec}(AXB) = (B' \otimes A) \operatorname{vec}(X)$
10. $\operatorname{tr}(A \otimes B) = \operatorname{tr}(A) \operatorname{tr}(B)$

7.2 ANOVA Decomposition for Two-Way Layout

Remark 7.3. ***Classical Form***

$$
\begin{aligned}
m_{ij} &= \mu + \alpha_i + \beta_j + \gamma_{ij}, \qquad 1 \le i \le p_1 \\
\alpha_\cdot &= \beta_\cdot = \gamma_{i\cdot} = \gamma_{\cdot j} = 0 \\
m_{ij} &= m_{\cdot\cdot} + (m_{i\cdot} - m_{\cdot\cdot}) + (m_{\cdot j} - m_{\cdot\cdot}) + (m_{ij} - m_{i\cdot} - m_{\cdot j} + m_{\cdot\cdot})
\end{aligned}
$$

In order to put this in projection form, we set

$$
\begin{aligned}
\underbrace{M}_{p_1 \times p_2} &= \{m_{ij}\} \\
\underbrace{m}_{p \times 1} &= \mathrm{vec}(M) = \text{ mirror dictionary vectorization of } \{m_{ij}\} \\
p &= p_1 p_2 \\
k &= 1, 2
\end{aligned}
$$

Let

$$
\begin{aligned}
\underbrace{u_k}_{p_k \times 1} &= p_k^{-1/2}(1, 1, \ldots, 1)' \\
J_k &= u u_k' \\
H_k &= I_{p_k} - J_k
\end{aligned}
$$

Note:

1. J_k, H_k are symmetric & idempotent
2. $H_k J_k = J_k H_k = 0$ (because $H_k u_k = (I_{p_k} - u_k u_k')u_k = u_k - u_k = 0$, since $u_k' u_k = I$)

Thus

$$
\begin{aligned}
I_p = I_{p_1} \otimes I_{p_2} &= (H_2 J_2) \otimes (H_1 + J_1) \\
&= (J_2 \otimes J_1) + (J_2 \otimes H_1) + (H_2 \otimes J_1) + (H_2 \otimes H_1)
\end{aligned}
$$

Definition 7.4. ***Two-Way Layout Projections***

$$
\begin{aligned}
P_0 &= J_2 \otimes J_1 \\
P_1 &= (J_2 \otimes H_1) \\
P_2 &= (H_2 \otimes J_1) \\
P_{12} &= (H_2 \otimes H_1)
\end{aligned}
$$

H-subscripts induce P subscripts.

Note:

1. P_0, P_1, P_2, P_{12} are each symmetric & idempotent
2. Thse 4 projections are mutually orthogonal. e.g. $P_1P_2 = P_1P_{12} = \cdots = 0$

Remark 7.5. ***Projection Form of the ANOVA Decomposition***

$$m = P_0m + P_1m + P_2m + P_{12}m$$

This is *equivalent* to the classical ANOVA decomposition because

$$\underbrace{P_0}_{p\times p}\underbrace{m}_{p\times 1} = \{m_{..},\ 1 \le i \le p_1\},\ 1 \le j \le p_2\} = (J_2 \otimes J_1)m$$
$$P_1m = \{\{m_{i.} - m_{..},\ 1 \le i \le p_1\},\ 1 \le j \le p_2\} = (J_2 \otimes H_1)m$$
$$P_2m = \{\{m_{.j} - m_{..},\ 1 \le i \le p_1\},\ 1 \le j \le p_2\} = (H_2 \otimes J_1)m$$
$$P_{12}m = \{\{m_{ij} - m_{i.} - m_{.j} + m_{..},\ 1 \le i \le p_1\},\ 1 \le j \le p_2\} = (H_2 \otimes H_1)m$$

This says

$$m_{ij} = m_{..} - (m_{i.} - m_{..}) + (m_{.j} - m_{..}) + (m_{ij} - m_{i.} - m_{.j} + m_{..}), \quad 1 \le i \le p_1,\ 1 \le j \le p_2$$

Method of Checking the four equations above

$$P_0m = (J_2 \otimes J_1)m = \mathrm{vec}(J_1MJ_2)$$
$$P_1m = (J_2 \otimes H_1)m = \mathrm{vec}(H_1MJ_2) = \mathrm{vec}[(I_{p_1} - J_1)MJ_2] = \mathrm{vec}(MJ_2) - \mathrm{vec}(J_1MJ_2)$$
$$P_2m = (H_2 \otimes J_1)m = \mathrm{vec}(J_1MH_2) = \mathrm{vec}[J_1M(I_{p_2} - J_2)] = \mathrm{vec}(J_1M) - \mathrm{vec}(J_1MJ_2)$$
$$P_{12}m = \mathrm{vec}(H_1MH_2) = \mathrm{vec}(M) - \mathrm{vec}(MJ_2) - \mathrm{vec}(J_1M) + \mathrm{vec}(J_1MJ_2)$$

Thus,

$$\mathrm{unvec}(P_0m) = \underbrace{J_1MJ_2}_{p_1\times p_2} = \begin{pmatrix} m_{..} & m_{..} & \cdots & m_{..} \\ \vdots & \vdots & \ddots & \dots \\ m_{..} & m_{..} & \cdots & m_{..} \end{pmatrix}$$
$$\mathrm{unvec}(P_1m) = MJ_2 - J_1MJ_2 = \begin{pmatrix} m_{1.} - m_{..} & \cdots & m_{1.} - m_{..} \\ \vdots & \ddots & \vdots \\ m_{p_1.} - m_{..} & \cdots & m_{p_1.} - m_{..} \end{pmatrix}$$

Remark 7.6. ***Some Standard ANOVA Submodels for 2-Way Layout***

Form is $m = \underbrace{Q}_{p\times p}\underbrace{\beta}_{p\times 1}$, $\beta \in \mathbb{R}^p$, Q symmetric & idempotent.

$Q = I_p = P_0 + P_1 + P_2 + P_{12}$	(General Model)
$Q = P_0 + P_1 + P_2 \leftrightarrow \{\gamma_{ij}\}$ all 0	(Additive Submodels)
$Q = P_0 + P_1 \leftrightarrow \{\beta_j\},\ \{\gamma_{ij}\}$ all 0	(Factor 1 effects)
$Q = P_0 + P_2 \leftrightarrow \{\alpha_i\},\ \{\gamma_{ij}\}$ all 0	(Factor 2 effects)
$Q = P_0 \leftrightarrow \{\alpha_i\}, \{\beta_j\}, \{\gamma_{ij}\}$ all 0	(No effects)

7.3 Least Squares Analysis

Model

$$\underbrace{y}_{n\times 1} = \underbrace{C}_{n\times p}\underbrace{m}_{p\times 1} + \underbrace{e}_{n\times 1},\ \mathrm{rank}(C) = p \le n$$

Note: It is *not* assumed that $n_{ij} = n_0,\ 1 \le i \le p_1,\ 1 \le j \le p_2$. Thus, $C'C \ne n_0 I_p$ for some n_0 (because $C'C = \mathrm{diag}\{n_{ij}\}$).

Under the general model: $y = Cm + e,\ m \in \mathbb{R}^p$. The LSEs of $m,\ \eta = Cm$ are

$$\hat{m} = C^+ y = (C'C)^{-1}C'y$$
$$\hat{\eta} = CC^+ y = C(C'C)^{-1}C'y$$

These have elementary forms, as in 1-way layout.

Under submodel (Q), the LSEs of m, η are

$$\hat{m}_0 = (CQ)^+ y$$
$$\hat{\eta}_0 = C(CQ)^+ y\ (= CQ(CQ)^+ y)$$

F-test
H: submodel Q holds
K: not so, general model holds

The test statistic is

$$T = \frac{|\hat{\eta} - \hat{\eta}_0|^2 / r_1}{\hat{\sigma}^2}$$

where $\hat{\sigma}^2 = \frac{1}{n-r}|y - \hat{\eta}|^2$, with $r = \mathrm{rank}(C) = p$ ($\leftarrow C$ is the design matrix in the general model).

$$r_1 = r - r_0 = p - \mathrm{tr}(Q)$$
$$r_0 = \mathrm{rank}(CQ) = \mathrm{rank}(Q) = \mathrm{tr}(Q)$$

where CQ is the design matrix in the submodel $y = CQ\beta + e$ and $n = \sum_{i=1}^{p_1}\sum_{j=1}^{p_2} n_{ij}$.

The Null distribution is $F_{p-\mathrm{tr}(Q),n-p}$.

Estimated Risks
of $\hat{\eta}$ and $\hat{\eta}_0$ as competing estimators of $\eta = Cm$ in the general model:

$$\hat{R}(\hat{\eta}) = \hat{\sigma}^2$$
$$\hat{R}(\hat{\eta}_0) = \underbrace{p^{-1}}_{r}[|y - \hat{\eta}_0|^2 + (2\underbrace{\mathrm{tr}(Q)}_{r_0} - n)\hat{\sigma}^2]$$

Note: Calculating $r_0 = \mathrm{tr}(Q)$ is easy algebraically in 2-way ANOVA models.

Example 7.7.

$Q = P_0 + P_1 + P_2$.

$$\begin{aligned}\operatorname{tr}(Q) &= \operatorname{tr}(P_0) + \operatorname{tr}(P_1) + \operatorname{tr}(P_2)\\ \operatorname{tr}(P_0) &= \operatorname{tr}(J_2 \otimes J_1) = \operatorname{tr}(J_2)\operatorname{tr}(J_1) = 1 \cdot 1 = 1\end{aligned}$$

where we have used that $\operatorname{tr}(J_k) = \operatorname{tr}(u_k u_k') = \operatorname{tr}(\underbrace{u_k' u_k}_{=1}) = 1$.

$$\begin{aligned}\operatorname{tr}(P_1) &= \operatorname{tr}(J_2 \otimes H_1) = \operatorname{tr}(J_2)\operatorname{tr}(H_1) = \operatorname{tr}(J_2)[\operatorname{tr}(I_p - J_1)]\\ &= \underbrace{\operatorname{tr}(J_2)}_{} = 1[\operatorname{tr}(I_{p_1}) - \underbrace{\operatorname{tr}(J_1)}_{=1}]\\ &= p_1 - 1\\ \operatorname{tr}(P_2) &= p_2 - 1 \text{ (by similar analysis)}\\ r_0 &= \operatorname{rank}(CQ) = \operatorname{rank}(Q) = \operatorname{tr}(Q) = 1 + (p_1 - 1) + (p_2 - 1)\end{aligned}$$

where we used that C is of full rank. For the F-test:

$$\begin{aligned}r_1 = r - r_0 &= \underbrace{p}_{p_1 p_2} - [1 + (p_1 - 1) + (p_2 - 1)] = p_1 p_2 - p_1 - p_2 + 1\\ &= (p_1 - 1)(p_2 - 1)\end{aligned}$$

The d.f. (degrees of freedom) are $(p_1 - 1)(p_2 - 1),\ n - p$.

Note: This analysis of 2-way layouts works *whatever* the $\{n_{ij}\}$ may be, i.e. in unbalanced complete layouts as well as in balanced complete layouts.

Definition 7.8. ***Complete Layout, Balanced Complete Layout, Unbalanced Complete Layout***

A *complete layout* has $n_{ij} \geq 1,\ 1 \leq i \leq p_1,\ 1 \leq j \leq p_2$.
A *balanced complete layout* has $n_{ij} = n_{\cdot} \geq 1\ \forall\ i, j$.
An *unbalanced complete layout* has $\{n_{ij}\}$ are not equal but $n_{ij} \geq 1\ \forall\ i, j$.

Comments:

- Balanced complete layout is classical & elementary
- Unbalanced complete layout is very difficult classically and not elementary

Recall:
Submodel $m = CQ\beta,\ \beta \in \mathbb{R}^p,\ Q$ symmetric & idempotent and of rank r_0.
$\updownarrow$
Submodel $m = CV\gamma,\ \gamma \in \mathbb{R}^{r_0},\ Q = \underbrace{V}_{p \times r_0}\underbrace{V'}_{r_0 \times p}$ is the spectral representation of Q.

Example 7.9. ***Spectral Forms of Projections for One-Way Layout***

Consider $P_0 = uu'$, $P_1 = I_p - uu' = I_p - P_0$, $\underbrace{u}_{p\times 1} = p^{-1/2}(1, 1, \ldots, 1)'$. Let $(\underbrace{u}_{p\times 1} \ \underbrace{U}_{p\times(p-1)})$ be an orthogonal matrix. This implies that $u'u + U'U = uu' + UU' = I_p$. The spectral representations of P_0 and P_1 are

$$P_0 = v_0 v_0', \quad \text{where } v_0 = u$$
$$P_1 = v_1 v_1', \quad \text{where } v_1 = U$$

Notes:

1. The columns of U are mutually orthogonal and are orthogonal to u, i.e. the columns of U are mutually orthogonal contrasts
2. Construction of U from Helmert contrasts:

$$\underbrace{H}_{p\times(p-1)} = \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ -1 & 1 & 1 & \cdots & 1 \\ 0 & -2 & 1 & \cdots & 1 \\ 0 & 0 & -3 & \cdots & 1 \\ 0 & 0 & 0 & \cdots & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & -(p-1) \end{pmatrix}$$

Define $U = \underbrace{H}_{p\times(p-1)} \left(\underbrace{H'H}_{\text{diagonal matrix}} \right)^{-1/2}$. Then

$$U'U = I_{p-1}$$
$$u'U = 0$$

So $(u \ U)$ is an orthogonal matrix.

3. Take U to have columns that are orthogonal polynomials of degrees 1 to $(p-1)$. poly$(\cdot) \rightarrow U$.

Remark 7.10. ***LSE's for the One-Way Layout Submodel***

Submodel Model Fit in One-Way Layout

$$\hat{m}_0 = V_0(CV_0)^+ y, \qquad V_0 = uu'$$
$$\hat{\eta}_0 = CV_0(CV_0)^+ y$$

Example 7.11. ***Two-Way Layout: Spectral Representations of*** P_0, P_1, P_2, P_{12}

Let $(\underbrace{u_k}_{p_k \times 1} \quad \underbrace{U_k}_{p_k \times (p_k - 1)})$ be orthogonal matrices, $k = 1, 2$, $u_k = p_k^{-1/2}(1, 1, \ldots, 1)'$. Then

$$I_p = \begin{pmatrix} u_k & U_k \end{pmatrix} \begin{pmatrix} u_k' \\ U_k' \end{pmatrix} = \underbrace{u_k u_k'}_{J_k} + \underbrace{U_k U_k'}_{H_k} = J_k + H_k.$$

Then we get the spectral representation

$$\begin{aligned}
P_0 &= J_2 \otimes J_1 = u_2 u_2' = V_0 V_0' \quad \text{for } \underbrace{V_0}_{p \times 1} = \underbrace{u_2}_{p_2 \times 1} \otimes \underbrace{u_1}_{p_1 \times 1} \\
P_1 &= J_2 \otimes H_1 = u_2 u_2' \otimes U_1 U_1' = V_1 V_1' \quad \text{for } V_1 = u_2 \otimes U_1 \\
P_2 &= V_2 V_2' \quad \text{for } V_2 = U_2 \otimes u_1 \\
P_{12} &= H_2 \otimes H_1 = U_2 U_2' \otimes U_1 U_1' = V_{12} V_{12}' \quad \text{for } V_{12} = U_2 \otimes U_1'
\end{aligned}$$

These yield spectral representations for standard submodels Q in the 2-way layout.

Example 7.12.

$$\begin{aligned}
Q &= P_0 + P_1 + P_2 = V_0 V_0' + V_1 V_1' + V_2 V_2' \\
&= V V' \qquad \text{for } V = \begin{pmatrix} V_0 & V_1 & V_2 \end{pmatrix}
\end{aligned}$$

and the columns of V are mutually orthogonal.

7.4 Review for Midterm

- Lab 1 material
- Existence and uniqueness of LSEs
- Algebraic stuff

8 10-18-11

8.1 Midterm Info

- Closed book. Can bring 2 double-sided sheets of notes.
- Coverage:
  - SVD, Moore-Penrose pseudoinverse
  - Least squares: normal equation, solution sets, distribution theory, linear estimability, optimality properties (Gauss-Markov, Scheffe), polynomial regression
  - Cutoff is just before F-test or estimated risk
  - Study labs 1 & 2
- 3 problems, relatively short, each with a nice solution

8.2 r & r_0

General Model: $y = \underbrace{C}_{n\times p}\underbrace{m}_{p\times 1} + e$

Submodel Q: $m = \underbrace{Q}_{p\times p}\underbrace{\beta}_{p\times 1},\ \beta \in \mathbb{R}^p$, where Q is symmetric & idempotent.

$$y = \underbrace{CQ\beta}_{X_0} + e$$

So

$$r = \operatorname{rank}(X) = \operatorname{rank}(C)$$
$$r_0 = \operatorname{rank}(X_0) = \operatorname{rank}(CQ)$$

Useful for complete layouts
When $\underbrace{C}_{n\times p}$ is of full rank, $\operatorname{rank}(C) = p$, then

$$r = p$$
$$r_0 = \operatorname{rank}(CQ) = \operatorname{rank}(Q) = \operatorname{tr}(Q)$$

8.3 Spectral Representations of P_0, P_1, P_2, P_{12}

Let $O = (\underbrace{u_k}_{p_k\times 1}\ \underbrace{U_k}_{p_k\times(p_k-1)})$, $k = 1, 2$, be a $p \times p$ orthogonal matrix: $O'O = OO' = I_{p_k} \Leftrightarrow O^{-1} = O'$. Then

$$I_{p_k} = \begin{pmatrix} u_k & U_k \end{pmatrix}\begin{pmatrix} u_k' \\ U_k' \end{pmatrix} = \underbrace{u_k u_k'}_{J_k} + \underbrace{U_k U_k'}_{I_{p_k}-J_k} = J_k + H_k$$

The columns of U_k are mutually orthonormal contrasts.

$$P_0 = J_2 \otimes J_1 = u_2u_2' \otimes u_1u_1' = V_0V_0'$$

for $\underbrace{V_0}_{p\times 1} = \underbrace{u_2}_{p_2\times 1} \otimes \underbrace{u_1}_{p_1\times 1},\ p = p_1p_2$. Similarly,

$$P_1 = J_2 \otimes H_1 = u_2u_2' \otimes U_1U_1' = V_1V_1'$$

for $\underbrace{V_1}_{p\times(p_1-1)} = u_2 \otimes U_1$.

$$P_2 = H_2 \otimes J_1 = U_2U_2' \otimes u_1u_1' = V_2V_2'$$

for $\underbrace{V_2}_{p\times(p_2-1)} = U_2 \otimes u_1$.

$$P_{12} = H_2 \otimes H_1 = U_2U_2' \otimes U_1U_1' = V_{12}V_{12}'$$

for $\underbrace{V_{12}}_{p\times(p_1-1)(p_2-1)} = U_2 \otimes U_1$.

