ST552: Linear Models and Variance Components

Mon/Wed 11:45am-1:00pm, Wed 1:30pm-2:45pm, SAS Hall 5270

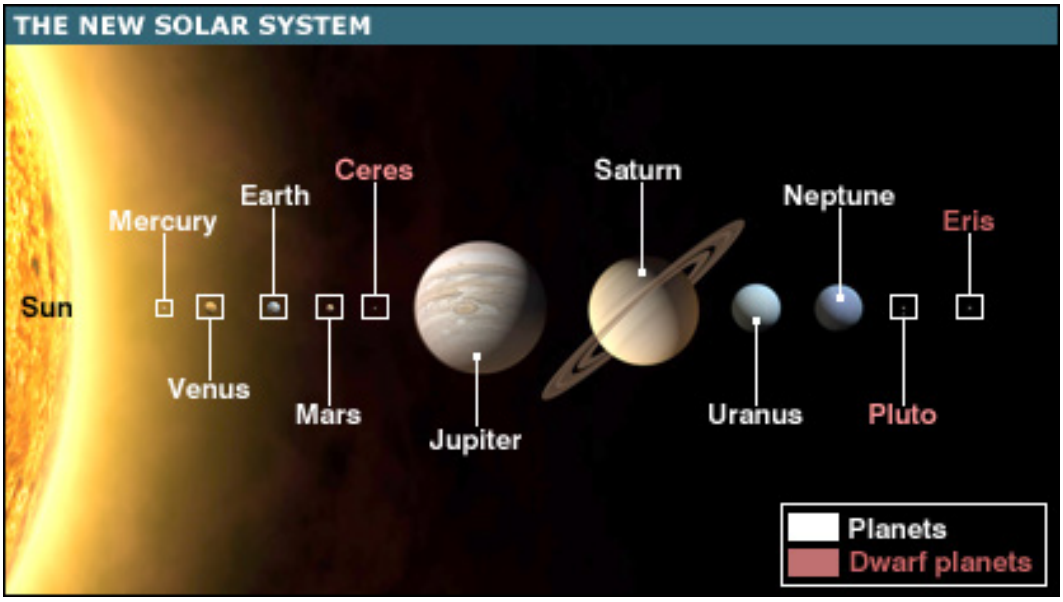
Instructor: Dr Hua Zhou, `hua_zhou@ncsu.edu`

1 Lecture 1: Aug 21

Today

- Introduction
- Course logistics
- Read JM Appendix A and chapter 1
- Linear algebra review
- No afternoon session today

How Gauss became famous?



- 1801, *Dr Carl Friedrich Gauss*, 24; proved Fundamental Theorem of Algebra; wrote the book *Disquisitiones Arithmetic*, which is still being studied today.
- 1801, Jan 1 - Feb 11 (41 days), astronomer Piazzi observed Ceres (a dwarf planet), which was then lost behind sun.
- 1801, Aug – Sep, futile search by top astronomers; Laplace claimed it unsolvable.
- 1801, Oct – Nov, young Gauss did calculations by *method of least squares*.
- 1801, Dec 31, astronomer von Zach relocated Ceres according to Gauss' calculation.
- 1802, *Summarische Übersicht der Bestimmung der Bahnen der beiden neuen Hauptplaneten angewandten Methoden*, considered the origin of linear algebra.
- 1807, Professor of Astronomy and Director of Göttingen Observatory in remainder of his life.
- 1809, *Theoria motus corporum coelestium in sectionibus conicis solem ambientum* (Theory of motion of the celestial bodies moving in conic sections around the Sun); birth of the Gaussian distribution, as an attempt to rationalize the method of least squares.

- 1810, Laplace consolidated importance of Gaussian distribution by proving the central limit theorem.
- 1829, Gauss-Markov Theorem. Under Gaussian error assumption (actually only uncorrelated and homoscedastic needed), least square solution is the best linear unbiased estimate (BLUE), i.e., it has the smallest variance and MSE among all linear unbiased estimators. Other estimators such as the James-Stein estimator may have smaller MSE, but they are *nonlinear*.

For more details

- `http://www.keplersdiscovery.com/Asteroid.html`
- Teets and Whitehead (1999)

ARTICLES

The Discovery of Ceres: How Gauss Became Famous

DONALD TEETS
KAREN WHITEHEAD
South Dakota School of Mines and Technology
Rapid City, SD 57701

"The Duke of Brunswick has discovered more in his country than a planet: a super-terrestrial spirit in a human body."

These words, attributed to Laplace in 1801, refer to the accomplishment of Carl Friedrich Gauss in computing the orbit of the newly discovered planetoid *Ceres Ferdinandea* from extremely limited data. Indeed, although Gauss had already achieved some fame among mathematicians, it was his work on the Ceres orbit that "made Gauss a European celebrity—this a consequence of the popular appeal which astronomy has always enjoyed..." **[2]**. The story of Gauss's work on this problem is a good one and is often told in biographical sketches of Gauss (e.g., **[2]**, **[3]**, **[6]**), but the mathematical details of how he solved the problem are invariably omitted from such historical works. We are left to wonder, how did he do it? *Just how did Gauss*

Gauss' story

- Motivated by a real problem.
- Heuristic solution: method of least squares.
- Solution readily verifiable: Ceres was re-discovered!

- Algorithmic development: linear algebra, Gaussian elimination, FFT (fast Fourier transform).
- Theoretical justification: Gaussian distribution, Gauss-Markov theorem.

What is this course about?

This course focuses on the "theoretical" aspect of the method of least squares and statistical inference under the normal assumption. Read JM chapter 1 for a few specific examples of linear models.

A hierarchy of linear models

- The linear mean model:

  $$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e},$$

  where $\mathbf{E}(\boldsymbol{e}) = \mathbf{0}$. Only assumption is that errors have mean 0.

- Gauss-Markov model:

  $$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e},$$

  where $\mathbf{E}(\boldsymbol{e}) = \mathbf{0}$ and $\mathbf{Var}(\boldsymbol{e}) = \sigma^2\boldsymbol{I}$ (uncorrelated errors with constant variance).

- Aitken model or general linear model:

  $$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e},$$

  where $\mathbf{E}(\boldsymbol{e}) = \mathbf{0}$ and $\mathbf{Var}(\boldsymbol{e}) = \sigma^2\boldsymbol{V}$. $\boldsymbol{V}$ is fixed and known.

- Linear models with joint normal errors: $\boldsymbol{y} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$ or $\boldsymbol{y} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{V})$ with $\boldsymbol{V}$ known.

- Variance components model: $\boldsymbol{y} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma_1^2\boldsymbol{V}_1 + \cdots \sigma_r^2\boldsymbol{V}_r)$ with $\boldsymbol{V}_1, \ldots, \boldsymbol{V}_r$ known.

- Multivariate linear model: $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{E}$.

- Generalized linear models (GLMs). Logistic regression, probit regression, log-linear model (Poisson regression), ... Note the difference from the general linear model. GLMs are generalization of the *concept* of linear models. They are typically covered in a categorical data analysis course.

Syllabus

Check course website frequently for updates and announcements.
`http://hua-zhou.github.io/teaching/st552-2013fall/`
Lecture notes will be updated and posted after each lecture.

Vector and vector space

- A set of vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are *linearly dependent* if there exist coefficients c_j such that $\sum_{j=1}^n c_j \boldsymbol{x}_j = \boldsymbol{0}$ and $\|\boldsymbol{c}\|_2 = \sum_{j=1}^n c_j^2 > 0$. They are *linearly independent* if $\sum_{j=1}^n c_j \boldsymbol{x}_j = \boldsymbol{0}$ implies $c_j = 0$ for all j.

- A *vector space* $\mathcal{V}$ is a set of vectors that are closed under addition and scalar multiplication (closed under *axpy* operation). Any vector space must contain the zero vector $\boldsymbol{0}$ (why?).

- A vector space $\mathcal{V}$ is *generated* or *spanned* by a set of vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, written as $\mathcal{V} = \text{span}\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$, if any vector $\boldsymbol{x}$ in the vector space is a linear combination of $\boldsymbol{x}_i$, $i = 1, \ldots, n$.

- A set of linearly independent vectors that generate or span a space $\mathcal{V}$ is called a *basis* of $\mathcal{V}$.

- Order and dimension. The *order* of a vectors space is simply the length of the vectors in that space. The *dimension* of a vector space is the maximum number of linearly independent vectors in that space.

- Two vector spaces $\mathcal{V}_1$ and $\mathcal{V}_2$ are *essentially disjoint* if the only element in $\mathcal{V}_1 \cap \mathcal{V}_2$ is the zero vector $\boldsymbol{0}$.

- If $\mathcal{V}_1$ and $\mathcal{V}_2$ are two vector spaces, then $\mathcal{V}_1 \cap \mathcal{V}_2$ are vector spaces.