Moreover, $(V_0\ V_1\ V_2\ V_{12})$ is an orthogonal matrix.

Example 8.1.

Suppose $Q = P_0+P_1+P_2 = V_0V_0'+V_1V_1'+V_2V_2' = VV'$ for $V = (V_0\ V_1\ V_2)$ = spectral representation of Q. Note that $V'V = I$. Hence,

$$m = CQ\beta,\ \beta \in \mathbb{R}^p \quad \leftrightarrow \quad m = CV\gamma,\ \gamma \in \mathbb{R}^{p_1+p_2-1}$$

In the case that $\text{rank}(C) = p$, the LSEs for model Q are

$$\hat{m}_0 = (CQ)^+y = V(CV)^+y$$
$$\hat{\eta}_0 = C\hat{m}_0 = C(CQ)^+y = CV(CV)^+y$$

8.4 Special Case: Balanced Complete Design

General model: $y = \underbrace{C}_{n\times p}\underbrace{m}_{p\times 1}+e$, $\text{rank}(C) = p$ (because it is complete), $C'C = n_0I_p$ (because it is balanced $\Leftrightarrow$ same number of observations for each mean). The LSEs are

$$\hat{m} = (C'C)^{-1}C'y = n_0^{-1}C'y$$
$$\hat{\eta} = C\hat{m} = n_0^{-1}CC'y$$

Submodel Q: $y = CQ\beta + e$ for $\beta \in \mathbb{R}^p$, i.e. $m = Q\beta$. The LSEs are

$$\hat{m}_0 = (CQ)^+y$$
$$\hat{\eta}_0 = C(CQ)^+y$$

Theorem 8.2.

Suppose $C'C = n_0I_p,\ n_0 \geq 1$ (i.e. complete balanced design). Then

$$\hat{m}_0 = n_0^{-1}QC'y = Q\hat{m}$$
$$\hat{\eta}_0 = n_0^{-1}CQC' = CQ\hat{m}$$

Proof. $\underbrace{Q}_{p\times p} = \underbrace{V}_{p\times r_0}\underbrace{V'}_{r_0\times p}$ spectral representation. $r_0 = \text{rank}(Q) = \text{tr}(Q)$.

$$
\begin{aligned}
(CQ)^+ &= (C\underbrace{VV'}_{Q})^+ = V(CV)^+ \qquad \text{Lab 1, 1e} \\
(CV)^+ &= (V'\underbrace{C'CV}_{n_0 I_p})^+ V'C' \qquad \text{Lab 1} \\
&= (n_0\underbrace{V'V}_{I_{r_0}})^+ V'C' = n_0^{-1} I_{r_0} V'C' = n_0^{-1} V'C' \\
(CQ)^+ &= V(CV)^+ = n_0^{-1}\underbrace{VV'}_{Q}C' = n_0^{-1}QC'
\end{aligned}
$$

Thus,

$$
\begin{aligned}
\hat{m}_0 &= n_0^{-1}QC'y = Q\hat{m} \qquad \text{because } \hat{m} = n_0^{-1}C'y \\
\hat{\eta}_0 &= C\hat{m}_0 = CQ\hat{m}
\end{aligned}
$$

□

Note: This explains nice formulas for balanced, complete designs.

8.5 Three-Way Layouts - Complete Layout

General model in classical form: $y_{ijkl} = m_{ijk} + e_{ijkl}$. Factor levels:

$$1 \le i \le p_1,\ 1 \le j \le p_2,\ 1 \le k \le p_3$$

Replications are labeled by

$$1 \le l \le n_{ijk}$$

Vectorize:

$$\underbrace{m}_{p\times 1} = \{\{\{m_{ijk} \mid 1 \le i \le p_1\},\ 1 \le j \le p_2\},\ 1 \le k \le p_3\}$$

This is called *mirror dictionary order* or *array order*.

$$
\begin{aligned}
y &= \{\{\{\{y_{ijkl} \mid 1 \le l \le n_{ijk}\},\ 1 \le i \le p_1\},\ 1 \le j \le p_2\},\ 1 \le k \le p_3\} \\
n &= \sum_{i=1}^{p_1}\sum_{j=1}^{p_2}\sum_{k=1}^{p_3} n_{ijk}
\end{aligned}
$$

e is defined similar to y.

$$C = \text{ data-incidence matrix of 1's and 0's}$$

8.5.1 Simple ANOVA Decomposition of Means

Classical Form:

$$m_{ijk} = \mu + \alpha_i^{(1)} + \alpha_j^{(2)} + \alpha_k^{(3)} + \alpha_{ij}^{(12)} + \alpha_{ik}^{(13)} + \alpha_{jk}^{(23)} + \alpha_{ijk}^{(123)}$$

where

$$
\begin{aligned}
&\alpha_{\cdot}^{(1)} = \alpha_{\cdot}^{(2)} = \alpha_{\cdot}^{(3)} = 0 \\
&\alpha_{i\cdot}^{(12)} = \alpha_{\cdot j}^{(12)} = \alpha_{\cdot k}^{(13)} = \alpha_{\cdot k}^{(23)} = \alpha_{j\cdot}^{(23)} = 0 \\
&\alpha_{ij\cdot}^{(123)} = \alpha_{i\cdot k}^{(123)} = \alpha_{\cdot kj}^{(123)} = 0
\end{aligned}
$$

This form gives a one-to-one mapping of $\{m_{ijk}\}$ into $\mu,\ \{\alpha_i^{(1)}\},\ \{\alpha_j^{(2)}\},\ \ldots$

$$\mu = m_{...},\ \alpha_i^{(1)} = m_{i..} - m_{...},\ \alpha_j^{(2)} = m_{.j.} - m_{...}$$
$$\alpha_{ij}^{(12)} = m_{ij.} - m_{i..} - m_{.j.} + m_{...}$$
$$\alpha_{jk}^{(12)} = m_{.jk} - m_{.j.} - m_{..k} + m_{...}$$
$$\alpha_{ik}^{(13)} = m_{i.k} - m_{i..} - m_{..k} + m_{...}$$
$$\alpha_{ijk}^{(123)} = m_{ijk} - m_{ij.} - m_{i.k} - m_{.jk} + m_{i..} + m_{.j.} + m_{..k} - m_{...}$$

9 10-20-11

9.1 Projection Form of ANOVA Decomposition for 3-Way Layout

As earlier, let $u_k = p_k^{-1/2}(1, 1, \ldots, 1)'$, $J_k = u_k u_k'$, $H_k = I_{p_k} - J_k$, $1 \le k \le 3$, $p = p_1 p_2 p_3 =$ # of means $\{m_{ijk}\}$. Let

$$\begin{aligned} P_0 &= J_3 \otimes J_2 \otimes J_1 \\ P_1 &= J_3 \otimes J_2 \otimes H_1 \\ P_2 &= J_3 \otimes H_2 \otimes J_1 \\ P_3 &= H_3 \otimes J_2 \otimes J_1 \\ P_{12} &= J_3 \otimes H_2 \otimes H_1 \\ P_{13} &= H_3 \otimes J_2 \otimes H_1 \\ P_{23} &= H_3 \otimes H_2 \otimes J_1 \\ P_{123} &= H_3 \otimes H_2 \otimes H_1 \end{aligned}$$

The subscripts on the LHS indicate the H-factors on the RHS.

Note:

1. These P's are symmetric & idempotent, and they are mutually orthogonal
2. $I_p = I_{p_3} \otimes I_{p_2} \otimes I_{p_1} = (J_3 + H_3) \otimes (J_2 + H_2) \otimes (J_1 + H_1) = P_0 + P_1 + P_2 + P_3 + P_{12} + P_{13} + P_{23} + P_{123}$. The ANOVA decomposition of m in the 3-way layout:

$$m = \underbrace{P_0 m}_{\text{overall mean}} + \underbrace{P_1 m + P_2 m + P_3 m}_{\text{main effects}} + \underbrace{P_{12} m + P_{13} m + P_{23} m}_{\text{2-way interactions}} + \underbrace{P_{123} m}_{\text{3-way interactions}}$$

9.2 Standard ANOVA Submodels

The form is: $\underbrace{m}_{p \times 1} = \underbrace{Q}_{p \times p} \beta$, $\beta \in \mathbb{R}^p$, Q is an orthogonal projection.

- $Q = I_p$ (general model) $\Rightarrow 1$
- $Q = P_0 + P_1 + P_2 + P_3 + P_{12} + P_{13} + P_{23}$ (no 3-way interactions) $\Rightarrow 1$
- $Q = P_0 + P_1 + P_2 + P_3 + P_{12} + P_{13}$ (no 3-way and no 2-3-way interactions) $\Rightarrow 1$
  - $+$ two more 2×2-way, no 3-way interactions $\Rightarrow 2$
- $Q = P_0 + P_1 + P_2 + P_3 + P_{12}$ (no 3-way interactions, only the 1-2 2-way interaction) $\Rightarrow 3$
- $Q = P_0 + P_1 + P_2 + P_3$ (additive) $\Rightarrow 1$
- $Q = P_0 + P_1 + P_2$ (+ 2 more) (subadditive, no factor 3 effect) $\Rightarrow 3$
- $Q = P_0 + P_1$ (+ 2 more) (only factor 1 matters) $\Rightarrow 3$
- $Q = P_0$ (no effects) $\Rightarrow 1$

Total = 16 submodels (more exist...)

9.3 LSE's under the General Model and Submodel

Under the general model, $y = \underbrace{C}_{n\times p}\underbrace{m}_{p\times 1} + e$, the LSE's are:

$$\hat{m} = C^+ y = (C'C)^{-1}C'y = \{m_{ijk.}\} \text{ in mirror dictionary order}$$
$$\hat{\eta} = CC^+ y = C\hat{m} = C(C'C)^{-1}C'y$$

We assume that it is a complete layout (1 observation for every cell, i.e. $n_{ijk} \geq 1$), so

$$r = \text{rank}(C) = p$$

Under submodel Q, the model is $y = CQ\beta + e,\ \beta \in \mathbb{R}^p,\ m = Q\beta$. The LSE's are

$$\hat{m}_0 = (CQ)^+ y$$
$$\hat{\eta}_0 = CQ(CQ)^+ y \overset{\text{Lab 1}}{=} C(CQ)^+ y$$

Note: $\eta_0 = Cm_0 = \mathbb{E}y$ under the submodel. Thus, $m_0 = (C'C)^{-1}C'\eta_0$, so m_0 is linearly estimable.

$$r_0 = \text{rank}(CQ) = \text{rank}(Q) = \text{tr}(Q)$$

Special Case: *Balanced* Complete Design (i.e. $C'C = n_0 I_p,\ n_0 \geq 1$)
As for the 2-way layout,

$$\hat{m}_0 = Q\hat{m}$$
$$\hat{m} = (C'C)^{-1}C'y$$
$$\hat{\eta}_0 = CQ\hat{m}$$

9.4 Spectral Representations for the Projections $P_0, P_1, P_2, \ldots, P_{123}$

Let $(u_k\ U_k)$ be a $p_k \times p_k$ orthogonal matrix. (The columns of U_k are orthonormal contrasts.)

$$I_{p_k} = \begin{pmatrix} u_k & U_k \end{pmatrix}\begin{pmatrix} u_k' \\ U_k' \end{pmatrix} = \underbrace{u_k u_k'}_{J_k} + \underbrace{U_k U_k'}_{H_k}, \qquad k = 1,2,3$$

$$J_k = u_k u_k' \qquad \text{by definition}$$
$$H_k = I_p - J_k = U_k U_k'$$
$$P_0 = J_3 \otimes J_2 \otimes J_1 = V_0 V_0' \quad \text{with } V_0 = u_3 \otimes u_2 \otimes u_1$$
$$P_1 = J_3 \otimes J_2 \otimes H_1 = V_1 V_1' \quad \text{with } V_1 = u_3 \otimes u_2 \otimes U_1$$
$$\text{similarly for } P_2, P_3$$
$$P_{12} = J_3 \otimes H_2 \otimes H_1 = V_{12} V_{12}' \quad \text{with } V_{12} = u_3 \otimes U_2 \otimes U_1$$
$$\text{similarly for } P_{13}, P_{23}$$
$$P_{123} = H_3 \otimes H_2 \otimes H_1 = V_{123} V_{123}' \quad \text{with } V_{123} = U_3 \otimes U_2 \otimes U_1$$

Note:

1. The matrix
$$\underbrace{V}_{p\times p} = \begin{pmatrix} V_0 & V_1 & V_2 & V_3 & V_{12} & V_{13} & V_{23} & V_{123} \end{pmatrix}$$
is orthogonal: $V'V = VV' = I_p$.

2. Submodel Q can be expressed in terms of V. For example, let $Q = P_0 + P_1 + P_{12} + P_{13} = V_Q V_Q'$, where $V_1 = (V_0\ V_1\ V_{12}\ V_{13})$. This is a spectral representation of Q.

9.5 Other Projection Decompositions

The ANOVA projections are designed for *nominal covariates*: the levels (values) of the covariates are only labels.

For *ordinal covariates* (the values are real numbers whose order and particular values matter), we may want other projections.

Example 9.1. ***Ordinate Covariates Example: vineyard.dat***

Vineyard data in Lab #2.

$$y_{ij} = m_{ij} + e_{ij},\ 1 \le i \le 52 = p_1,\ 1 \le j \le 3 = p_2$$

i labels the vineyard row ($\leftarrow$ ordinal covariates).
j labels the year (‘ nominal covariates).
y_{ij} = the year from row i in year j.

We conjecture that m_{ij} varies “slowly” in i for a given j. E.g. polynomial regression for each year?

We can replace ANOVA projections by alternative projections that implement polynomial regression as follows:

1. Let $\underbrace{U_1(d)}_{p_1 \times d}$ = orthogonal polynomial of degree $d-1$.
2. Define $\underbrace{J_1(d)}_{p_1 \times p_1} = U_1(d)U_1'(d),\ \underbrace{H_1(d)}_{p_1 \times p_1} = I_{p_1} - U_1(d)$.
3. Define $J_2 = u_2 u_2'$ (where $u_2 = p_2^{-1/2}(1, 1, \ldots, 1)'$), $H_2 = I_{p_2} - J_2$ as earlier.
4. Then $I_p = I_{p_2} \otimes I_{p_1} = P_0(d) + P_1(d) + P_2(d) + P_{12}(d)$

$$\begin{aligned} P_0(d) &= J_2 \otimes J_1(d) \\ P_1(d) &= J_2 \otimes H_1(d) \\ P_2(d) &= H_2 \otimes J_1(d) \\ P_{12}(d) &= H_2 \otimes H_1(d) \\ m &= P_0(d)m + P_1(d)m + P_2(d)m + P_{12}(d)m \end{aligned}$$

This gives a projection decomposition of m that extends the ANOVA decomposition $\leftrightarrow\ d = 1$.

Idea: Extend the ANOVA fits to $d = 2, 3, \ldots, 6$. Compare these least squares fits via estimated risk.

9.6 Incomplete Designs

Of interest is the mean vector $\underbrace{m}_{p \times 1}$. We have one or more observations only on a subvector $\underbrace{m_D}_{q \times 1}$ of m, where $q \le p$. We get this subvector by deleting components of m. Thus, $\underbrace{m_D}_{q \times 1} = \underbrace{D}_{q \times p}\ \underbrace{m}_{p \times 1}$.

Note: D is obtained by deleting suitable rows of I_p. The rows of D are therefore orthonormal: $\underbrace{D}_{q\times p}\underbrace{D'}_{p\times q} = I_q$. We call D the *deletion matrix*.

Example 9.2. ***Deletion Matrix Example***

$$m = \begin{pmatrix} m_1 \\ m_2 \\ m_3 \\ m_4 \end{pmatrix}, \qquad m_D = \begin{pmatrix} m_1 \\ m_4 \end{pmatrix}, \qquad D_2 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \qquad p = 4,\ q = 2$$

Note that D is I_4 with rows 2 and 3 deleted.

$$DD' = I_2, \qquad m_D = Dm$$

Remark 9.3. ***General Model for the Incomplete Design***

$$\begin{aligned} \underbrace{y}_{n\times 1} &= \underbrace{C}_{n\times q}\underbrace{m_D}_{q\times 1} + e, \qquad \operatorname{rank}(C) = q = \operatorname{rank}(D) = \operatorname{rank}(CD) = r \\ &= CDm + e \\ \hat{m}_D &= C^+ y = (C'C)^{-1}C'y \\ \hat{\eta} &= C(C'C)^{-1}C'y = CC^+y \\ \hat{m} &= (CD)^+y + [I_p - CD(CD)^+]c,\ c \in \mathbb{R}^p \end{aligned}$$

Question: How do we make an intelligent choice of c?
Question: $D\hat{m} \stackrel{?}{=} \hat{m}_D$. Yes.

- Lab 1.l: $(CD)^+ = D'C^+$ because $DD' = I_q$.
- Hence

$$\begin{aligned} D\hat{m} &= D(CD)^+y + [D - D(CD)^+CD]c \\ &= \underbrace{DD'}_{I_q}C^+y + [D - \underbrace{DD'}_{I_q}C^+CD]c \\ &= C^+y + [D - \underbrace{(C'C)^{-1}C'CD}_{I}]c \\ &= C^+y = \hat{m}_D \end{aligned}$$

9.7 Submodel Q for m_D

If m_D is a subvector of m, which is a vectorized array of means, submodel specification is often obscure.

Proposed Idea: Start with the submodel Q for m: $m = Q\beta,\ \beta \in \mathbb{R}^p,\ Q$ is symmetric & idempotent. This

implies that $m_D = DQ\beta,\ \beta \in \mathbb{R}^p$. So the model is $y = \underbrace{C}_{n\times q}\underbrace{D}_{q\times p}\underbrace{Q}_{p\times p}\underbrace{\beta}_{p\times 1} + e$.

Let $\hat{m}_{D,0}$ = LSE of m_D under the submodel. Then

$$\hat{m}_{D,0} = DQ(CDQ)^+ y = D(CDQ)^+ y$$

Let $\hat{\eta}$ = LSE of $\eta = CDQ\beta$ under the submodel. Then

$$\hat{\eta} = CDQ(CDQ)^+ y = CD(CDQ)^+ y = \text{ unique}$$

Because $m_D = (C'C)^{-1}C'\eta_0$, it is linearly estimable.

9.8 Midterm Comments

- Recommended: have Lab 1 results on cheat sheets, also major results from in class
  - Can cite any lab 1 results or results we proved in class
- 3 problems, about an hour's worth of work
- <u>Material Covered</u>
  - SVD, Moore-Penrose pseudoinverse
  - Solutions to consistent linear equations (including how to test for consistency)
  - Least squares (whatever the rank of X $\Rightarrow$ normal equation and its solutions)
  - Linear parametric functions of β ($y = X\beta + e$)
    - Linear estimability $\leftrightarrow$ unique LSE's of linear parametric functions
    - $\Psi = \lambda'\beta$
      1. Is $\hat{\Psi} = \lambda'\hat{\beta}$ unique for every LSE $\hat{\beta}$?
      2. Link to linear unbiased estimators of Ψ
         - Boils down to 2 theorems we wrote down: uniqueness and something about $\lambda'\hat{\beta}$ being unique
  - Gauss-Markov/Lehmann-Scheffé theorems
- The algebra is simple, but it requires understanding
- 2 problems are statistical in nature, the 3rd is algebraic

10 10-27-11

10.1 General Model for Incomplete Design

$$\underbrace{y}_{n\times 1} = \underbrace{C}_{n\times q}\underbrace{m_D}_{q\times 1} + e = \underbrace{C}_{n\times q}\underbrace{D}_{q\times p}\underbrace{m}_{p\times 1} + e$$
$$\text{rank}(C) = q \le p \le n$$
$$D = \text{ depletion matrix}$$
$$DD' = I_q$$

Unique LSE's

- The LSE of η is $\hat{\eta} = CC^+y = C(C'C)^{-1}C'y$
  - $r = \text{rank}(CD) = \text{rank}(D) = q$
- The LSE of m_D is $\hat{m}_D = (C'C)^{-1}C'y$.

Not Unique LSE

- The LSEs of m are $\hat{m} = (CD)^+y + [I_p - (CD)^+(CD)]c$, where $c \in \mathbb{R}^p$

10.2 Submodels for the Incomplete Design

Start with the submodel $\underbrace{m}_{p\times 1} = \underbrace{Q}_{p\times p}\underbrace{\beta}_{p\times 1}$ (where Q is symmetric & idempotent) for the associated complete design. This induces submodel Q for m_D:

$$m_{D,0} = DQ\beta, \qquad \beta \in \mathbb{R}^p.$$

$$y = C\underbrace{DQ\beta}_{m_{D,0}} + e$$

The LSE of $\eta_0 = \mathbb{E}(y)$ in this submodel is

$$\hat{\eta}_0 = CDQ(CDQ)^+y = CD(CDQ)^+y \qquad \text{Lab 1, 1.m}$$
$$r_0 = \text{rank}(CDQ) = \text{rank}(DQ)$$

The LSE of $m_{D,0}$ is

$$\hat{m}_{D,0} = (C'C)^{-1}C'\hat{\eta}_0 = DQ(CDQ)^+y = D(CDQ)^+y$$

because $m_{D,0} = DQ\beta = (C'C)^{-1}C'\eta_0$; i.e. $m_{D,0}$ is linearly estimable.

10.3 Balanced Incomplete Design

Here, in addition, $\underbrace{C'}_{q\times p}\underbrace{C}_{p\times q} = n_0 I_q,\ n_0 \ge 1$.

Simplifications:

$$CDQ(CDQ)^+ = CDQ(QD'\underbrace{C'C}_{n_0 I_p}DQ)^+QD'C'$$
$$= n_0^{-1}CD(QD'DQ)^+D'C'$$

Let

$$\begin{aligned}\hat{m}_D &= (C'C)^{-1}C'y = \text{ general model LSE of } m_D \\ &= n_0^{-1}C'y\end{aligned}$$

Hence, in the submodel Q

$$\begin{aligned}\hat{\eta}_0 &= CDQ(CDQ)^+y \\ &= n_0^{-1}CD(QD'DQ)^+D'C'y \\ &= CD(QD'DQ)^+D'\hat{m}_D \\ &= CDQ(DQ)^+\hat{m}_D\end{aligned}$$

i.e.

$$\begin{aligned}\hat{\eta}_0 &= CDQ(DQ)^+\hat{m}_D \\ \hat{m}_{D,0} &= (C'C)^{-1}C'\hat{\eta}_0 = DQ(DQ)^+\hat{m}_D\end{aligned}$$

10.4 Interpolating Among Submodel Fits in Complete Balanced Designs

General model:

$$\underbrace{y}_{n\times 1} = \underbrace{C}_{n\times p}\underbrace{m}_{p\times 1} + \underbrace{e}_{n\times 1}, \qquad C'C = n_0 I_p,\ n_0 \geq 1$$

Consider the projection decomposition

$$\sum_{k=1}^{s} P_k = I_p$$

where $\{P_k\}$ are symmetric and idempotent: $P_kP_j = 0$ if $j \neq k$. For example, the ANOVA decomposition.

Let $d_q \subset \{1, 2, \ldots, s\}$. Let $Q = \sum_{k\in d_Q} P_k$. Q is symmetric & idempotent. The submodel $m = Q\beta$ has LSE's

$$\begin{aligned}\hat{m}_0 &= Q\hat{m} \\ \hat{\eta}_0 &= CD\hat{m} \\ \hat{m} &= n_0^{-1}C'y = \text{ general model LSE of } m.\end{aligned}$$

Note: $\hat{m}_D = \sum_{k\in d_Q} P_k\hat{m}$. Thus, $\hat{m}_D = \sum_{k=1}^{s} a_kP_k\hat{m}$, where

$$a_k = \begin{cases} 1 & k \in d_Q \\ 0 & k \notin d_Q \end{cases}$$

New idea: Consider the class of multiple shrinkage estimators

$$\hat{m}(a) = \sum_{k=1}^{s} a_kP_k\hat{m}$$

for m in the general model $y = Cm + e,\ C'C = n_0I_p$ (where $a = (a_1, a_2, \ldots, a_s),\ a_k \in [0, 1]$).

Note: In this discussion, the general model is balanced and complete.

Aim: Choose the $\{a_k\}$ to minimize risk and estimated risk of $\hat{\eta}(a) = C\hat{m}(a)$.

Formally:
General model $y = Cm + e,\ C'C = n_0 I_p$ (balanced, complete design).
Strong Gauss-Markov model on e: the components $\{e_i\}$ are i.i.d. with $\mathbb{E}(e_i) = 0,\ \mathrm{Var}(e_i) = \sigma^2,\ 0 < \sigma^2 < \infty$.

Definition 10.1. ***Quadratic Risk***

The *quadratic risk* of any estimator $\tilde{\eta}$ for η is

$$R(\tilde{\eta}, \eta, \sigma^2) = p^{-1}\mathbb{E}|\tilde{\eta} - n|^2$$

Remark 10.2. ***References***

- Stein (1966)
- Beran (2008) AISM

We will calculate the risk of $\hat{\eta}(a) = C\hat{m}(a)$ and then minimize it by choice of $a = (a_1, a_2, \ldots, a_s)$.

Theorem 10.3.

$$R(\hat{\eta}(a), \eta, \sigma^2) = \sum_{k=1}^{s} r(a_k, \tau_k, w_k)$$

where

$$\begin{aligned}
\tau_k &= p^{-1}\sigma^2 \,\mathrm{tr}(P_k) \\
w_k &= p^{-1}n_0|P_k m|^2 \\
r(a_k, \tau_k, w_k) &= \tau_k a_k^2 + (1 - a_k)^2 w_k = (a_k - \tilde{a}_k)^2(\tau_k + w_k) + \tau_k \tilde{a}_k \\
\tilde{a}_k &= \frac{w_k}{\tau_k + w_k}
\end{aligned}$$

Proof.

$$
\begin{aligned}
R(\hat{\eta}(a), \eta, \sigma^2) &= p^{-1}\mathbb{E}|\hat{\eta}(a) - n|^2 \\
\mathbb{E}|\hat{\eta}(a) - \eta|^2 &= |C(\hat{m}(a) - m)|^2 = (\hat{m}(a) - m)' \underbrace{C'C}_{=n_0 I_p}(\hat{m}(a) - m) \\
&= n_0|\hat{m}(a) - m(a)|^2 = n_0 \left|\sum_{k=1}^{s}(a_k P_k \hat{m} - P_k m)\right|^2 \\
&= n_0 \sum_{k=1}^{s} |a_k P_k \hat{m} - P_k m|^2 \\
&= n_0 \sum_{k=1}^{s} |a_k P_k(\hat{m} - m) - (1 - a_k)P_k m|^2 \\
R(\hat{\eta}(a), \eta, \sigma^2) &= p^{-1}\eta_0 \sum_{k=1}^{s} \mathbb{E}|a_k(\hat{m} - m) - (1 - a_k)P_k m|^2 \\
&= p^{-1}n_0 \sum_{k=1}^{s} \mathbb{E}\,\mathrm{tr}\left[\{a_k P_k(\hat{m} - m) - (1 - a_k)P_k m\}\{a_k P_k(\hat{m} - m) - (1 - a_k)P_k m\}'\right] \\
&= p^{-1}n_0 \sum_{k=1}^{s} \mathrm{tr}\,\mathbb{E}\left[\{a_k P_k(\hat{m} - m) - (1 - a_k)P_k m\}\{a_k P_k(\hat{m} - m) - (1 - a_k)P_k m\}'\right] \\
&= p^{-1}n_0 \sum_{k=1}^{s} \mathrm{tr}\left[a_k^2 P_k \left(\frac{\sigma^2}{n_0} I_p\right) P_k + (1 - a_k)^2 P_k m (P_k m)'\right] \\
&= \sum_{k=1}^{s} \underbrace{[\tau_k a_k^2 + (1 - a_k)^2 w_k]}_{r(a_k, \tau_k, w_k)}
\end{aligned}
$$

where we got the last line by using

$$
\begin{aligned}
\mathbb{E}(\hat{m}) &= m \\
\mathbb{E}\left[(\hat{m} - m)(\hat{m} - m)'\right] &= \mathrm{Cov}(\hat{m}) = \left(\frac{\sigma^2}{n_0}\right) I_p
\end{aligned}
$$

To complete the argument:

$$
\begin{aligned}
\tilde{a}_k &= \frac{w_k}{\tau_k + w_k} \\
(a_k - \tilde{a}_k)^2(\tau_k + w_k + \tau_k \tilde{a}_k &= a_k^2(\tau_k + w_k) - 2a_k \tilde{a}_k(\tau_k + w_k) + \tilde{a}_k^2(\tau_k + w_k) + \tau_k \tilde{a}_k \\
&= \tau_k a_k^2 + a_k^2 w_k - 2a_k w_k + \frac{w_k^2}{\tau_k + w_k} + \frac{\tau_k w_k}{\tau_k + w_k} \\
&= \tau_k a_k^2 + \underbrace{a_k^2 w_k - 2a_k w_k + w_k}_{\text{quadratic}} \\
&= \tau_k a_k^2 + (1 - a_k)^2 w_k
\end{aligned}
$$

□

10.5 Oracle Estimation

Oracle Estimation (not fully realizable)

Definition 10.4. ***Oracle Shrinkage Estimator***

The *oracle shrinkage estimator* $\tilde{m}_{\text{shr}}$ is the candidate shrinkage estimator $\hat{m}(a)$ that minimizes risk over all $a \in [0,1]^s$.

Theorem 10.5.

$$\tilde{m}_{\text{shr}} = \sum_{k=1}^{s} \tilde{a}_k P_k \hat{m} = \sum_{k=1}^{s} \left(\frac{w_k}{\tau_k + w_k} \right) P_k \hat{m}$$

where $\hat{m} = n_0^{-1} C' y$ is the LSE of m in the general model. Moreover, for $\tilde{\eta}_{\text{shr}} = C\tilde{m}_{\text{shr}}$, the risk is

$$R(\tilde{\eta}_{\text{shr}}, \eta, \sigma^2) = \sum_{k=1}^{s} \tau_k \tilde{a}_k = \sum_{k=1}^{s} \left(\frac{\tau_k w_k}{\tau_k + w_k} \right)$$

Proof. Taking $a_k = \tilde{a}_k$ minimizes $r(a_k, \tau_k, w_k)$ (i.e. the summand). □

Definition 10.6. ***Oracle Projection Estimator***

The *oracle projection estimator* $\tilde{m}_{\text{pro}}$ is the candidate shrinkage estimator $\hat{m}(a)$ that minimizes risk over all $a \in \{0,1\}^s \Leftrightarrow$ each $a_k = 0$ or 1.

If not unique, pick the one with the smallest $\{a_k\}$.

Note: This identifies the submodel fit(s) that minimizes risk.

Theorem 10.7.

$$\tilde{m}_{\text{pro}} = \sum_{k:\tilde{a}_k > 1/2}^{s} P_k \hat{m} = \sum_{k:w_k > \tau_k} P_k \hat{m}$$

$$R(\tilde{\eta}_{\text{pro}}, \eta, \sigma^2) = \sum_{k=1}^{s} \min\{\tau_k, w_k\}$$

Proof.

$$R(\hat{\eta}(a), \eta, \sigma^2) = \sum_{k=1}^{s} r(a_k, \tau_k, w_k)$$

$$r(a_k, \tau_k, w_k) = (a_k - \tilde{a}_k)^2 (\tau_k + w_k) + \tau_k \tilde{a}_k$$

The minimizing choice of $a_k = 0$ or 1 is

$$a_k = 1 \quad \text{if } \tilde{a}_k > \frac{1}{2}$$
$$a_k = 1 \quad \text{if } \tilde{a}_k \leq \frac{1}{2}$$

Either choice will do if $\tilde{a}_k = \frac{1}{2}$, but we take $a_k = 0$ by convention to simplify the summation. This gives the first form of $\tilde{m}_{\text{pro}}$.

Next, $\tilde{a}_k > \frac{1}{2} \Leftrightarrow \frac{w_k}{\tau_k + w_k} > \frac{1}{2} \Leftrightarrow w_k > \tau_k$.

Finally, to simplify the risk:

If $w_k > \tau_k$, then $a_k = 1$ and so $r(a_k, \tau_k, w_k) = \tau_k = \min(\tau_k, w_k)$. If $w_k \leq \tau_k$, then $a_k = 0$ and so $r(a_k, \tau_k, w_k) = w_k = \min(\tau_k, w_k)$. □

10.6 Comparison of the Oracle Estimators and the LSE

For the LSE, $a = \mathbf{1}$, and

$$\hat{m} = \left(\underbrace{\sum_{k=1}^{s} P_k}_{=I_p} \right) \hat{m} = \sum_{k=1}^{s} \mathbf{1} \cdot P_k \hat{m} = \hat{m}(\mathbf{1})$$

where $\mathbf{1} = (1, 1, \ldots, 1)'$.

Its risk is

$$\begin{aligned} R(\hat{\eta}(\mathbf{1}), \eta, \sigma^2) &= \sum_{k=1}^{s} r(1, \tau_k, w_k) \\ &= \sum_{k=1}^{s} \tau_k \\ &= \sum_{k=1}^{s} \underbrace{p^{-1}\sigma^2 \operatorname{tr}(P_k)}_{\tau_k} \\ &= p^{-1}\sigma^2 \operatorname{tr}(\underbrace{\sum P_k}_{I_p}) \\ &= \sigma^2 \end{aligned}$$

Theorem 10.8.

$$\frac{1}{2} R(\tilde{\eta}_{\text{pro}}, \eta, \sigma^2) \overset{1}{\leq} R(\tilde{\eta}_{\text{shr}}, \eta, \sigma^2) \overset{2}{\leq} R(\tilde{\eta}_{\text{pro}}, \eta, \sigma^2) \overset{3}{\leq} \underbrace{R(\hat{\eta}, \eta, \sigma^2)}_{\text{risk of LSE}} = \sigma^2$$

Proof. We already know 3 to be true. The kth summand in $R(\tilde{\eta}_{\text{shr}}, \eta, \sigma^2)$ and in $R(\tilde{\eta}_{\text{pro}}, \eta, \sigma^2)$ are, respectively, $\frac{\tau_k w_k}{\tau_k + w_k}$ and $\min\{\tau_k, w_k\}$. Obviously $\frac{\tau_k w_k}{\tau_k + w_k} \leq \tau_k$ and $\leq w_k$ and so $\leq \min\{\tau_k, w_k\}$.

$$\tau_k \underbrace{\left(\frac{w_k}{\tau_k + w_k}\right)}_{\leq 1}, \qquad w_k \underbrace{\left(\frac{\tau_k}{\tau_k + w_k}\right)}_{\leq 1}$$

On the other hand,

- if $\tau_k \leq w_k$, then $\frac{\tau_k w_k}{\tau_k + w_k} \geq \frac{\tau_k w_k}{2 w_k} = \frac{1}{2}\tau_k = \frac{1}{2}\min\{\tau_k, w_k\}$.