- If $\mathcal{V}_1$ and $\mathcal{V}_2$ are two vector spaces of same order, then $\mathcal{V}_1 + \mathcal{V}_2 = \{\boldsymbol{v} : \boldsymbol{v} = \boldsymbol{v}_1 + \boldsymbol{v}_2, \boldsymbol{v}_1 \in \mathcal{V}_1, \boldsymbol{v}_2 \in \mathcal{V}_2\}$ is a vector space. If furthermore $\mathcal{V}_1$ and $\mathcal{V}_2$ are essentially disjoint, the sum is called the *direct sum* and denoted by $\mathcal{V}_1 \oplus \mathcal{V}_2$.

- If $\mathcal{V}_1$ and $\mathcal{V}_2$ are two vector spaces, $\mathcal{V}_1 \cup \mathcal{V}_2$ is not necessarily a vector space. (Exercise: find a counter example).

- Affine spaces. Consider a system of m linear equations in variable $\boldsymbol{x} \in \mathbb{R}^n$

$$\begin{aligned} \boldsymbol{c}_1^T \boldsymbol{x} &= b_1 \\ \vdots &= \vdots \\ \boldsymbol{c}_m^T \boldsymbol{x} &= b_m, \end{aligned}$$

  where $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_m \in \mathbb{R}^n$ are linearly independent (and hence $m \leq n$). The set of solutions is called an *affine space*. The intersection of two affine spaces is an affine space (why?). If the zero vector $\mathbf{0}_n$ belongs to the affine space, i.e., $b_1 = \cdots = b_m = 0$, then it is a vector space. Thus any affine space containing the origin $\mathbf{0}$ is a vector space, but other affine spaces are not vector spaces.

- If $m = 1$, the affine space is called a *hyperplane*. A hyperplane through the origin is an $(n-1)$-dimensional vector space.

- If $m = n - 1$, the affine space is a line. A line through the origin is a one-dimensional vector space.

- The mapping $\boldsymbol{x} \mapsto \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}$ is called an *affine function*. If $\boldsymbol{b} = \mathbf{0}$, it is called a *linear function*.

Orthogonality and orthogonalization

- Vector $\boldsymbol{x}_1$ is orthogonal to another vector $\boldsymbol{x}_2$, denoted by $\boldsymbol{x}_1 \perp \boldsymbol{x}_2$, if $\langle \boldsymbol{x}_1, \boldsymbol{x}_2 \rangle = \boldsymbol{x}_1^{\mathsf{T}} \boldsymbol{x}_2 = 0$. They are *orthonormal* if $\boldsymbol{x}_1 \perp \boldsymbol{x}_2$ and $\|\boldsymbol{x}_1\|_2 = \|\boldsymbol{x}_2\|_2 = 1$.

- The projection of a vector $\boldsymbol{x}_2$ onto the vector $\boldsymbol{x}_1$ is

$$\hat{\boldsymbol{x}}_2 = \frac{\langle \boldsymbol{x}_2, \boldsymbol{x}_1 \rangle}{\|\boldsymbol{x}_1\|_2^2} \boldsymbol{x}_1 = \langle \tilde{\boldsymbol{x}}_1, \boldsymbol{x}_2 \rangle \tilde{\boldsymbol{x}}_1,$$

  where $\tilde{\boldsymbol{x}}_1 = \boldsymbol{x}_1 / \|\boldsymbol{x}_1\|_2$ is the normalized $\boldsymbol{x}_1$ vector.

- Gram-Schmidt transformation: orthonormalize two vectors $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$.

$$\begin{aligned} \tilde{\boldsymbol{x}}_1 &= \frac{1}{\|\boldsymbol{x}_1\|_2} \boldsymbol{x}_1 \\ \tilde{\boldsymbol{x}}_2 &= \frac{1}{\|\boldsymbol{x}_2 - \langle \tilde{\boldsymbol{x}}_1, \boldsymbol{x}_2 \rangle \tilde{\boldsymbol{x}}_1\|_2} (\boldsymbol{x}_2 - \langle \tilde{\boldsymbol{x}}_1, \boldsymbol{x}_2 \rangle \tilde{\boldsymbol{x}}_1) \end{aligned}$$