- if $w_k \leq \tau_k$, then $\frac{\tau_k w_k}{\tau_k + w_k} \geq \frac{\tau_k w_k}{2 \tau_k} = \frac{1}{2} w_k = \frac{1}{2}\min\{\tau_k, w_k\}$.

Note: The oracle estimators are unrealizable because the risk function and so $\{w_k\}, \{\tau_k\}$ depend on m and σ^2, which are unkown.

The next step is to devise trustworthy estimators of the risk function. □

11 11-1-11

11.1 Estimators

Model:

$$\underbrace{y}_{n\times 1} = \underbrace{C}_{n\times p}\underbrace{m}_{p\times 1} + \underbrace{e}_{n\times 1}$$

where $\mathbb{E}(e) = 0,\ \mathrm{Cov}(e) = \sigma^2 I_n,\ \{e_i\}$ are i.i.d., $\eta = \mathbb{E}(y) = Cm,\ C'C = n_0 I_p$ with $n_0 \geq 1 \ \Rightarrow$ complete balanced design.

Candidate estimators:

$$\hat{m}(a) = \sum_{k=1}^{s} a_k P_k m \quad \Leftrightarrow \quad \hat{\eta}(a) = C\hat{m}(a)$$

where $\{P_k\}$ are mutually orthogonal projections and $\sum_{k=1}^{s} P_k = I_n$.

Risk:

$$\begin{aligned} R(\hat{\eta}(a), \eta, \sigma^2) &= p^{-1}\mathbb{E}|\hat{\eta}(a) - \eta|^2 \\ &= \sum_{k=1}^{s} r(a_k, \tau_k, w_k) \\ \tau_k &= p^{-1}\sigma^2 \operatorname{tr}(P_k) \\ w_k &= p^{-1} n_0 |P_k m|^2 \\ r(a_k, \tau_k, w_k) &= \tau_k a_k^2 + (1-a_k)^2 w_k \\ &= (a_k - \tilde{a}_k)^2(\tau_k + w_k) + \tau_k \tilde{a}_k \\ \tilde{a}_k &= \frac{w_k}{\tau_k + w_k} \end{aligned}$$

11.2 Adaptive Estimators

Let $\hat{\sigma}^2$ be an asymptotically (as $p \to \infty$) consistent estimator of σ^2. Estimate τ_k by

$$\hat{\tau}_k = p^{-1}\hat{\sigma}^2 \operatorname{tr}(P_k).$$

The naive estimator of w_k is

$$\hat{w}_k = p^{-1} n_0 |P_k \hat{m}|^2$$

where $\hat{m} = n_0^{-1} C' y$.

Know:

$$\begin{aligned} \mathbb{E}(\hat{m}) = m,\ \mathrm{Cov}(\hat{m}) &= n_0^{-1}\sigma^2 I_p \\ \mathbb{E}(\hat{m}\hat{m}') &= mm' + n_0^{-1}\sigma^2 I_p \end{aligned}$$

Hence

$$\begin{aligned} n_0 \mathbb{E}|P_k \hat{m}|^2 &= n_0 \mathbb{E}(\hat{m}' P_k \hat{m}) \\ &= n_0 \mathbb{E} \operatorname{tr}(P_k \hat{m}\hat{m}') \\ &= n_0 \operatorname{tr}[P_k \mathbb{E}(\hat{m}\hat{m}')] \\ &= n_0 |P_k m|^2 + \sigma^2 \operatorname{tr}(P_k) \end{aligned}$$

Thus,

$$\begin{aligned}\mathbb{E}[p^{-1}n_0|P_k\hat{m}|^2] &= p^{-1}n_0|P_k m|^2 + p^{-1}\sigma^2\,\mathrm{tr}(P_k)\\ &= w_k + I_k\end{aligned}$$

This suggests estimating w_k by

$$\check{w}_k = p^{-1}n_0|P_k\hat{m}|^2 - \hat{\tau}_k$$

(cf. Mallows 1973 adjustment)
More convenient is

$$\hat{w}_k = \check{w}_{k+} = \begin{cases} \check{w}_k & \check{w}_k \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Evidently

$$|\hat{w}_k - w_k| \leq |\check{w}_k - w_k|$$

because $\hat{w}_k \geq 0$, as is w_k.

Definition 11.1. ***Estimated Risk***

The *estimated risk* is

$$\begin{aligned}\hat{R}(\hat{\eta}(a)) &= \sum_{k=1}^{s} r(a_k, \hat{\tau}_k, \hat{w}_k)\\ \hat{\tau}_k &= p^{-1}\hat{\sigma}^2\,\mathrm{tr}(P_k)\\ \hat{w}_k &= [p^{-1}n_0|P_k\hat{m}|^2 - \hat{\tau}_k]_+\\ r(a_k, \tau_k, w_k) &= \tau_k a_k^2 + (1-a_k)^2 w_k\\ &= (a_k - \tilde{a}_k)^2(\tau_k + w_k) + \tau_k\tilde{a}_k\\ \tilde{a}_k &= \frac{w_k}{\tau_k + w_k}\end{aligned}$$

By analogy with the oracle estimator:

Definition 11.2. ***Adaptive Shrinkage Estimator***

The *adaptive shrinkage estimator* $\hat{m}_{\mathrm{shr}}$ is the candidate shrinkage estimator that minimizes the estimated risk over all $a \in [0,1]^s$ (or over all $a \in \mathbb{R}^s$).

Theorem 11.3.

$$\hat{m}_{\mathrm{shr}} = \sum_{k=1}^{s} \hat{a}_k P_k\hat{m} = \sum_{k=1}^{s}\left(\frac{\hat{w}_k}{\hat{\tau}_k + \hat{w}_k}\right) P_k\hat{m}$$

where

$$\begin{aligned}\hat{m} &= n_0^{-1}C'y\\ \hat{a}_k &= \frac{\hat{w}_k}{\hat{\tau}_k + \hat{w}_k}\end{aligned}$$

Moreover, for $\hat{\eta}_{\text{shr}} = C\hat{m}_{\text{shr}}$, the estimated risk of the adaptive shrinkage estimator is

$$\hat{R}(\hat{\eta}_{\text{shr}}) = \sum_{k=1}^{s} \hat{\tau}_k \hat{a}_k = \sum_{k=1}^{s} \frac{\hat{\tau}_k \hat{w}_k}{\hat{\tau}_k + \hat{w}_k}$$

Definition 11.4. ***Adaptive Projection Estimator***

The *adaptive projection estimator*, $\hat{m}_{\text{pro}}$, is the candidate shrinkage estimator that minimizes estimated risk over all $a_k \in \{0, 1\}$.

Theorem 11.5.

$$\hat{m}_{\text{pro}} = \sum_{\substack{k \\ \hat{a}_k > 1/2}} P_k \hat{m} = \sum_{\substack{k \\ \hat{w}_k > \hat{\tau}_k}} P_k \hat{m}$$

For $\hat{\eta}_{\text{pro}} = C\hat{m}_{\text{pro}}$, the corresponding estimated risk is

$$\hat{R}(\hat{\eta}_{\text{pro}}) = \sum_{k=1}^{s} \min\{\hat{\tau}_k, \hat{w}_k\}$$

Theorem 11.6.

$$\frac{1}{2}\hat{R}(\hat{\eta}_{\text{pro}}) \leq \hat{R}(\hat{\eta}_{\text{shr}}) \leq \hat{R}(\hat{\eta}_{\text{pro}}) \leq \hat{\sigma}^2 = \hat{R}(\hat{\eta})$$

11.3 Link to Stein Shrinkage (1956, 1961, 1966)

$$\begin{aligned}
\hat{w}_k &= [p^{-1} n_0 |P_k \hat{m}|^2 - \hat{\tau}_k]_+ \\
&= \begin{cases} p^{-1} n_0 |P_k \hat{m}|^2 - \hat{\tau}_k & p^{-1} n_0 |P_k \hat{m}|^2 \geq \hat{\tau}_k \\ 0 & \text{otherwise} \end{cases} \\
\hat{\tau}_k + \hat{w}_k &= \begin{cases} p^{-1} n_0 |P_k \hat{m}|^2 & p^{-1} n_0 |P_k \hat{m}|^2 \geq \hat{\tau}_k \\ \hat{\tau}_k & \text{otherwise} \end{cases} \\
a_k &= \frac{\hat{w}_k}{\hat{\tau}_k + \hat{w}_k} \\
&= \begin{cases} 1 - \frac{\hat{\tau}_k}{p^{-1} n_0 |P_k \hat{m}|^2} & p^{-1} n_0 |P_k \hat{m}|^2 \geq \hat{\tau}_k \\ 0 & \text{otherwise} \end{cases} \\
&= \left[1 - \frac{\hat{\tau}_k}{p^{-1} n_0 |P_k \hat{m}|^2}\right]_+ \\
\hat{m}_{\text{shr}} &= \sum_{k=1}^{s} \left[1 - \frac{P_k \hat{\tau}_k}{n_0 |P_k \hat{m}|^2}\right]_+ P_k \hat{m}
\end{aligned}$$

Notes:

- Apart from small p refinements in the Gaussian error model on e, $\hat{m}_{\text{shr}}$ applies James-Stein (1961) positive-part to each $P_k\hat{m}$.
- Stein (1966) gave an exact treatment under the Gaussian error model with an independent estimate $\hat{\sigma}^2$ of σ^2.
- Our approach supports an asymptotic rationale under the strong Gauss-Markov error model. It also motivates further developments such as penalized least squares.

11.4 Estimating σ^2 in Complete Balanced Designs

Balanced complete design $\Rightarrow C'C = n_0 I_p$

1. When $n_0 > 1$ (replication), the LSE of σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n-p}|y - C\hat{m}|^2$$

   where$n = n_0 p$, $\hat{m} = n_0^{-1}C'y$. This works well if $n \gg p$

2. When $n - p = 0 \Leftrightarrow n_0 = 1$, we might use the estimator of σ^2 associated with a submodel $m = Q\beta$ (where Q is symmetric & idempotent):

$$\hat{\sigma}_0^2 = \frac{1}{n - \text{tr}(Q)}|y - CQ\hat{m}|^2$$

   because

   - $\text{rank}(CQ) = \text{rank}(Q) = \text{tr}(Q)$
   - $Q\hat{m} =$ LSE of $Q\beta$

   in a complete balanced layout. Note: $\hat{\sigma}_0^2$ is usually biased upwards.

11.5 Section: Lab #5 Comments

11.5.1 Part a

P_{GD} is a projection $\Rightarrow$ symmetric & idempotent

$$\begin{aligned}
\hat{\sigma}^2 &= \frac{|P_{\text{GD}}y|^2}{\text{tr}(P_{\text{GD}})} = ? \\
\mathbb{E}(\hat{\sigma}^2) &= \frac{\mathbb{E}|P_{\text{GD}}y|^2}{\text{tr}(P_{\text{GD}})} = \frac{\mathbb{E}(y'P_{\text{GD}}y)}{\text{tr}(P_{\text{GD}})} \\
\mathbb{E}(y'P_{\text{GD}}y) &= \mathbb{E}[\text{tr}(y'P_{\text{GD}}y)] \\
&= \mathbb{E}[\text{tr}(P_{\text{GD}}yy')] \\
&= \text{tr}[P_{\text{GD}}\mathbb{E}(yy')] \\
\text{tr}(AB) &= \text{tr}(BA) \\
y &= m + e, \quad e \sim N(0, \sigma^2 I) \\
\mathbb{E}(y) &= m \\
\mathbb{E}(y'y) &= ?
\end{aligned}$$

11.5.2 Parts f & g

perspective plot $\Rightarrow$ function persp($\cdot$)
Use arguments "phi" and "theta" in persp to adjust the angle of the graphs

11.6 Section: Lab #3 Comments

$$t = \frac{|\hat{\eta} - \hat{\eta}_Q|/(p - r_Q)}{\hat{\sigma}^2}$$
$$T \sim F_{p-r_Q, n-p}$$
$$\text{p-value} = p(T > t)$$

12 11-3-11

12.1 General Problem of Estimating σ^2

Model:

$$\underbrace{y}_{n\times 1} = \underbrace{X}_{n\times p}\underbrace{\beta}_{p\times 1} + \underbrace{e}_{n\times 1}$$

$\operatorname{rank}(X) = r \le p \le n$. The $\{e_i\}$ are i.i.d. $\mathbb{E}(e_i) = 0,\ \operatorname{Var}(e_i) = \sigma^2 < \infty$.

Case 1:

$n > p$, $n - r$ not small. LSE:

$$\begin{aligned} \hat{\sigma}^2 &= \frac{1}{n-r}|y - \hat{\eta}|^2 \\ \hat{\eta} &= XX^+y = Py \qquad \text{with } P = XX^+ \\ \operatorname{tr}(P) &= \operatorname{rank}(P) = r \end{aligned}$$

Case 2:

$n - r$ not small, or even zero. Strategy: fit a submodel to the data and construct the associated σ^2 estimator.

Submodel:

$$\begin{aligned} y &= X_0\beta_0 + e \\ \operatorname{rank}(X_0) &= r_0 < r \\ \mathcal{R}(X_0) &\subset \mathcal{R}(X) \end{aligned}$$

The LSE of $\eta_0 = \mathbb{E}(y)$ is now $\hat{\eta} = X_0X_0^+y = P_0y$, where $P_0 = X_0X_0^+$. $\operatorname{tr}(P_0) = \operatorname{rank}(X_0) = r_0$. The associated estimator of σ^2 in the submodel is

$$\hat{\sigma}_0^2 = \frac{1}{n-r_0}|y - \hat{\eta}_0|^2.$$

Question: Is $\hat{\sigma}_0^2$ a sensible estimator of σ^2 under the general model?

Theorem 12.1.

Under the general model:

$$\mathbb{E}(\hat{\sigma}_0^2) = \sigma^2 + \underbrace{\frac{|\eta - P_0\eta|^2}{n-r_0}}_{\text{bias}}$$

where $\eta = X\beta$.

Proof.

$$\begin{aligned}
\mathbb{E}|y-\hat{\eta}_0|^2 &= \mathbb{E}|y-P_0y|^2 && (P_0 = X_0X_0^+)\\
&= \mathbb{E}|(I_n-P_0)y|^2\\
&= \mathbb{E}\,\mathrm{tr}[y'\underbrace{(I_n-P_0)}_{\substack{\text{symmetric}\\ \text{idempotent}}}y] && (\text{trace trick})\\
&= \mathbb{E}\,\mathrm{tr}[(I_n-P_0)yy']\\
&= \mathrm{tr}[(I_n-P_0)\mathbb{E}(yy')]\\
&= \mathrm{tr}[(I_n-P_0)(\eta\eta'+\sigma^2I_n)]\\
&= \sigma^2\,\mathrm{tr}(I_n-P_0)+\mathrm{tr}[(I_n-P_0)\eta\eta']\\
&= (n-r_0)\sigma^2+|(I_n-P_0)\eta|^2
\end{aligned}$$

Note:

1. Further analysis yields
$$\hat{\sigma}_0^2 \to \mathbb{E}(\hat{\sigma}_0^2) \quad \text{as } n-r_0\to\infty$$

2. In practice we construct $\hat{\sigma}_0^2$ for several of the larger submodels and use the smallest value obtained. We want $n-r_0\geq 30$ to control the variability of $\hat{\sigma}_0^2$.

3. $\hat{\sigma}_0^2$ quantifies the level of variability in y that is deemed unimportant.

4. In the complete balanced design setting,

$$\begin{aligned}
X &= \underbrace{C}_{n\times p}, \qquad r=p\\
X_0 &= CQ \qquad (\text{because } m=Q\beta \text{ describes the submodel})\\
r_0 &= \mathrm{rank}(X_0) = \mathrm{tr}(Q)\\
n &= n_0p \qquad (\text{balanced design})\\
n &= Cm\\
P_0 &= X_0X_0^+\\
&= CQ(Q\underbrace{C'C}_{n_0I}Q)^+Q'C'\\
&= n_0^{-1}CQ\underbrace{Q^+}_{=Q}QC'\\
&= n_0^{-1}CQC'\\
P_0\eta &= n_0^{-1}CQC'\underbrace{CM}_{\eta}\\
&= CQm \qquad \text{as expected.}
\end{aligned}$$

($\mathbb{E}y$ under the submodel where $m=Q\beta \;\Leftrightarrow\; m=Qm$)

$$\mathbb{E}(\hat{\sigma}_0^2) = \sigma^2 + \frac{|Cm-CQm|^2}{n-\mathrm{tr}(Q)}$$

□

12.2 Penalized Least Squares

Model:

$$\underbrace{y}_{n\times 1} = \underbrace{X}_{n\times p}\underbrace{\beta}_{p\times 1} + \underbrace{e}_{n\times 1}, \qquad \text{rank}(X) = r \le p \le n$$

Goal: To estimate $\eta = X\beta$ and, if possible, β.

Definition 12.2. ***Penalized Least Squares***

The *penalized least squares* criterion is

$$T(\beta) = |y - X\beta|^2 + \underbrace{\beta'}_{1\times p}\underbrace{W}_{p\times p}\underbrace{\beta}_{p\times 1}$$

where W is symmetric positive semi-definite.

We consider β values that minimize $T(\beta)$ over $\beta \in \mathbb{R}^p$.

Note:

1. $W = \mathbf{0}$ gives classical least squares.
2. Existence and uniqueness of minimizers has to be resolved.
3. The strategy is to transform the penalized least squares (PLS) problem to the least squares (LS) problem.

Theorem 12.3.

Let $\hat{\beta}_0 = (X'X + W)^+X'y$. Then

1. The minimizers of $T(\beta)$ as β ranges over $\mathbb{R}^p$ are

$$\hat{\beta}(c) = \hat{\beta}_0 + [I_p - (X'X + W)^+(X'X + W)]c, \qquad c \in \mathbb{R}^p$$

   (This formula reduces to the classical LS formula when $W = \mathbf{0}$.)
2. $X\hat{\beta}(c) = X\hat{\beta}_0$ and $W^{1/2}\hat{\beta}(c) = W^{1/2}\hat{\beta}_0$ for all $c \in \mathbb{R}^p$.
3. $\hat{\beta}(c) = \hat{\beta}_0$ for every $c \in \mathbb{R}^p$ iff $\text{rank}(X'X + W) = p$, in which case the Moore-Penrose pseudoinverse is the regular inverse. (i.e. $\text{rank}(X) = p$ or $\text{rank}(W) = p$ or both)

Proof. Key idea: Let

$$\underbrace{\tilde{y}}_{(n+p)\times 1} = \begin{pmatrix} y \\ \mathbf{0}_p \end{pmatrix}, \qquad \underbrace{\tilde{X}}_{(n+p)\times p} = \begin{pmatrix} X \\ W^{1/2} \end{pmatrix}$$

Then

$$\begin{aligned}|\tilde{y} - \tilde{X}\beta|^2 &= \left|\begin{pmatrix} y - X\beta \\ -W^{1/2}\beta \end{pmatrix}\right|^2 \\ &= |y - X\beta|^2 + |-W^{1/2}\beta|^2 \\ &= |y - X\beta|^2 + \beta' W \beta \\ &= T(\beta).\end{aligned}$$

Thus, the PLS criterion is also an LS criterion! So by LS theory:

1. The minimizing values of β are

$$\hat{\beta}(c) = \underbrace{\tilde{X}^+\tilde{y}}_{\hat{\beta}_0} + (I_p - \tilde{X}^+\tilde{X})c \qquad c \in \mathbb{R}^p$$

   Using Lab 1 results, we get

$$\begin{aligned}\tilde{X}^+ &= (\tilde{X}'\tilde{X})^+\tilde{X}' \qquad\qquad = \left[\begin{pmatrix} X' & W^{1/2} \end{pmatrix}\begin{pmatrix} X \\ W^{1/2} \end{pmatrix}\right]^+ \begin{pmatrix} X' & W^{1/2} \end{pmatrix}\begin{pmatrix} y \\ \mathbf{0} \end{pmatrix} \\ &= (X'X + W)^+ X'y\end{aligned}$$

   and

$$\begin{aligned}\tilde{X}^+\tilde{X} &= \underbrace{(\tilde{X}'\tilde{X})^+\tilde{X}'}_{\tilde{X}^+}\tilde{X} \\ &= (X'X + W)^+(X'X + W)\end{aligned}$$

2. LS theory says that $\tilde{X}\hat{\beta}(c)$ is unique for all $c \in \mathbb{R}^p$. In particular,

$$\begin{aligned}\tilde{X}\hat{\beta}(c) = \tilde{X}\hat{\beta}_0 &= \tilde{X}\tilde{X}^+ y \\ &= \begin{pmatrix} X \\ W^{1/2} \end{pmatrix}\hat{\beta}(c) = \begin{pmatrix} X\hat{\beta}(c) \\ W^{1/2}\hat{\beta}(c) \end{pmatrix} \\ &= \begin{pmatrix} X \\ W^{1/2} \end{pmatrix}\hat{\beta}_0 = \begin{pmatrix} X\hat{\beta}_0 \\ W^{1/2}\hat{\beta}_0 \end{pmatrix}\end{aligned}$$

3. The LS form of $T(\beta)$ has a unique minimizer iff $\text{rank}(\tilde{X}) = p$, in which case the minimizer is $(\tilde{X}'\tilde{X})^{-1}\tilde{X}'y = \tilde{X}^+ y = \hat{\beta}_0$. i.e.

$$p = \text{rank}(\underbrace{\tilde{X}}_{(n+p)\times p}) = \text{rank}(\underbrace{\tilde{X}'\tilde{X}}_{p\times p}) = \text{rank}(X'X + W)$$

   Both $X'X$ and W are positive semi-definite ($x'Wx \geq 0 \;\forall\; x \neq \mathbf{0} \;\Rightarrow$ all eigenvalues are nonnegative). So $X'X + W$ is positive semi-definite. $X'X + W$ is positive definite iff at least one of $X'X$ or W is positive definite. Equivalently, either $\text{rank}(X'X) = \text{rank}(X) = p$ or $\text{rank}(W) = p$.

□

12.3 Interpolating Among Submodel Fits Using PLS in Complete Balanced Designs

General Model:

$$\underbrace{y}_{n\times 1} = \underbrace{C}_{n\times p}\underbrace{m}_{p\times 1} + \underbrace{e}_{n\times 1}, \qquad C'C = n_0 I_p,\ n_0 \geq 1$$

Consider again the projection decomposition $\sum_{k=1}^{s} P_k = I_p$, where the $\{P_k\}$ are symmetric and idempotent and mutually orthogonal. Let $l_Q \subset \{1, 2, \ldots, s\}$ and let $Q = \sum_{k\in l_Q} P_k$. The submodel is $m = Q\beta,\ \beta \in \mathbb{R}^p$. To interpolate among LS fits to such submodels, consider the *penalty matrix*:

$$Q(t) = \sum_{k=1}^{s} t_k P_k \qquad \text{where } t_k \geq 0,\ 1 \leq k \leq s.$$

(Q and $Q(t)$ are not the same.) Note that $Q(t)$ is a symmetric positive semi-definite matrix. Consider the PLS criterion:

$$T(m) = |y - Cm|^2 + m'Q(t)m.$$

Note:

$$m'Q(t)m = \sum_{k=1}^{s} t_k m'P_k m = \sum_{k=1}^{s} t_k |P_k m|^2.$$

Since $\text{rank}(C) = p$, by the previous theorem $T(m)$ has a unique minimizer,

$$\hat{m}(t) = \underset{m\in\mathbb{R}^p}{\arg\min}\, T(m) = [C'C + Q(t)]^{-1}C'y$$

(where $t = (t_1, t_2, \ldots, t_s) \in [0, \infty)^s$). This simplifies greatly because $C'C = n_0 I_p$, and so

$$\begin{aligned}
\hat{m}(t) &= \Big[n_0 \underbrace{I_p}_{\sum_{k=1}^{s} P_k} + \sum_{k=1}^{s} t_k P_k\Big]^{-1} C'y = \Big[\sum_{k=1}^{s}(n_0 + t_k)P_k\Big]^{-1} C'y \\
&= \sum_{k=1}^{s}(n_0 + t_k)^{-1} P_k C'y \\
&= \sum_{k=1}^{s} a_k P_k \hat{m}
\end{aligned}$$

where $a_k = \frac{n_0}{n_0 + t_k}$, $\hat{m} = n_0^{-1}C'y$ = LSE of m in $y = Cm + e$. Observe that $a_k \in (0, 1]$.

Note:

1. By adding the values $a_k = 0,\ 1 \leq k \leq s$, we obtain the candidate shrinkage estimators considered earlier as a slight extension of the candidate PLS estimators.

2. The PLS estimator $\hat{m}(t)$ is defined for unbalanced complete layouts:

   $$\hat{m}(t) = \underset{m\in\mathbb{R}^p}{\arg\min}\, T(m) = [C'C + Q(t)]^{-1}C'y$$

   by the previous theorem. $t = (t_1, t_2, \ldots, t_s) \in [0, \infty)^s$. Now $C'C = \text{diag}\{n_i\}$, where n_i is the number of observations on m_i.

3. As in the balanced subcase, we seek to find $\hat{t} \in [0, \infty)^s$ that minimizes the estimated risk of $\hat{m}(t)$. Unfortunately, simplification of $[C'C + Q(t)]^{-1}$ is not obvious.

4. Other penalty matrices can be considered usefully beyond $Q(t)$.

13 11-8-11

13.1 PLS Estimators in Possibly Unbalanced Layouts

Model:

$$\underbrace{y}_{n\times 1} = \underbrace{C}_{n\times p}\underbrace{m}_{p\times 1} + e, \quad C'C = \text{diag}\{n_i\},\ n_i \geq 1\ \forall\ i$$

e satisfies the strong Gauss-Markov model. The PLS estimator of m is

$$\hat{m}(t) = \arg\min_{m\in\mathbb{R}^p}[|y - Cm|^2 + m'\underbrace{Q(t)}_{\substack{\text{penalty}\\ \text{matrix}}} m]$$

$\{P_k\}$ are orthogonal projections (symmetric & idempotent), with $\sum_{k=1}^{s} P_k = I_p$. For example, ANOVA projections.

$$Q(t)m = \sum_{k=1}^{s} t_k \underbrace{|P_k m|^2}_{=m'P_k^2 m} = m'\left(\sum_{k=1}^{s} t_k P_k\right) m$$

$$Q(t) = \sum_{k=1}^{s} t_k P_k = \text{ spectral representation}$$

$$\hat{m}(t) = [C'C + Q(t)]^{-1}C'y, \qquad t \in [0,\infty)^s$$

$$\hat{\eta}(t) = C\hat{m}(t) = C[C'C + Q(t)]^{-1}C'y$$

13.2 Numerical Issues in Computing $\hat{m}(t)$

13.2.1 Aside from numerical analysis

Suppose A $(m \times m)$ is a nonsingular square matrix with SVD $A = ULV'$, where $U'U = V'V = I_m$ and $L = \text{diag}\{l_i\},\ l_1 \geq l_2 \geq \ldots \geq l_m$.

Definition 13.1. ***Matrix Norm***

$$\|A\| = \sup_{x\neq 0}\frac{|Ax|}{|x|}$$

where $|\cdot|$ is the Euclidean norm. (spectral norm, 2-norm)

Definition 13.2. ***Condition Number***

The *condition number* of A is

$$\kappa(A) = \frac{l_1}{l_m} = \|A\|\|A^{-1}\|.$$

Theorem 13.3.

$$\lim_{\epsilon \to 0} \sup_{\|\Delta A\| \leq \epsilon \|A\|} \frac{\|(A + \Delta A)^{-1} - A^{-1}\|}{\epsilon} = \|A^{-1}\| \kappa(A)$$

Notes

1. Results like this are cited in Matrix Computations (Golub & van Loan), 3rd Edition, page 80.
2. Large $\kappa(A)$ entails relatively large errors in $(A + \Delta A)^{-1}$ versus A^{-1}.
3. When A is symmetric,
$$\kappa^2(A) = \frac{l_1^2}{l_m^2} = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}$$

13.2.2 Apply This Aside to PLS Estimation

$$\begin{aligned} \hat{m}(t) &= [C'C + Q(t)]^{-1} C'y \\ &= A^{-1} C'y \end{aligned}$$

for $A = C'C + Q(t)$.

Theorem 13.4.

$$\kappa^2[C'C + Q(t)] \geq \frac{t_{\max}}{n_{\max} + t_{\min}}$$

where $t_{\max} = \max_{1 \leq k \leq s} t_k$, $t_{\min} = \min_{1 \leq k \leq s} t_k$, $n_{\max} = \max_{1 \leq i \leq p} n_i$.

Proof.

$$\begin{aligned} \lambda_{\max}(C'C + Q(t)) &= \max_{|a|=1} a'[C'C + Q(t)]a \\ &= \max_{|a|=1} [\underbrace{a'C'Ca}_{\geq 0} + a'Q(t)a] \\ &\geq \max_{|a|=1} a'Q(t)a \\ &= \lambda_{\max}(Q(t)) = t_{\max} \end{aligned}$$

because $\{t_k\}$ are the eigenvalues of $Q(t)$.

On the other hand,

$$\begin{aligned}
\lambda_{\min} &= \min_{|a|=1} a'[\underbrace{C'C}_{=\text{diag}\{n_i\}} + Q(t)]a \\
&= \min\left[\sum_{i=1}^{p} n_i a_i^2 + a'Q(t)a\right] \\
&\le \min_{|a|=1}[n_{\max}|a|^2 + a'Q(t)a] \\
&= n_{\max} + \min_{|a|=1} a'Q(t)a \\
&\le n_{\max} + \lambda_{\min}
\end{aligned}$$

Thus, for $A = [C'C + Q(t)]$

$$\kappa^2(A) = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)} \ge \frac{t_{\max}}{n_{\max} + t_{\min}}$$

□

Notes

1. Thus, $\kappa[C'C + Q(t)]$ can be very large as t wanders through $[0, \infty)^s$. e.g. $t_{\max}$ large and other $t_k \approx 0$ makes $\kappa[C'C + Q(t)]$ very large.
2. In the balanced design, where $C'C = n_0 I_n$, we have re-expressed the PLS estimator as a shrinkage estimator by computing $[C'C + Q(t)]^{-1}$ algebraically. We seek an analogous strategy for all C such that $C'C = \text{diag}\{n_i\}$.

13.3 Reparameterizing PLS Estimators in the General Unbalanced Case

$$\begin{aligned}
\hat{m}(t) &= [C'C + Q(t)]^{-1}C'y \\
Q(t) &= \sum_{k=1}^{s} t_k P_k, \qquad \sum_{k=1}^{s} P_k = I_p
\end{aligned}$$

$\{P_k\}$ are mutually orthogonal, symmetric & idempotent. $t \in [0, \infty)$. Let

$$d_k^2 = \frac{1}{1 + t_k}, \qquad 1 \le k \le n,\ d_k \in (0, 1].$$

This is a one-to-one reparameterization:

$$\begin{aligned}
1 - d_k^2 &= \frac{t_k}{1 + t_k} \\
t_k &= \frac{1 - d_k^2}{d_k^2}
\end{aligned}$$

Hence,

$$\begin{aligned}
Q(t) = \sum_{k=1}^{s} t_k P_k = \sum_{k=1}^{s} \left(\frac{1 - d_k^2}{d_k^2}\right) P_k &= Q^{-1}(d)Q(\mathbf{1} - d^2)Q^{-1}d, \qquad \text{where} \\
d &= (d_1, d_2, \ldots, d_k)^T \\
\mathbf{1} - d^2 &= (1 - d_1, 1 - d_2, \ldots, 1 - d_k)^T \\
Q^{-1}(d) = \sum_{k=1}^{\infty} d_k^{-1} P_k &= Q(d^{-1})
\end{aligned}$$

d_k^{-1} is positive definite because $\frac{1}{d_k} \geq 1$. Thus,

$$\begin{aligned}
[C'C + Q(t)]^{-1} &= [C'C + Q^{-1}(d)Q(\mathbf{1} - d^2)Q^{-1}(d)]^{-1} \\
&= Q(d)[Q(d)C'CQ(d) + Q(\mathbf{1} - d^2)]^{-1}Q(d) \\
&= Q(d)[\underbrace{Q(d)\overbrace{(C'C - I_p)}^{\text{PSD}}Q(d)}_{\text{PSD}} + I_p]^{-1}Q(d)
\end{aligned}$$

using

$$\begin{aligned}
Q(\mathbf{1} - d^2) &= \sum_{k=1}^{s}(1 - d_k^2)P_n \\
&= \underbrace{\sum_{k=1}^{s} P_k}_{I_p} - \underbrace{\sum_{k=1}^{s} d_k^2 P_k}_{Q^2(d)}
\end{aligned}$$

Definition 13.5. ***Hypercubed***

The *hypercubed* PLS estimators are

$$\begin{aligned}
\hat{m}(d) &= Q(d)[Q(d)(C'C - I_p)Q(d) + I_p]^{-1}Q(d)C'y \\
&= Q(d)[Q(d)C'CQ(d) + Q(\mathbf{1} - d^2)]^{-1}Q(d)C'y
\end{aligned}$$

where $d \in [0, 1]^s$

Theorem 13.6.

The PLS estimators are the subclass where $d \in (0, 1]^s$.

13.4 Numerical Issues for Hypercubed PLS Estimators

Theorem 13.7.

$$\kappa^2[Q(d)C'CQ(d) + Q(\mathbf{1} - d^2)] \leq n_{\max} \qquad \text{for } d \in [0, 1]^k$$

where $n_{\max} = \max_{1 \leq i \leq p} n_i$. i.e., the matrix $Q(d)C'CQ(d) + D(\mathbf{1} - d^2)$ should invert stably.

13.5 Section 11-8-11: Lab 6 Comments

For parts (a)-(d), use matrix results about rank and trace.

For parts (e)-(i), use the results from (a)-(d) to perform data analysis.

For part (h), the fourth difference matrix is

$$D = \begin{pmatrix} 1 & -4 & 6 & -4 & 1 & & \\ & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & & 1 & -4 & 6 & -4 & 1 \end{pmatrix}$$

14 11-10-11

14.1 Penalized Least Squares

Recall:

1. $\underbrace{y}_{n\times 1} = \underbrace{C}_{n\times p}\underbrace{m}_{p\times 1} + e$
2. $C'C = \operatorname{diag}\{n_i\},\ n_i \geq 1$, where C is the data-incidence matrix
3. $n = \sum_{i=1}^{p} n_i$
4. $Q(t) = \sum_{k=1}^{s} t_k P_k$
5. $I_p = \sum_{k=1}^{s} P_k$, where the P_k's are mutually orthogonal projections (e.g. ANOVA)

PLS Estimator

1. $$\hat{m}_{\text{PLS}}(t) = \arg\min_{m\in\mathbb{R}^p} \left[|y - Cm|^2 + m'Q(t)m\right]$$
$$Q(t) = \sum_{k=1}^{s} t_k P_k$$

   so $m'Q(t)m = \sum_{k=1}^{s} t_k |P_k m|^2$

2. $$\hat{m}_{\text{PLS}}(t) = [C'C + Q(t)]^{-1}C'y, \qquad t \in [0,\infty)^s$$

14.2 Hypercubed Penalized Least Squares Estimator (HPLS)

$$\begin{aligned}\hat{m}_{\text{HPLS}} &= Q(d)[Q(d)C'CQ(d) + Q(\mathbf{1} - d^2)]^{-1}Q(d)C'y \\ &= Q(d)[Q(d)(C'C - I_p)Q(d) + I_p]^{-1}Q(d)C'y, \qquad d \in [0,1]^s\end{aligned}$$

HPLS estimators contain PLS estimates, but should promote numerical stability.

14.2.1 Numerical Conditioning of HPLS

Question: Is the matrix inversion well-conditioned?

Theorem 14.1.

$$\kappa^2[Q(d)C'CQ(d) + Q(1 - d^2)] \leq n_{\max}, \qquad n_{\max} = \max_{1\leq i\leq p} n_i, \quad d \in [0,1]^s$$

(where d^2 means coordinate-wise squaring)

Remark 14.2.