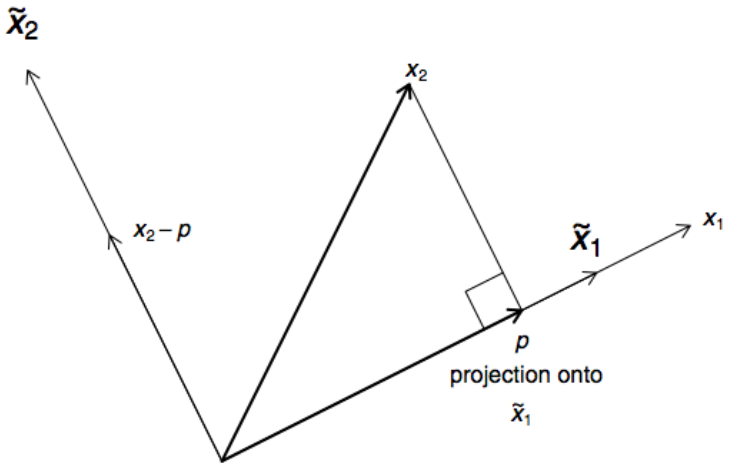


Fig. 2.2. Orthogonalization of x_1 and x_2

- Gram-Schmidt algorithm orthonormalizes a set of of vectors.
- A set of nonzero, mutually orthogonal vectors are linearly independent. Proof by contradiction.
- Two vector spaces $\mathcal{V}_1$ and $\mathcal{V}_2$ are *orthogonal*, written $\mathcal{V}_1 \perp \mathcal{V}_2$, if each vector in $\mathcal{V}_1$ is orthogonal to every vector in $\mathcal{V}_2$.
- The intersection of two orthogonal vector spaces consists only of the zero vector $\mathbf{0}$.
- If $\mathcal{V}_1 \perp \mathcal{V}_2$ and $\mathcal{V}_1 \oplus \mathcal{V}_2 = \mathbb{R}^n$, then $\mathcal{V}_2$ is the *orthogonal complement* of $\mathcal{V}_1$. This is written as $\mathcal{V}_2 = \mathcal{V}_1^{\perp}$.

Rank

Assume $\boldsymbol{A} \in \mathbb{R}^{m \times n}$.

- $\text{rank}(\boldsymbol{A})$ is the maximum number of linearly independent rows of a matrix.
- Alternatively, $\text{rank}(\boldsymbol{A})$ is the maximum number of linearly independent columns of a matrix. (Exercise: show that these two definitions are equivalent. Hint: form the full rank partition of $\boldsymbol{A}$.)
- $\text{rank}(\boldsymbol{A}) \leq \min\{m, n\}$.

- A matrix is *full rank* if $\text{rank}(\boldsymbol{A}) = \min\{m, n\}$. It is *full row rank* if $\text{rank}(\boldsymbol{A}) = m$. It is *full column rank* if $\text{rank}(\boldsymbol{A}) = n$.
- A square matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ is *singular* if $\text{rank}(\boldsymbol{A}) < n$ and *non-singular* if $\text{rank}(\boldsymbol{A}) = n$.
- $\text{rank}(\boldsymbol{A}) = \text{rank}(\boldsymbol{A}^T) = \text{rank}(\boldsymbol{A}^T\boldsymbol{A}) = \text{rank}(\boldsymbol{A}\boldsymbol{A}^T)$. (Exercise: show it.)
- $\text{rank}(\boldsymbol{A}\boldsymbol{B}) \leq \min\{\text{rank}(\boldsymbol{A}), \text{rank}(\boldsymbol{B})\}$. (Hint: Columns of $\boldsymbol{A}\boldsymbol{B}$ are spanned by columns of $\boldsymbol{A}$ and rows of $\boldsymbol{A}\boldsymbol{B}$ are spanned by rows of $\boldsymbol{B}$.)
- $\text{rank}(\boldsymbol{A}\boldsymbol{B}) = \text{rank}(\boldsymbol{A})$ if $\boldsymbol{B}$ is square and of full rank. More general, pre-multiplying by a matrix with full column rank or post-multiplying by a matrix with full row rank does not change rank. (Exercise: show it.)
- $\text{rank}(\boldsymbol{A} + \boldsymbol{B}) \leq \text{rank}(\boldsymbol{A}) + \text{rank}(\boldsymbol{B})$.

$$\boldsymbol{A} + \boldsymbol{B} = (\boldsymbol{A}\,\boldsymbol{B}) \begin{pmatrix} \boldsymbol{I}_n \\ \boldsymbol{I}_n \end{pmatrix}.$$

- If $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}_m$ for some $\boldsymbol{x} \neq \boldsymbol{0}_n$, then $\text{rank}(\boldsymbol{A}) \leq n - 1$.

2 Lecture 2: Aug 26

Last time

- How Gauss became famous? Method of least squares, Gauss-Markov theorem, Gaussian distribution, ...
- What do we mean by "linear models"? Different layers of assumptions: linear mean model, Gauss-Markov model, Aitken or general linear model, Gaussian assumption, variance components (linear mixed) model, multivariate linear model, GLM, ...
- Course logistics
- Linear algebra: vector and vector space, rank of a matrix

Today

- Continue review of linear algebra
- TA office hours are posted
- Homework 1 is posted, due Sep 4
- Reach JM chapter 2

Column space and null space

Assume $\boldsymbol{A} \in \mathbb{R}^{m \times n}$.

- The *column space* of a matrix $\boldsymbol{A}$, denoted by $\mathcal{C}(\boldsymbol{A})$, is the vector space (of order m) spanned by the columns of the matrix. Other names: *range* of $\boldsymbol{A}$, or the *manifold* of $\boldsymbol{A}$.
- The *null space* of a matrix $\boldsymbol{A}$, denoted by $\mathcal{N}(\boldsymbol{A})$, is the vector space (of order n) $\{\boldsymbol{y} : \boldsymbol{A}\boldsymbol{y} = \boldsymbol{0}\}$.
- $\dim(\mathcal{C}(\boldsymbol{A})) = r$ and $\dim(\mathcal{N}(\boldsymbol{A})) = n - r$, where $r = \text{rank}(\boldsymbol{A})$.
  See JM Theorem A.1 for the proof.
  Interpretation: "dimension of column space + dimension of null space = #

columns"

*Mis*interpretation: Columns space and null space are orthogonal complement to each other. They are of different orders in general! Next result gives the correct statement.

- $\mathcal{C}(\boldsymbol{A})$ and $\mathcal{N}(\boldsymbol{A}^T)$ are orthogonal complement to each other in $\mathbb{R}^m$: $\mathcal{C}(\boldsymbol{A}) = \mathcal{N}(\boldsymbol{A}^T)^\perp$.

  Proof: (1) Orthogonality is trivial. (2) (Essentially disjoint) For any $\boldsymbol{v} \in \mathcal{C}(\boldsymbol{A}) \cap \mathcal{N}(\boldsymbol{A}^T)$, there exists a $\boldsymbol{c}$ such that $\boldsymbol{A}\boldsymbol{c} = \boldsymbol{v}$ and $\boldsymbol{A}^T\boldsymbol{v} = \boldsymbol{0}_n$. Therefore $\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{c} = \boldsymbol{A}^T\boldsymbol{v} = \boldsymbol{0}_n$ and thus $\boldsymbol{c}^T\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{c} = 0$, implying that $\boldsymbol{v} = \boldsymbol{A}\boldsymbol{c} = \boldsymbol{0}$. Therefore $\mathcal{C}(\boldsymbol{A}) \cap \mathcal{N}(\boldsymbol{A}^T) = \{\boldsymbol{0}\}$. (3) Since $\dim(\mathcal{C}(\boldsymbol{A})) = \text{rank}(\boldsymbol{A})$ and $\dim(\mathcal{N}(\boldsymbol{A}^T)) = m - \text{rank}(\boldsymbol{A})$, we have $\mathcal{C}(\boldsymbol{A})) \cup \mathcal{N}(\boldsymbol{A}^T) = \mathbb{R}^m$.

- $\mathcal{C}(\boldsymbol{A}\boldsymbol{B}) \subset \mathcal{C}(\boldsymbol{A})$.

- If $\mathcal{C}(\boldsymbol{B}) \subset \mathcal{C}(\boldsymbol{A})$, then there exists a matrix $\boldsymbol{C}$ such that $\boldsymbol{B} = \boldsymbol{A}\boldsymbol{C}$.

- $\mathcal{C}(\boldsymbol{A}\boldsymbol{A}^T) = \mathcal{C}(\boldsymbol{A})$ and thus $\text{rank}(\boldsymbol{A}\boldsymbol{A}^T) = \text{rank}(\boldsymbol{A})$. (Exercise: show it.)
  "Multiplication by its transpose does not change the rank."

Trace of a square matrix

Assume $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is a square matrix.

- $\text{tr}(\boldsymbol{A}) = \sum_{i=1}^n a_{ii}$ is the sum of diagonal entries.

- $\text{tr}(\boldsymbol{A}^T) = \text{tr}(\boldsymbol{A})$.

- $\text{tr}(\boldsymbol{A} + \boldsymbol{B}) = \text{tr}(\boldsymbol{A}) + \text{tr}(\boldsymbol{B})$.

- $\text{tr}(c\boldsymbol{A}) = c\text{tr}(\boldsymbol{A})$, where c is a scalar.