Inversion should be stable for all d.

Proof.

$$\kappa^2(A) = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}, \qquad \text{where } A \text{ is symmetric}$$
$$A = Q(d)\underbrace{[C'C - I_p]}_{B}Q(d) + I_p$$

Let $B = CC' - I_p = \text{diag}\{n_i - 1\}$.

$$\begin{aligned}\lambda_{\max} &= \max_{|a|=1} a'Aa \\ &= \max_{|a|=1}[a'Q(d)BQ(d)a + \underbrace{a'a}_{1}] \\ &= \max_{|a|=1} a'Q(d)BQ(d) + 1\end{aligned}$$

Note:

$$|Q(d)a|^2 = a'Q^2(d)a = a'Q(d^2)a \le \lambda_{\max}(Q(d^2))\underbrace{|a|^2}_{1} = d^2_{\max}.$$

Hence

$$\begin{aligned}\lambda_{\max}(A) &\le \max_{|b|\le 1} b'Bb + 1 \\ &= \max_{|b|=1}\sum_{i=1}^{p}(n_i - 1)b_i^2 + 1 \\ &= (n_{\max} - 1) + 1 = n_{\max}\end{aligned}$$

Note: $\lambda_{\max}(A)$ does not depend on d!

$$\begin{aligned}\lambda_{\min}(A) &= \min_{|a|=1}[a' \underbrace{Q(d)BQ(d)}_{\substack{\ge 0 \text{ b/c} \\ B \text{ is pos. semi-def.} \\ \text{b/c } n_i - 1}} a + a'a] \\ &\ge 0 + 1 = 1\end{aligned}$$

□

14.3 HPLS Estimators Include Submodel Fits

Model: $y = Cm + e$.

Let $l_Q \subset \{1, 2, \ldots, s\}$ and $\underbrace{Q}_{p\times p} = \sum_{k\in l_Q} P_k$. Q is symmetric and idempotent and defines the submodel $m = Q\beta,\ \beta \in \mathbb{R}^p$.

Recall: In submodel Q, the LSE of m is $\hat{m}_0 = (CQ)^+ y$.

Theorem 14.3.

Suppose

$$d_k = \begin{cases} 1 & k \in l_Q \\ 0 & \text{otherwise} \end{cases}$$

Then

$$\hat{m}_{\text{HPLS}}(d) = (CQ)^+ y.$$

Proof.

$$\begin{aligned} \hat{m}(d) &= Q(d)[Q(d)C'CQ(d) + Q(1-d)^2]^{-1}Q(d)C'y \\ &\quad \text{by the choice of } d \\ Q(d) &= Q \\ Q(1-d^2) &= \sum_{k=1}^{s}(1-d_k)^2 P_k = I_p - Q \\ Q(I-Q) &= 0 \\ \hat{m}(d) &= Q[\underbrace{QC'CQ}_{A} + \underbrace{(I_p - Q)}_{B}]^{-1}QC'y \end{aligned}$$

A and B are symmetric with $AB = 0$ (think of the midterm problem). Hence

$$\begin{aligned} (A+B)^{-1} &= (A+B)^+ = A^+ + B^+ \\ \hat{m}(d) &= Q\,\underbrace{[(CQ)'(CQ)]^+}_{A^+}\,QC'y + Q\,\underbrace{(I_p - Q)^+}_{B^+}\,QC'y \\ &= C(CQ)^+ y + 0 \qquad \text{because } (I_p - Q)^+ = I_p - Q \\ &= (CQ)^+ y \qquad \text{by Lab 1 part l} \end{aligned}$$

□

14.4 Symmetric Linear Estimators

Linear model of full rank:

$$y = \underbrace{X}_{n\times p}\,\underbrace{\beta}_{p\times 1} + e$$

1. Assume $\text{rank}(X) = p \le n \;\Leftrightarrow$ full rank
2. Assume Gauss-Markov error model: the e_i are i.i.d. and $\mathbb{E}(e_i) = 0,\ \text{Var}(e_i) = \sigma^2 < \infty,\ \mathbb{E}[e_i^4] < \infty$.

Definition 14.4. ***Linear Estimator***

A *linear estimator* of $\eta = \mathbb{E}(y) = X\beta$ has the form Ay, where A does not depend on y, and A is an $n \times n$ matrix.

Loss of Ay is $p^{-1}|Ay - \eta|^2$.

Risk of Ay is $R(Ay, \eta, \sigma^2) = \mathbb{E}[p^{-1}|Ay - \eta|^2] = p^{-1}\mathbb{E}[|Ay - \eta|^2]$.

Theorem 14.5.

1. The risk of Ay is

$$\begin{aligned} R(Ay, \eta, \sigma^2) &= p^{-1}[\sigma^2 \operatorname{tr}(A'A) + \eta'(I_n - A)'(I_n - A)\eta] \\ &= p^{-1}[\sigma^2 |A|^2 + |\eta - A\eta|^2] \\ &= p^{-1}\Big[\sigma^2 \underbrace{\operatorname{tr}(A'A)}_{\text{Frobenius norm}} + \operatorname{tr}[(I_n - A)'(I_n - A)\eta\eta']\Big] \end{aligned}$$

2. The oracle $\tilde{A}$ that minimizes this risk is

$$\begin{aligned} \tilde{A} &= I_n - (I_n + \sigma^{-2}\eta\eta')^{-1} \\ &= (\sigma^2 + |\eta|^2)^{-1}\eta\eta' \end{aligned}$$

   $\tilde{A}$ is a symmetric positive semi-definite $n \times n$ matrix.

3. The canonical form of $\tilde{A}$. Let $H = X'X,\ U = XH^{-1/2}$. Then

$$\begin{aligned} \eta &= U\xi, & \xi &= H^{1/2}\beta \\ U'U &= I_p \\ \tilde{A} &= U\tilde{S}U, & \tilde{S} &= (\sigma^2 + |\xi|^2)^{-1}\xi\xi' \end{aligned}$$

   Here $\tilde{S}$ is symmetric with eigenvalues in $[0, 1]$.

Proof. (Sketch)

1.

$$\begin{aligned} R(Ay, \eta, \sigma^2) &= p^{-1}\mathbb{E}[|Ay - \eta|^2] = p^{-1}\mathbb{E}[|A(y - \eta) + (I_n - A)\eta|^2] \\ &= p^{-1}[\sigma^2 \operatorname{tr}(A'A) + \underbrace{\eta'(I_n - A)'(I_n - A)\eta}_{\eta'\eta - 2\eta'A\eta + \eta'A'A\eta}] \\ &= \text{variance} + \text{bias} \\ &= \text{convex functions of } A \end{aligned}$$

2. We use matrix derivatives to minimize the risk (c.f. Rao & Toutenberg (1995) or "Matrix Codebook" online)

$$\begin{aligned} \frac{\partial \operatorname{tr}(A'A)}{\partial A} &= 2A \\ \frac{\partial \eta'A\eta}{2A} &= \eta\eta' \\ \frac{\partial \eta'A'A\eta}{\partial A} &= 2A\eta\eta' \end{aligned}$$

   Hence,

$$\frac{\partial R(Ay, \eta, \sigma^2)}{\partial A} = p^{-1}[2\sigma^2 A - 2\eta\eta' + 2A\eta\eta']$$

Set $\frac{\partial R(Ay,\eta,\sigma^2)}{\partial A} = 0$ and solve for $\tilde{A}$.

$$\begin{aligned}
\tilde{A}[\sigma^2 I_n + \eta\eta'] &= \eta\eta' \\
\tilde{A} &= \eta\eta'(\sigma^2 I_n + \eta\eta')^{-1} \\
&= \frac{\eta\eta'}{\sigma^2}\left(I_n + \frac{\eta\eta'}{\sigma^2}\right)^{-1} = W(I_n + W)^{-1}, \qquad \text{where } W = \frac{\eta\eta'}{\sigma^2}.
\end{aligned}$$

Thus, $\tilde{A} = I_n - (I_n + W)^{-1}$ = symmetric matrix because

$$\begin{aligned}
I_n &= (I_n + W)(I_n + W)^{-1} \\
&= (I_n + W)^{-1} + W(I_n + W)^{-1}
\end{aligned}$$

This is the first formula for $\tilde{A}$. To get the second formula, we use the identity (see Schott page 10)

$$(I_n + \underbrace{c}_{n\times 1}\underbrace{d'}_{1\times n})^{-1} = I_n - \frac{cd'}{1 + d'c}.$$

Set $c = d = \sigma^{-1}\eta$ to get

$$\begin{aligned}
(I_n + \sigma^2\eta\eta') &= I_n - \frac{\sigma^{-2}\eta\eta'}{1 + \sigma^{-2}|\eta|^2} \\
&= I_n - (\sigma^2 + |\eta|^2)^{-1}\eta\eta'
\end{aligned}$$

Thus, $\tilde{A} = (\sigma^2 + |\eta|^2)^{-1}\eta\eta'$, a symmetric positive semi-definite matrix.

3. Canonical form of $\tilde{A}$. Since $\eta = U\xi$, with $U'U = I_p$,

$$\begin{aligned}
|\eta|^2 &= \xi' U' U \xi = |\xi|^2 \\
\eta\eta' &= U\xi\xi' U' \\
\tilde{A} &= (\sigma^2 + |\xi|^2)^{-1} U\xi\xi' U' = U\tilde{S}U', \qquad \tilde{S} = (\sigma^2 + |\xi|^2)^{-1}\xi\xi' \\
\lambda_{\min}(\tilde{S}) &\geq 0 \qquad \text{because } \tilde{S} \text{ is p.s.d.} \\
\lambda_{\max}(\tilde{S}) &= \max_{|a|=1} a'\tilde{S}a = \max_{|a|=1} \frac{a'\xi\xi' a}{\sigma^2 + |\xi|^2} \\
&= \frac{\max_{|a|=1} |a'\xi|^2}{\sigma^2 + |\xi|^2} \overset{\text{C.S.}}{\leq} \frac{|\xi|^2}{\sigma^2 + |\xi|^2} < 1
\end{aligned}$$

Note

(a) This motivates interest in studying symmetric linear estimates, Ay, where $A = USU'$ with U constructed from X as discussed and S symmetric with eigenvalues in $[0, 1]$.

(b) Symmetric linear estimators with this structure arise naturally.

□

Example 14.6.

In the model $y = X\beta + e,\ \text{rank}(X) = p \leq n$, the LSE of $\eta = X\beta$ is

$$\begin{aligned}\hat{\eta} &= U \underbrace{H^{1/2}H^{-1}H^{1/2}}_{I} U'y \\ &= UU'y \\ &= U \underbrace{S}_{=I_p} U'y\end{aligned}$$

We have shrinkage when we replace eigenvalues less than 1.

Example 14.7.

PLS estimators in $y = Cm + e$, where C is the data-incidence matrix, $\underbrace{C'C}_{p\times p} = \text{diag}\{n_i\},\ n_i \geq 1$.

Penalty matrix from earlier:

$$\begin{aligned}Q(t) &= \sum_{k=1}^{s} t_k P_k, \quad \text{etc.} \\ \hat{\eta}_{\text{PLS}} &= \underbrace{C[C'C + Q(t)]^{-1}C'}_{\text{symmetric}} y\end{aligned}$$

So $\hat{\eta}_{\text{PLS}}$ is a symmetric linear estimator.

14.4.1 Canonical Structure

Let $H = C'C,\ U = CH^{-1/2} \Leftrightarrow C = UH^{1/2}$. Then

$$\begin{aligned}C[C'C + Q(t)]^{-1}C' &= UH^{1/2}[H + Q(t)]^{-1}H^{1/2}U \\ &= U[\underbrace{I_p + H^{-1/2}Q(t)H^{-1/2}}_{S(t)}]^{-1}U'\end{aligned}$$

$S(t)$ is symmetric, and

$$\begin{aligned}\min_{|a|=1} a'[I_p + H^{-1/2}Q(t)H^{-1/2}]a &= \text{smallest eigenvalue of } S^{-1}(t) \\ &\geq 1\end{aligned}$$

Thus, the larges eigenvalues of $S(t)$ are ≤ 1. The smallest eigenvalues of $S(t)$ are ≥ 0 because $S(t)$ is positive semi-definite.

14.5 Discussion

If canonical structure, we reduce numerical error.

- HPLS are numerically stable.
- HPLS has canonical form.

15 11-15-11

15.1 Last Time

Last time: $y = X\beta + e$. Found oracle linear estimator $\tilde{A}y$ that minimizes risk. Motivates symmetric linear estimators of the form Ay, where

$$\begin{aligned} A &= \underbrace{U}_{n\times p}\underbrace{S}_{p\times p}\underbrace{U'}_{p\times n} \\ U &= X(X'X)^{-1/2} \\ \mathcal{R}(U) &= \mathcal{R}(X) \\ U'U &= I_p \end{aligned}$$

where S is symmetric with eigenvalues in $[0,1]$.

15.2 Actual Examples

1. LSE of η is
$$\hat{\eta}_{\text{LS}} = X(X'X)^{-1}X'y = UU'y = U\underbrace{I_p}_{S}U'y$$

2. PLS in $\underbrace{y}_{n\times 1} = \underbrace{C}_{n\times p}\underbrace{m}_{p\times 1} + \underbrace{e}_{n\times 1}$ with penalty matrix $Q(t) = \sum_{k=1}^{s} t_k P_k$, with $\sum_{k=1}^{s} P_k = I_p$, $t_k \geq 0$, $P_k^2 = P_k$, $P_k P_j = 0$ if $j \neq k$.
$$\hat{\eta}_{\text{PLS}}(t) = C[C'C + Q(t)]^{-1}C'y$$
Let $H = C'C,\ U = CH^{-1/2}$.
$$\begin{aligned} \hat{\eta}_{\text{PLS}}(t) &= UH^{1/2}[H + Q(t)]^{-1}H^{1/2}U'y \\ &= U[I_p + H^{-1/2}Q(t)H^{-1/2}]^{-1}U'y \\ &= US(t)U'y \end{aligned}$$
with
$$\underbrace{S(t)}_{\text{symm}} = [I_p + H^{-1/2}Q(t)H^{-1/2}]^{-1}.$$
Note:
$$a'S^{-1}(t)a = a'[I_p + \underbrace{H^{-1/2}Q(t)H^{-1/2}}_{\text{psd}}]a \geq 1 \qquad \text{if } |a| = 1,$$
so the eigenvalues of $S^{-1}(t)$ are ≥ 1 (because they are the reciprocals of the eigenvalues of $S(t)$, which lie in $[0,1]$).

Bujua, Hastie, Tibshirani (1989) discussed through examples symmetric linear estimators.

3. Hypercubed PLS can also be put into the form $US(d)U'y$, where $d \in [0,1]^s$.
PLS $\Leftrightarrow$ HPLS when $d \in (0,1]^s$.
The eigenvalues of $S(d)$ are in $[0,1]$ for $d \in [0,1]^k$.

15.3 Risk and Estimated Risk of a Symmetric Linear Estimator

Theorem 15.1.

Model: $y = X\beta + e$, $\text{rank}(X) = p$, $\eta = X\beta$. Let A be an $n \times n$ symmetric matrix.

1. The risk of Ay as an estimator of η is

$$\begin{aligned} R(Ay, \eta, \sigma^2) &= p^{-1}\mathbb{E}|Ay - \eta|^2 \\ &= p^{-1}\left[\sigma^2 \operatorname{tr}(A^2) + \operatorname{tr}\left((I_n - A)^2\eta\eta'\right)\right] \\ &= p^{-1}[\sigma^2|A|^2 + |\eta - A\eta|^2] \end{aligned}$$

2. Suppose $\hat{\sigma}^2$ is the LSE of σ^2 (assume $n > p$). Then the estimated risk is

$$\hat{R}(A) = p^{-1}\left[\hat{\sigma}^2 \operatorname{tr}(A^2) + \operatorname{tr}\left((I_n - A)(yy' - \hat{\sigma}^2 I_n)\right)\right] \tag{15.1}$$

$$= p^{-1}\left[|y - Ay|^2 + \left(2\operatorname{tr}(A) - n\right)\hat{\sigma}^2\right] \tag{15.2}$$

is unbiased for $R(Ay, \eta, \sigma^2)$.

Proof. 1. Follows by specializing the risk of a linear estimator, using $A' = A$.

2. Consider

$$\begin{aligned} \mathbb{E}|y - Ay|^2 &= \mathbb{E}|\underbrace{(I_n - A)}_{\text{symm}} y|^2 \\ &= \mathbb{E}[y'(I_n - A)^2 y] \\ &= \mathbb{E}\operatorname{tr}[(I_n - A)^2 yy'] \\ &= \operatorname{tr}[(I_n - A)^2 \underbrace{(\eta\eta' + \sigma^2 I_n)}_{\mathbb{E}(yy')}] \\ &= \operatorname{tr}[(I_n - A)^2\eta\eta'] + \sigma^2 \operatorname{tr}[(I_n - A)^2] \\ &= \operatorname{tr}[(I_n - A)^2\eta\eta'] + \sigma^2[n - 2\operatorname{tr}(A) + \operatorname{tr}(A^2)] \\ p^{-1}\mathbb{E}|y - Ay|^2 &= R(Ay, \eta, \sigma^2) + (n - 2\operatorname{tr}(A))\sigma^2 \end{aligned}$$

i.e. (15.2) is an unbiased estimator for risk. For (15.1), note that $\mathbb{E}(\hat{\sigma}^2) = \sigma^2$ and

$$\mathbb{E}[yy' - \hat{\sigma}^2 I_n] = (\eta\eta' + \sigma^2 I_n) - \sigma^2 I_n = \eta\eta'.$$

□

Note:

1. This result generalizes the Mallows (1973) argument for the estimated risk of submodel fits.
2. The C_p-criterion for $\hat{\eta} = Ay$ is $|y - Ay|^2 + 2\operatorname{tr}(A)\hat{\sigma}^2$

15.4 Specialization for Canonical Symmetric Linear Estimators $USU'y$

Model: $y = X\beta + e$, $\text{rank}(X) = p \leq n$. Let U $(n \times p)$ be such that $\mathcal{R}(U) = \mathcal{R}(X)$ and $U'U = I_p$.

$$\hat{\eta} = USU'y = Ay$$

where $A = USU'$.