- Invariance of trace function under cyclic permutation: $\text{tr}(\boldsymbol{A}\boldsymbol{B}) = \text{tr}(\boldsymbol{B}\boldsymbol{A})$ for $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ and $\boldsymbol{B}^{n\times m}$. In general, $\text{tr}(\boldsymbol{A}_1 \cdots \boldsymbol{A}_k) = \text{tr}(\boldsymbol{A}_{j+1} \cdots \boldsymbol{A}_k \boldsymbol{A}_1 \cdots \boldsymbol{A}_j)$, $j = 1, \ldots, k-1$, for matrices of compatible sizes.

Inner (dot) product between two matrices

- For $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{m \times n}$, $\langle \boldsymbol{A}, \boldsymbol{B} \rangle := \operatorname{tr}(\boldsymbol{A}^T \boldsymbol{B}) = \operatorname{tr}(\boldsymbol{B}\boldsymbol{A}^T)$. Analog of the inner product between two vectors $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$.
- $\langle \boldsymbol{A}, \boldsymbol{B} \rangle = \langle \boldsymbol{A}^T, \boldsymbol{B}^T \rangle$.
- We say $\boldsymbol{A}$ is orthogonal to $\boldsymbol{B}$ if $\langle \boldsymbol{A}, \boldsymbol{B} \rangle = 0$.
- *Trace norm* (or *Frobenius norm*, or *Euclidean norm*) is the norm on the space of $m \times n$ matrices induced by the inner product: $\|\boldsymbol{A}\|_{\mathrm{F}} = (\langle \boldsymbol{A}, \boldsymbol{A} \rangle)^{1/2} = \operatorname{tr}(\boldsymbol{A}^T \boldsymbol{A})^{1/2} = (\sum_{i,j} a_{ij}^2)^{1/2}$.
- Cauchy-Schwartz: $\langle \boldsymbol{A}, \boldsymbol{B} \rangle \leq \langle \boldsymbol{A}, \boldsymbol{A} \rangle^{1/2} \langle \boldsymbol{B}, \boldsymbol{B} \rangle^{1/2}$.
  Proof: Expand $\operatorname{tr}((\boldsymbol{A} - x\boldsymbol{B})^T (\boldsymbol{A} - x\boldsymbol{B})) \geq 0$.

Matrix inverses

Assume $\boldsymbol{A} \in \mathbb{R}^{m \times n}$.

- The *Moore-Penrose inverse* of $\boldsymbol{A}$ is a matrix $\boldsymbol{A}^+ \in \mathbb{R}^{n \times m}$ with following properties
  (a) $\boldsymbol{A}\boldsymbol{A}^+\boldsymbol{A} = \boldsymbol{A}$. (*Generalized inverse*, g_1 *inverse*, or *inner pseudo-inverse*)
  (b) $\boldsymbol{A}^+\boldsymbol{A}\boldsymbol{A}^+ = \boldsymbol{A}^+$. (*Outer pseudo-inverse*. Any g_1 inverse that satisfies this condition is called a g_2 *inverse*, or *reflexive generalized inverse* and is denoted by $\boldsymbol{A}^*$.)
  (c) $\boldsymbol{A}^+\boldsymbol{A}$ is symmetric.
  (d) $\boldsymbol{A}\boldsymbol{A}^+$ is symmetric.
- $\boldsymbol{A}^+$ exists and is unique for any matrix $\boldsymbol{A}$.
- *Generalized inverse* (or g_1 *inverse*, denoted by $\boldsymbol{A}^-$ or $\boldsymbol{A}^g$): property (a).
- g_2 *inverse* (denoted by $\boldsymbol{A}^*$): properties (a)+(b).
- *Moore-Penrose inverse* (denoted by $\boldsymbol{A}^+$): properties (a)+(b)+(c)+(d).
- If $\boldsymbol{A}$ is square and full rank, then the generalized inverse is unique and denoted by $\boldsymbol{A}^{-1}$ (inverse).

- How to find a generalized inverse (may be non-unique) for a matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ of rank r? Permute rows and columns to form the *full rank partitioning*

$$\boldsymbol{PAQ} = \begin{pmatrix} \boldsymbol{C}_{r\times r} & \boldsymbol{D}_{r\times(n-r)} \\ \boldsymbol{E}_{(m-r)\times r} & \boldsymbol{F}_{(m-r)\times(n-r)} \end{pmatrix},$$

  where $\boldsymbol{C} \in \mathbb{R}^{r\times r}$ is of full rank. Then the matrix

$$\boldsymbol{Q} \begin{pmatrix} \boldsymbol{C}^{-1}_{r\times r} & \boldsymbol{0}_{r\times(m-r)} \\ \boldsymbol{0}_{(n-r)\times r} & \boldsymbol{0}_{(n-r)\times(m-r)} \end{pmatrix} \boldsymbol{P}$$

  is a generalized inverse of $\boldsymbol{A}$. See JM Result A.11 (p247) for the argument.

- In practice, the Moore-Penrose inverse $\boldsymbol{A}^+$ is easily computed from the singular value decomposition (SVD) of $\boldsymbol{A}$.

- For any nonzero k, $(1/k)\boldsymbol{A}^-$ is a generalized inverse of $k\boldsymbol{A}$.

- $(\boldsymbol{A}^-)^T$ is a generalized inverse of $\boldsymbol{A}^T$.

- $\mathcal{C}(\boldsymbol{A}) = \mathcal{C}(\boldsymbol{AA}^-)$ and $\mathcal{C}(\boldsymbol{A}^T) = \mathcal{C}((\boldsymbol{A}^-\boldsymbol{A})^T)$.
  $\text{rank}(\boldsymbol{A}) = \text{rank}(\boldsymbol{AA}^-) = \text{rank}(\boldsymbol{A}^-\boldsymbol{A})$.
  "Multiplication by generalized inverse does not change rank."

  Proof. We already know $\mathcal{C}(\boldsymbol{A}) \supset \mathcal{C}(\boldsymbol{AA}^-)$. Now since $\boldsymbol{A} = \boldsymbol{AA}^-\boldsymbol{A}$, we also have $\mathcal{C}(\boldsymbol{A}) \subset \mathcal{C}(\boldsymbol{A}^-\boldsymbol{A})$. □

- $\text{rank}(\boldsymbol{A}^-) \geq \text{rank}(\boldsymbol{A})$. "Generalized inverse has equal or a larger rank than original matrix."

  Proof. $\text{rank}(\boldsymbol{A}) = \text{rank}(\boldsymbol{AA}^-\boldsymbol{A}) \leq \text{rank}(\boldsymbol{AA}^-) = \text{rank}(\boldsymbol{A}^-)$. □

System of linear equations

$\boldsymbol{Ax} = \boldsymbol{b}$ where $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, $\boldsymbol{x} \in \mathbb{R}^n$, $\boldsymbol{b} \in \mathbb{R}^m$.

- When is there a solution? The following statements are equivalent.

  1. The linear system $\boldsymbol{Ax} = \boldsymbol{b}$ has a solution (*consistent*)
  2. $\boldsymbol{b} \in \mathcal{C}(\boldsymbol{A})$.

3. $\text{rank}((\boldsymbol{A}, \boldsymbol{b})) = \text{rank}(\boldsymbol{A})$.
4. $\boldsymbol{A}\boldsymbol{A}^{-}\boldsymbol{b} = \boldsymbol{b}$.

Proof. Equivalence between 1, 2 and 3 is trivial. 4 implies 1: apparently $\tilde{\boldsymbol{x}} = \boldsymbol{A}^{-}\boldsymbol{b}$ is a solution to $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$. 1 implies 4: if $\tilde{\boldsymbol{x}}$ is a solution, then $\boldsymbol{b} = \boldsymbol{A}\tilde{\boldsymbol{x}} = \boldsymbol{A}\boldsymbol{A}^{-}\boldsymbol{A}\tilde{\boldsymbol{x}} = \boldsymbol{A}\boldsymbol{A}^{-}\boldsymbol{b}$. □

The last equivalence gives some intuition why $\boldsymbol{A}^{-}$ is called an inverse.

- How to characterize the solution set? Let's first study the homogenous case $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}$, which is always consistent (why?).
  $\tilde{\boldsymbol{x}}$ is a solution to $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}$ if and only if

$$\tilde{\boldsymbol{x}} = (\boldsymbol{I}_n - \boldsymbol{A}^{-}\boldsymbol{A})\boldsymbol{q}$$

  for some $\boldsymbol{q} \in \mathbb{R}^n$.

  Proof. "If": Apparently $(\boldsymbol{I}_n - \boldsymbol{A}^{-}\boldsymbol{A})\boldsymbol{q}$ is a solution regardless value of $\boldsymbol{q}$ since $\boldsymbol{A}(\boldsymbol{I}_n - \boldsymbol{A}^{-}\boldsymbol{A}) = \boldsymbol{A} - \boldsymbol{A} = \boldsymbol{0}_{m\times n}$. "Only if": If $\tilde{\boldsymbol{x}}$ is a solution, then $\tilde{\boldsymbol{x}} = (\boldsymbol{I}_n - \boldsymbol{A}^{-}\boldsymbol{A})\boldsymbol{q}$ by taking $\boldsymbol{q} = \tilde{\boldsymbol{x}}$. □

- Rephrasing above result we have $\mathcal{N}(\boldsymbol{A}) = \mathcal{C}(\boldsymbol{I}_n - \boldsymbol{A}^{-}\boldsymbol{A})$.