Definition 15.2. *Canonical*

A symmetric linear estimator Ay is *canonical* iff $A = USU'$ (as above).

Theorem 15.3.

Suppose Ay is a canonical symmetric estimator.

1. The risk of Ay is
$$\begin{aligned} R(Ay, \eta, \sigma^2) &= p^{-1}\left[\sigma^2 \operatorname{tr}(S^2) + \operatorname{tr}\left((I_p - S)^2 \xi\xi'\right)\right] \\ &= p^{-1}[\sigma^2 |S|^2 + |\xi - S\xi|^2] \end{aligned}$$
where $\xi = U'\eta$.
2. Suppose $\hat{\sigma}^2 = \frac{1}{n-p}|y - UU'y|^2$ is the LSE of σ^2 $(n > p)$. Then
$$\begin{aligned} \hat{R}(A) &= p^{-1}\left[\hat{\sigma}^2 \operatorname{tr}(S^2) + \operatorname{tr}\left((I_p - S)^2 (zz' - \hat{\sigma}^2 I_p)\right)\right] \\ &= p^{-1}\left[|z - Sz|^2 + \left(2\operatorname{tr}(S) - p\right)\hat{\sigma}^2\right] \end{aligned}$$
where $z = \underbrace{U'}_{p\times n}\underbrace{y}_{n\times 1}$ is unbiased for $R(Ay, \eta\sigma^2)$.

15.5 Section 11-15-11

Construct V_d. The first d columns of V_d are the normalized orthogonal polynomials supported on 1 to p of degrees 0 to $d-1$.
$v_2 = \text{poly}(x, \text{degree}) \to d-1$ vectors from degree 1 to $d-1$

$x = 1:p$
degree $= d-1$
$v_1 = \text{rep}(1,p)/\sqrt{p}$
cbind(v_1, v_2)
For part (c), under what condition on m in the model will $\lim_{p\to\infty} \underbrace{\mathbb{E}|\hat{\sigma}^2_{\text{H}} - \sigma^2|}_{\text{mean square error}} = 0$? Don't need rigorous argument, just perception.

For part (h), explain in terms of your findings. Look for a seasonal pattern and at sales on holidays.

16 11-17-11

16.1 Comments on Lab 7

For part (c), assume that the errors are Gaussian. This simplifies things because then the z's are Gaussian. The estimator is clearly biased, but what is the nature of the bias and when is it manageable? When is the estimator useful?

16.2 Symmetric Linear Estimators (Continued)

$$y = \underbrace{X}_{n\times p}\underbrace{\beta}_{p\times 1} + e, \qquad \operatorname{rank}(X) = p \le n, \quad e \text{ strong Gauss-Markov}$$

$\hat{\eta} = Ay$, where A is symmetric, is a candidate linear symmetric estimator.

We saw that the risk of Ay is

$$\begin{aligned} R(Ay, \eta\sigma^2) &= p^{-1}[\sigma^2 \operatorname{tr}(A^2) + \operatorname{tr}[(I_n - A)^2\eta\eta']] \\ &= p^{-1}[\sigma^2|A|^2 + |\eta - A\eta|^2] \end{aligned}$$

The estimated risk is

$$\begin{aligned} \hat{R}(A) &= \left[\hat{\sigma}^2 \operatorname{tr}(A^2) + \operatorname{tr}\left((I_n - A)(yy' - \hat{\sigma}^2 I_n)\right)\right] \\ &= p^{-1}\left[|y - Ay|^2 + \left(2\operatorname{tr}(A) - n\right)\hat{\sigma}^2\right] \qquad \text{(Mallows)} \end{aligned}$$

where $\hat{\sigma}^2$ is the LSE of σ^2 (i.e. $n > p$).

16.3 Canonical Symmetric Linear Estimators

Let $\underbrace{U}_{n\times p}$ be such that $\mathcal{R}(U) = \mathcal{R}(X)$ and $U'U = I_p$ (e.g. $U = X(X'X)^{-1/2}$ OR $X = \underbrace{U}LV'$ SVD). Since $\mathcal{R}(U) = \mathcal{R}(X)$, $\eta \in \mathcal{R}(X)$ has the form $\eta = U\xi$ for $\xi \in \mathbb{R}^p \;\Rightarrow\; \xi = U'\eta$.

Definition 16.1. ***Canonical***

The symmetric linear estimator Ay is *canonical* iff $A = USU'$ for some symmetric $p \times p$ matrix S whose eigenvalues lie in $[0, 1]$.

Theorem 16.2. ***Simplified Risk & Estimated Risk Formulas for Canonical Symmetric Linear Estimators***

Suppose that AY is a canonical symmetric linear estimator and $\hat{\sigma}^2$ is the LSE of σ^2 $(n > p)$. Then

1.
$$\begin{aligned} R(Ay, \eta, \sigma^2) &= p^{-1}\left[\sigma^2 \operatorname{tr}(S^2) + \operatorname{tr}\left((I_p - S)^2 \xi\xi'\right)\right] \\ &= p^{-1}[\sigma^2 |S|^2 + |\xi - S\xi|^2] \end{aligned}$$

2. Also, for $z = U'y$
$$\begin{aligned} \hat{R}(A) &= p^{-1}\left[\hat{\sigma}^2 \operatorname{tr}(S^2) + \operatorname{tr}\left((I_p - S)^2 (zz' - \hat{\sigma}^2 I_p)\right)\right] \\ &= p^{-1}\left[|z - Sz|^2 + \big(2\operatorname{tr}(S) - p\big)\hat{\sigma}^2\right] \end{aligned}$$

is an unbiased estimator of $R(Ay, \eta, \sigma^2)$.

Proof. Second risk formula (which implies the first form):

$$\begin{aligned} |A|^2 = \operatorname{tr}(A^2) &= \operatorname{tr}[US\underbrace{U' \cdot U}_{I_p}SU'] \\ &= \operatorname{tr}[S^2\underbrace{U'U}_{I_p}] = \operatorname{tr}(S^2) = |S|^2 \\ |\eta - A\eta| = (\eta - A\eta)'(\eta - A\eta) &\qquad \text{use } A\eta = US\underbrace{U' \cdot U}_{I_p}\xi = US\xi \\ &= (U\xi - US\xi)'(U\xi - US\xi) = (\xi - S\xi)'\underbrace{U'U}_{I_p}(\xi - S\xi) \\ &= |\xi - S\xi|^2 \end{aligned}$$

The second estimated risk formula:

$$\begin{aligned} |y - Ay|^2 &= |\underbrace{UU'y + (I_n - UU')y}_{y} - \underbrace{USU'}_{A}y|^2 \\ &= |UU'y - USU'y + (I_n - UU')y|^2 = |UU'y - US\underbrace{U'y}_{z}|^2 + |(I_n - UU')y|^2 \qquad \text{(cross-product = 0)} \\ &= |U(z - Sz)|^2 + |y - \underbrace{UU'y}_{\text{LSE of } \eta}|^2 \\ &= |z - Sz|^2 + (n - p)\hat{\sigma}^2 \qquad \text{where } \hat{\sigma}^2 = \text{LSE of } \sigma^2 \end{aligned}$$

Also,

$$\begin{aligned} \big(2\operatorname{tr}(\underbrace{A}_{USU'}) - n\big)\sigma^2 &= \big(2\operatorname{tr}(S) - n\big)\hat{\sigma}^2 \\ \hat{R}(A) &= |z - Sz|^2 + (\cancel{n} - p)\hat{\sigma}^2 + \big(2\operatorname{tr}(S) - \cancel{n}\big)\hat{\sigma}^2 \\ &= |z - Sz|^2 + \big(\operatorname{tr}(S) - p\big)\hat{\sigma}^2 \end{aligned}$$

□

Notes

1. The S-form uses smaller $p \times p$ matrices than the $n \times n$ A-form.
2. The S-form does not identically match the A-form when $\hat{\sigma}^2$ is not the LSE of σ^2.
   - However, both forms of the estimated risk still converge correctly to risk if $\hat{\sigma}^2$ is suitably consistent.
3. To avoid negative estimated risks, we might also consider the modified formula

$$\hat{R}_+(A) = p^{-1}\left[\hat{\sigma}^2 \operatorname{tr}(S^2) + \left(\operatorname{tr}[(I_p - S)(zz' - \hat{\sigma}^2 I_p)]\right)_+\right]$$

   Remark: feel free to use the uncorrected estimated risk formulas in the final project. Smaller is better, even if it means negative.

16.4 Applications to PLS

16.4.1 Interpolating Among Submodel Fits (Complete Design)

Model:

$$y = \underbrace{C}_{n\times p}\underbrace{m}_{p\times 1} + \underbrace{e}_{n\times 1}$$

where C is the data-incidence matrix, $C'C = \operatorname{diag}\{n_i\},\ n_i \geq 1$. $\{P_k \mid 1 \leq k \leq s\}$ are mutually orthogonal projections with $\sum_{k=1}^{s} P_k = I_p$. (e.g. ANOVA decomposition) $\{P_k\}$ are symmetric & idempotent: $P_k P_j = 0$ if $j \neq k$. Let

$$Q(t) = \sum_{k=1}^{s} t_k P_k, \qquad t_k \geq 0.$$

The candidate PLS estimator of m is

$$\hat{m}_{\text{PLS}}(t) = [C'C + Q(t)]^{-1} C'y.$$

The hypercubed PLS estimator of m is

$$\hat{m}_{\text{HPLS}}(d) = Q(d)\left[Q(d)(C'C - I_p)Q(d) + I_p\right]^{-1} Q(d)C'y, \qquad d \in [0,1]^s$$

Put $H = C'C$. Let $U = CH^{-1/2} \Leftrightarrow C = UH^{1/2}$. Then

$$\begin{aligned} \hat{\eta}(d) &= C\hat{m}(d) = US(d)U'y \\ \text{with} \quad S(d) &= H^{1/2}Q(d)\left[Q(d)(H - I_p)Q(d) + I_p\right]^{-1} Q(d)H^{1/2} \end{aligned}$$

Note:

1. When $d \in (0,1]^s$, the eigenvalues of $S(d)$ lie in $[0,1]$. Then

$$\hat{\eta}(d) = U \underbrace{H^{1/2}\left[H + Q(t)\right]^{-1} H^{1/2}}_{S(d) \text{ in } t\text{-parameterization}} Uy$$

2. This property is preserved if we let $d_k \to 0$.

Hence, $US(d)U'y$ is a canonical symmetric linear estimator for every $d \in [0,1]^s$.

The estimated risk of HPLS estimator $\hat{\eta}(d)$ (or $\hat{m}(d)$) is now

$$\begin{aligned} \hat{R}(d) &= p^{-1}\left[\hat{\sigma}^2 \operatorname{tr}\left(S^2(d)\right) + \operatorname{tr}\left((I_p - S(d))^2(zz' - \hat{\sigma}^2 I_p)\right)\right] \\ &= p^{-1}\left[|z - S(d)z|^2 + 2\operatorname{tr}\left(S(d) - p\right)\hat{\sigma}^2\right] \end{aligned}$$

where $z = U'y,\ U = C\underbrace{(C'C)^{-1/2}}_{H^{-1/2}},\ \hat{\sigma}^2 =$ LSE of σ^2 if $n > p$ or another consistent estimator of σ^2. We now minimize estimated risk by choice of $d \in [0,1]^s$.

Note

1. If we restrict each d_k to be either 0 or 1, this method identifies the submodel fit with smallest estimated risk.
2. This is computationally efficient because all of these submodel fits are treated simultaneously.
3. Moreover, minimizing estimated risk over all $d \in [0,1]^s$ may be reduce risk substantially.

16.4.2 Ordinary PLS with One Covariate (Complete Design)

Model:

$$y = \underbrace{C}_{n\times p}\underbrace{m}_{p\times 1} + e, \qquad C'C = \text{diag}\{n_i\},\ n_i \geq 1$$
$$m_i = \mu(x_i)$$

WLOG, assume $x_1 < x_2 < \cdots < x_p$ are distinct values of a covariate. μ is an unknown function. Either

1. The covariate is ordinal: the values and order of the covariate matter. Example: Canadian earnings data.
2. The covariate is nominal: it is just a label; permutation of labels is harmless. Example: rat litter data.

PLS candidate estimators – general strategy

Let $\underbrace{A}_{?\times p}$ be an annihilator matrix: $Au = 0$ for $\underbrace{u}_{p\times 1} = p^{-1/2}(1,1,\ldots,1)'$. Let

$$\hat{m}(\nu) = \underset{m\in\mathbb{R}^p}{\arg\min}\left[|y - Cm|^2 + \nu|Am|^2\right] = \underset{m\in\mathbb{R}^p}{\arg\min}\left[|y - Cm|^2 + \nu m'\underbrace{B}_{A'A}m\right]$$

From earlier, we know that

$$\hat{m}(\nu) = [C'C + \nu B]^{-1}C'y$$

(In Lab 6, $A =$ 4th difference matrix.) Can we hypercube the parameterization to stabilize computation of $\hat{m}(\nu)$ for large ν?

Since $\underbrace{B}_{p\times p} = A'A$ is positive semi-definite, it has spectral representation

$$B = \sum_{k=1}^{s}\lambda_k P_k$$

where P_k is the eigenprojection for eigenvalue λ_k. Order the eigenvalues as

$$0 \leq \lambda_1 < \lambda_2 < \cdots < \lambda_s, \qquad s \leq p(\text{b/c of possible multiplicities})$$

$P_kP_j = 0$ if $j \neq k$. $P_k^2 = P_k$. $\sum_{k=1}^{s} P_k = I_p$. Thus,

$$\nu B = \sum_{k=1}^{s}(\nu\lambda_k)P_k = \sum_{k=1}^{s} t_k Q_k$$

where $t = (\nu\lambda_1, \nu\lambda_2, \ldots, \nu\lambda_s)$. i.e. $t \in something = \{(\nu\lambda_1, \nu\lambda_2, \ldots, \nu\lambda_s) \mid \nu \geq 0\} \subset [0, \infty)^s$. The candidate PLS estimators amount to the estimators

$$\hat{m}(t) = [C'C + Q(t)]^{-1}C'y \qquad \text{for } t \in something.$$

The hypercubed PLS estimators are:

$$\begin{aligned}\hat{m}(d) &= Q(d)\left[Q(d)C'CQ(d) + Q(1-d^2)\right]Q(d)C'y \\ &= Q(d)\left[Q(d)(C'C - I_p)Q(d) + I_p\right]^{-1}Q(d)C'y\end{aligned}$$

where $d_i^2 = \frac{1}{1+t_i}$, $d \in \mathcal{D}$ is a restricted subset of $[0,1]^s$. Let

$$\mathcal{D}_0 = \{(1+\nu\lambda_1)^{-1/2}, (1+\nu\lambda_2)^{-1/2}, \ldots, (1+\nu\lambda_s)^{-1/2}\}$$

Cases:

1. Suppose $\lambda_i = 0$. Then $\mathcal{D} = \mathcal{D}_0 \cup (1, 0, 0, \ldots, 0)$. $\leftarrow$ Adds the fit to the submodel $m = P_1\beta,\ \beta \in \mathbb{R}^p$, to the candidate class.
2. Suppose $\lambda_1 > 0$. Then $\mathcal{D} = \mathcal{D}_0 \cup (0, 0, \ldots, 0)$. (Think of Lab 4.)

Numerical Stability when $t_k = \nu\lambda_k$

From earlier, the condition number for PLS satisfies

$$\kappa^2[C'C + Q(t)] \geq \frac{\nu\lambda_{\max}}{n_{\max} + \nu\lambda_{\min}}$$

which can be large; for example, if $\lambda_{\min} = 0$ and ν is large. The condition number for HPLS satisfies

$$\kappa^2[Q(d)C'CQ(d) + Q(1-d^2)] \leq n_{\max}$$

which does not depend on ν.

17 11-22-11

17.1 Simple PLS for One Covariate $\Leftrightarrow$ One-Way Layout

Candidate PLS:

$$\begin{aligned}\hat{m}(\nu) &= \underset{m\in\mathbb{R}^p}{\arg\min}[|y-Cm|^2+\nu|Am|^2], \quad \nu\geq 0\\ &= [C'C+\nu m'Bm]\end{aligned}$$

where $B=A'A$, A is an annihilialtor: $Au=\mathbf{0},\ u=p^{-1/2}(1,1,\ldots,1)'$. Candidate HPLS extension...

17.1.1 Constructions of A

The rows of A are *contrasts* (meaning they sum up to zero). $m=\mu(x_i)$, $1\leq i\leq p$. $x_1<x_2<\cdots<x_p$ are distinct values of the covariate. μ is unknown.

1. Ordinal Covariate. Suppose the $\{x_i\}$ are equally-spaced. To penalize departures in m from a local polynomial of degree h_0-1 in the $\{x_i\}$, we take A $(?\times p)$ to be the h_0th difference matrix.

   Explicitly, define

   $$\underbrace{\Delta}_{(g-1)\times g}(g)=\{\delta_{u,w}\}, \qquad \delta_{u,u}=1,\ \delta_{u,u+1}=-1,\ \text{all other } \delta_{u,w}=0$$

   $$=\begin{pmatrix}1 & -1 & 0 & \cdots & 0\\ 0 & 1 & -1 & \cdots & 0\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots 1 & -1 & \end{pmatrix}$$

   Define recursively

   $$\begin{aligned}\text{1st difference} \quad \underbrace{D(1,p)}_{(p-1)\times p} &= \Delta(p)\\ \text{2nd difference} \quad \underbrace{D(2,p)}_{(p-2)\times p} &= \Delta(p-1)D(1,p)\\ &\vdots\\ h\text{th difference} \quad \underbrace{D(h,p)}_{(p-h+1)\times p} &= \Delta(p-h+1)D(h-1,p), \quad 2\leq h\leq p-1\end{aligned}$$

   Set

   $$\underbrace{A}_{(p-h_0)\times p}=D(h_0,p)$$

   to achieve the goal of penalizing departures from a local polynomial of degree h_0-1.

   Let $c=(x_1,x_2,\ldots,x_p)'$. Write $c^h=(x_1^h,x_2^h,\ldots,x_p^h)'$. Then $Ac^h=\mathbf{0}$ for $0\leq h\leq h_0-1$. $u\ \leftrightarrow\ c^0$.

   More generally, if the elements of c (the distinct covariate values) are not equally spaced, we have to generalize the concept of differencing. We construct A to satisfy three conditions:

   (i) For every possible i, the elements in row i of A that are not in columns $i,i+1,\ldots,i+h_0$ are zero.

(ii) $Ac^h = 0$ for $0 \le h \le h_0 - 1$.

(iii) Each row vector in A has length 1.

These 3 conditions are achieved by putting the nonzero elements in row i to be the basis vector of degree h_0 in the orthonormal polynomial (MATLAB: `orth`) on the $h_0 + 1$ design points $(x_i, x_{i+1}, \ldots, x_{i+h_0})$.

Note: When the $\{x_i\}$ are equally spaced, this general construction of A yields a multiple of $D(h_0, p)$.

2. Nominal Covariate. The values of a nominal covariate are merely labels that can be permuted without loss of information. (e.g. one-way ANOVA.) The candidate PLS estimators should also be invariant under permutation of nominal covariates. This motivates setting $A = I_p - uu' \equiv H$ (from the ANOVA discussion), with $u = p^{-1/2}(1, 1, \ldots, 1)'$. (Remark: think of the final project as an extension of this to two-way ANOVA.)

17.2 Simple PLS with Two Covariates ⇔ Two-Way Layout

(complete design)
Model: $y = \underbrace{C}_{n\times p}\underbrace{m}_{p\times 1} + e$
Covariate k $(k = 1, 2)$ has p_k distinct values, $x_{k1} < x_{k2} < \cdots < x_{k,p_k}$.
Let $\mathcal{I}$ be all pairs $i = (i_1, i_2)$ such that $1 \le i_k \le p_k$ for $k = 1, 2$. WLOG, order the $p = p_1 p_2$ elements of $\mathcal{I}$ in mirrored dictionary order. Let $x_i = (x_{1i_1}, x_{2i_2})$.

Assume that m is such that $m_i = \mu(x_i)$, where μ is unknown.

17.3 Comments on the Final Project

If you get stuck, look at what we did for the rat litter data and the hypercubed LS and how setting $d = 0$ or $d = 1$ leads to submodel fits.

Each part gets equal value, 9 points (except the last part, which gets 8).

17.4 Section 11-22-11

17.4.1 Lab 6 Comments

(b)

$$X(d) = \begin{pmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^d \\ 1 & x_2 & x_2^2 & \cdots & x_2^d \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_p & x_p^2 & \cdots & x_p^d \end{pmatrix}_{p\times(d+1)}$$

$\operatorname{rank}(X(d)) \stackrel{?}{=} d + 1$. If $\operatorname{rank}(X(d)) < d + 1$, then there exists $\gamma \in \mathbb{R}^{d+1}$ such that $X(d)\gamma = \mathbf{0}$. $\Rightarrow$ polynomial $\sum_{j=0}^{d} \gamma_j x^j = 0$ has p roots, $x_1, \ldots, x_p$. But it should have at most d roots $\Rightarrow\Leftarrow$.

(d)

$$\min_m \underbrace{|y - Cm|^2 + \lambda |Dm|^2}_{S(m)}$$

$$S(m) = y'y - 2y'Cm + m'(C'C + \lambda D'D)m$$

$$\frac{\partial S(m)}{\partial m} = -2C'y + 2(C'C + \lambda D'D)m = 0$$

$\underbrace{C'C}_{\text{pos. def.}} + \underbrace{\lambda}_{\geq 0} \underbrace{D'D}_{\text{pos. def.}}$ is invertible.

17.4.2 Lab 7 Comments

(c) $V_4 = [V_{1:79}\ V_{80:p}],\ z = [z'_{1:79}\ z'_{80:p}]'$. Then

$$\sum_{i=80}^{p} z_i^2 = |z_{80:p}|^2 = |V'_{80:p}y|^2$$

$$\mathbb{E}\left(\sum_{i=80}^{p} z_i^2\right) = \mathbb{E}|V'_{80:p}y|^2 = \mathbb{E}(y'V_{80:p}V'_{80:p}\underbrace{y}_{m+e})$$

$$= m'V_{80:p}V'_{80:p}m + \mathbb{E}(e'V_{80:p}V'_{80:p}e)$$

$$\mathbb{E}(e'V_{80:p}V'_{80:p}e) = \mathbb{E}(\operatorname{tr}(e'V_{80:p}V'_{80:p}e)) = \operatorname{tr}(\underbrace{\mathbb{E}(ee')}_{\sigma^2 I} V_{80:p}V'_{80:p})$$

$$= \sigma^2 \operatorname{tr}(V'_{80:p}V_{80:p}) = (p-79)\sigma^2$$

$$\Rightarrow \quad \mathbb{E}(\hat{\sigma}_H^2) = \sigma^2 + \frac{m'V_{80:p}V'_{80:p}m}{p-79}$$

Claim: when $\lim_{p\to\infty} \frac{m'V_{80:p}V'_{80:p}m}{p-79} = 0$, we have $\mathbb{E}|\hat{\sigma}_H^2 - \sigma^2| \to 0$.

Proof. First, $\operatorname{Cov}(z_{80:p}) = \operatorname{Cov}(V'_{80:p}y) = \sigma^2 I_{p-79}$. This means the z_i's are uncorrelated, so

$$\operatorname{Var}(\hat{\sigma}_H^2) = \frac{\sum_{i=80}^{p} \operatorname{Var}(z_i^2)}{(p-79)^2} \leq \frac{k}{p-79} \to 0$$

$$\text{where} \quad k = \max_i \operatorname{Var}(z_i^2) < \infty \text{ because of the finite fourth moment assumption}$$

$$\operatorname{Var}(x) = \mathbb{E}x^2 - (\mathbb{E}x)^2 = \mathbb{E}|x|^2 - (\mathbb{E}x)^2$$

$$\Rightarrow \quad \mathbb{E}|\hat{\sigma}_H^2 - \sigma^2| = \underbrace{(\mathbb{E}(\hat{\sigma}_H^2 - \sigma^2))^2}_{\to 0} + \underbrace{\operatorname{Var}(\hat{\sigma}_H^2)}_{\to 0}$$

So $\mathbb{E}|\hat{\sigma}_H^2 - \sigma^2|$ is satisfied. □

18 11-29-11

18.1 (Simple) PLS with 2 Covariates

Model:

$$\underbrace{y}_{n\times 1} = \underbrace{C}_{n\times p}\underbrace{m}_{p\times 1} + \underbrace{e}_{n\times 1}, \quad \text{rank}(C) = p, \quad C'C = \text{diag}\{n_i\}$$

Covariate k ($k = 1, 2$) takes on p_k distinct values, $x_{k_1} < x_{k_2} < \cdots < x_{k_{p_k}}$. Let $\mathcal{I}$ = all pairs (i_1, i_2) such that $1 \le i_k \le p_k$, $k = 1, 2$. WLOG, order the $p = p_1 p_2$ elements of $\mathcal{I}$ in mirror dictionary order. Let $x_i = (x_{1,i_1}, x_{2,i_2})$ for $i \in \mathcal{I}$. Then

$$\underbrace{m}_{p\times 1} = (m_1, m_2, \ldots, m_p)'$$

where $m_i = \mu(x_i)$, μ is unknown.

Candidate PLS Estimators

Let A_k be an annihilator for covariate k:

$$A_k u_k = 0, \qquad \underbrace{u_k}_{p_k\times 1} = p_k^{-1/2}(1, 1, \ldots, 1)', \qquad k = 1, 2$$

Let

$$B_1 = u_2 u_2' \otimes A_1 A_1', \qquad B_2 = A_2 A_2' \otimes u_1 u_1', \qquad B_{12} = A_2 A_2' \otimes A_1 A_1'$$

Let $\nu = (\nu_1, \nu_2, \nu_{12}) \in [0, \infty)^3$. Define

$$\begin{aligned}\hat{m}(\nu) &= \underset{m\in\mathbb{R}^p}{\arg\min}\left[|y - Cm|^2 + \nu_1 m' B_1 m + \nu_2 m' B_2 m + \nu_{12} m' B_{12} m\right] \\ &= \left[C'C + \underbrace{\nu_1 B_1 + \nu_2 B_2 + \nu_{12} B_{12}}_{Q(\nu)}\right]^{-1} C'y\end{aligned}$$

- Candidate PLS estimators
- Choose ν_1, ν_2, ν_{12} to minimize estimated risk

18.1.1 Spectral Representation of $Q(\nu)$

Spectral representation:

$$A_1 A_1' = \sum_{a=1}^{s_1} \lambda_{1a} P_{1a}, \qquad \text{where } P_{11} = u_1 u_1', \quad 0 = \lambda_{11} \le \lambda_{12} \le \cdots \le \lambda_{1 s_1}$$

$$A_2 A_2' = \sum_{b=1}^{s_2} \lambda_{2b} P_{2b}, \qquad \text{where } P_{21} = u_2 u_2', \quad 0 = \lambda_{21} \le \lambda_{22} \le \cdots \le \lambda_{2 s_2}$$

Let

$$\delta_{jk} = \begin{cases} 1 & j = k \\ 0 & j \ne k \end{cases}$$

Then

$$\begin{aligned}
B_1 &= \underbrace{P_{21}}_{u_2 u_2'} \otimes A_1 A_1' \\
&= \sum_{a=1}^{s_1} \lambda_{1a} (P_{21} \otimes P_{1a}) \\
&= \sum_{a=1}^{s_1} \sum_{b=1}^{s_2} \lambda_{1a} \delta_{b1} (P_{2b} \otimes P_{1a}) \\
B_2 &= \sum_{b=1}^{s_2} \lambda_{2b} (P_{2b} \otimes P_{11}) \\
&= \sum_{a=1}^{s_1} \sum_{b=1}^{s_2} \delta_{a1} \lambda_{2b} (P_{2b} \otimes P_{1a}) \\
B_{12} &= A_2 A_2' \otimes A_1 A_1' \\
&= \sum_{a=1}^{s_1} \sum_{b=1}^{s_2} \lambda_{1a} \lambda_{2b} (P_{2b} \otimes P_{1b})
\end{aligned}$$

Hence,

$$Q(\nu) = \nu_1 B_1 + \nu_2 B_2 + \nu_{12} B_{12} = \underbrace{\sum_{a=1}^{s_1} \sum_{b=1}^{s_2} t_{ab} (P_{2b} \otimes P_{1a})}_{\text{spectral rep. of } Q(\nu)}$$

where

$$t_{ab} = t_{ab}(\nu) = \nu_1 \lambda_{1a} \delta_{b1} + \nu_2 \delta_{a1} \lambda_{2b} + \nu_{12} \lambda_{1a} \lambda_{1b}$$

So $t_{ab} \in [0, \infty)^{s_1 s_2}$, i.e. $t_{ab} \in \mathcal{T} \subset [0, \infty)^{s_1 s_2}$. Continue much as for 1 covariate.

18.2 Sketch of Supporting Asymptotics

(c.f. Beran (2007) JSPI)

Model:

$$y = \underbrace{X}_{n \times p} \underbrace{\beta}_{p \times 1} + e, \qquad \operatorname{rank}(X) = p \leq n$$

e satisfies strong Gauss-Markov. Let

$$H = X'X, \quad U = XH^{-1/2} \;\Rightarrow\; U'U = I_p$$

The candidate estimator is

$$\hat{\eta}(t) = US(t)U'y$$

(this is a symmetric linear estimator in canonical form) $t \in [0,1]^s$, or more generally $t \in \mathcal{T}$ is a closed subset of $[0,1]^s$. Let

$$\underbrace{z}_{p \times 1} = U'y$$

Then $\mathbb{E}(z) = \xi = U'n,\ \operatorname{Cov}(z) = \sigma^2 I_p$.

Loss:

$$\begin{aligned} L(\hat{\eta}(t), \eta) &= p^{-1}|\hat{\eta}(t) - \eta|^2 \\ &= p^{-1}|US(t)z - \underbrace{U\xi}_{\eta}|^2 \\ &= p^{-1}|S(t)z - \xi|^2 \end{aligned}$$

Let $T(t) = S^2(t),\ \overline{T}(t) = [I_p - S(t)]^2$.

Risk:

$$\begin{aligned} R(\hat{\eta}(t), \eta, \sigma^2) &= \mathbb{E}L(\hat{\eta}(t), \eta) \\ &= p^{-1} \operatorname{tr}\left[\sigma^2 T(t) + \overline{T}(t)\xi\xi'\right] \end{aligned}$$

Let $\hat{\sigma}^2$ be an L_1-consistent estimator of σ^2 as $p \to \infty$. Estimate $\xi\xi'$ by $zz' - \hat{\sigma}^2 I_p$. The estimated risk is

$$\hat{r}(t) = p^{-1} \operatorname{tr}\left[\hat{\sigma}^2 T(t) + \overline{T}(t)(zz' - \hat{\sigma}^2 I_p)\right].$$

18.3 Assumptions for the Asymptotes

1. $\mathcal{T} = [0,1]^s$. The symmetric matrices $\{S(t) \mid t \in \mathcal{T}\}$ satisfy

   $$\sup_p \sup_{t \in \mathcal{T}} |S(t)|_{\text{sp}} < \infty$$

   ($|B|_{\text{sp}} = \sup_{|x| \neq 0} \frac{|Bx|_{\text{Euclidean}}}{|x|}$.) $S(t)$ is continuous on $\mathcal{T}$ and is differentiable on the interior of $\mathcal{T}$ with partial derivatives $\nabla_i S(t) = \frac{\partial S(t)}{\partial t_i},\ 1 \leq i \leq s$.

2. The strong Gauss-Markov model holds.

3. Under the strong Gauss-Markov model,

   $$\lim_{p \to \infty} \sup_{p^{-1}|\eta|^2 \leq a} \mathbb{E}|\hat{\sigma}^2 - \sigma^2| = 0$$

   for every finite $a > 0$ and $\sigma^2 > 0$.

Theorem 18.1.

Suppose Assumptions 1-3 hold. Let $W(t)$ be either the loss $L(\hat{\eta}(t), \eta)$ or the estimated risk $\hat{r}(t)$ of $\hat{\eta}(t)$. Then for all finite $a > 0$ and $\sigma^2 > 0$,

$$\lim_{p \to \infty} \sup_{p^{-1}|\eta|^2 \leq a} \mathbb{E}\left[\sup_{t \in \mathcal{T}} |W(t) - r(t)|\right] = 0$$

where $r(t) = R(\hat{\eta}(t), \eta, \sigma^2)$.

19 12-1-11

19.1 Asymptotics (Continued)

Theorem 19.1.

$$y = X\beta + e, \qquad \hat{\eta}(t) = US(t)U'y$$

Suppose Assumptions 1 to 3 hold. Let $W(t)$ denote either the loss $L(\hat{\eta}(t), \eta)$ or the estimated risk $\hat{r}(t)$ of $\hat{\eta}(t)$. Then for every finite $a > 0$ and $\sigma^2 > 0$,

$$\lim_{p\to\infty} \sup_{p^{-1}|\eta|^2 \le a} \mathbb{E}\left[\sup_{t\in\mathcal{T}} |W(t) - r(t)|\right] = 0$$

where $r(t) = R(\hat{\eta}(t), \eta, \sigma^2) =$ risk of $\hat{\eta}(t)$. $\eta = \mathbb{E}y = X\beta$, $|\eta|^2 = \beta' X' X\beta = |X\beta|^2$.

Proof. (Steps)

1. Pointwise convergence of $W(t) - r(t) \xrightarrow{p} 0$.
2. Show $\sup_{t\in\mathcal{T}} |W(t) - r(t)| \xrightarrow{p} 0$. Uses weak convergence in $C[0,1]^s$.
3. Strengthen this to $\mathbb{E}\sup_{t\in\mathcal{T}} |W(t) - r(t)| \to 0$ (uniform integrability).

□

Theorem 19.2.

Let $\hat{t}$ minimize the estimated risk $\hat{r}(t)$, and let $\tilde{t}$ minimize the risk $r(t) = R(\hat{\eta}(t), \eta, \sigma^2)$. Suppose Assumptions 1 to 3 hold. Then for every finite $a > 0$ and $\sigma^2 > 0$,

$$\lim_{p\to\infty} \sup_{p^{-1}|\eta|^2 \le a} |R(\hat{\eta}(\hat{t}), \eta, \sigma^2) - r(\tilde{t})| = 0$$

where $\hat{t} = \arg\min_{t\in\mathcal{T}} \hat{r}(t)$, $\tilde{t} = \arg\min_{t\in\mathcal{T}} r(t)$.

Moreover, for V equal to either the loss $L(\hat{\eta}(\hat{t}), \eta)$ or the risk $R(\hat{\eta}(\hat{t}), \eta, \sigma^2)$,

$$\lim_{p\to\infty} \sup_{p^{-1}|\eta|^2 \le 0} \mathbb{E}|V - \hat{r}(\hat{t})| = 0.$$

where

$$\begin{aligned} \hat{r}(\hat{t}) &= \min_{t\in\mathcal{T}} \hat{r}(t) \\ &= \text{ estimated risk of the candidate estimator with smallest estimated risk} \end{aligned}$$

19.2 Summary

Theorem 19.1: The estimated risk is a function is a trustworthy approximation to the true risk function over $t \in \mathcal{T}$.

Theorem 19.2: (follows from Theorem 19.1) Hence

1. The risk of the adaptive estimator $\hat{\eta}(\hat{t})$ converges to the risk of the best estimator in the candidate class.
2. $\hat{r}(\hat{t})$ converges to this risk.

19.3 Statistics on Manifolds

Prominent researcher: Victor Patrangenara

Index