- Now we study the inhomogeneous case.
  If $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ is consistent, then $\tilde{\boldsymbol{x}}$ is a solution to $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ if and only if

$$\tilde{\boldsymbol{x}} = \boldsymbol{A}^{-}\boldsymbol{b} + (\boldsymbol{I}_n - \boldsymbol{A}^{-}\boldsymbol{A})\boldsymbol{q}$$

  for some $\boldsymbol{q} \in \mathbb{R}^n$.
  Interpretation: "a specific solution" + "a vector in the null space of $\boldsymbol{A}$".

  Proof.

$$\begin{aligned} & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \Leftrightarrow \boldsymbol{A}\boldsymbol{x} = \boldsymbol{A}\boldsymbol{A}^{-}\boldsymbol{b} \Leftrightarrow \boldsymbol{A}(\boldsymbol{x} - \boldsymbol{A}^{-}\boldsymbol{b}) = \boldsymbol{0} \\ \Leftrightarrow \quad & \boldsymbol{x} - \boldsymbol{A}^{-}\boldsymbol{b} = (\boldsymbol{I}_n - \boldsymbol{A}^{-}\boldsymbol{A})\boldsymbol{q} \Leftrightarrow \boldsymbol{x} = \boldsymbol{A}^{-}\boldsymbol{b} + (\boldsymbol{I}_n - \boldsymbol{A}^{-}\boldsymbol{A})\boldsymbol{q}. \end{aligned}$$

  □

- $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ is consistent for *all* $\boldsymbol{b}$ if and only if $\boldsymbol{A}$ has full row rank.

Proof. "If": $\dim(\mathcal{C}(\boldsymbol{A})) = \text{rank}(\boldsymbol{A}) = m$. Thus $\mathcal{C}(\boldsymbol{A}) = \mathbb{R}^m$ and contains any $\boldsymbol{b} \in \mathbb{R}^m$. "Only if": $\boldsymbol{A}\boldsymbol{A}^-\boldsymbol{b} = \boldsymbol{b}$ for any $\boldsymbol{b}$. Take $\boldsymbol{b} = \boldsymbol{e}_i$ gives $\boldsymbol{A}\boldsymbol{A}^- = \boldsymbol{I}_m$. Thus $m \geq \text{rank}(\boldsymbol{A}) \geq \text{rank}(\boldsymbol{A}\boldsymbol{A}^-) = m$. □

- If a system is consistent, its solution is unique if and only if $\boldsymbol{A}$ has full column rank.

  Proof. In previous proof, we see there is a one-to-one correspondence between the solution set to the inhomogenous system $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ and the solution set to the homogeneous system $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}$. Now $\boldsymbol{A}$ has full column rank if and only if $\dim(\mathcal{N}(\boldsymbol{A})) = n - \dim(\mathcal{C}(\boldsymbol{A}^T)) = n - \text{rank}(\boldsymbol{A}) = n - n = 0$ if and only if there is a unique solution to $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}$ if and only if there is a unique solution to $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$. □

- If $\boldsymbol{A}$ has full row and column rank, then $\boldsymbol{A}$ is non-singular and the unique solution is $\boldsymbol{A}^{-1}\boldsymbol{b}$.

3 Lecture 3: Aug 28

Last time

- Linear algebra: column and null spaces $\mathcal{C}(\boldsymbol{A}) = \mathcal{N}(\boldsymbol{A}^T)^\perp$, trace (invariance under cyclic permutation), generalized matrix inverse $\boldsymbol{A}\boldsymbol{A}^-\boldsymbol{A} = \boldsymbol{A}$, $\mathcal{C}(\boldsymbol{A}) = \mathcal{C}(\boldsymbol{A}\boldsymbol{A}^-)$, consistency of a linear system $\boldsymbol{A}\boldsymbol{A}^-\boldsymbol{b} = \boldsymbol{b}$, solution set of a consistent linear system $\boldsymbol{A}^-\boldsymbol{b}+(\boldsymbol{I}_n-\boldsymbol{A}^-\boldsymbol{A})\boldsymbol{q}$, $\mathcal{N}(\boldsymbol{A}) = \mathcal{C}(\boldsymbol{I}_n-\boldsymbol{A}^-\boldsymbol{A})$, $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ is consistent for all $\boldsymbol{b}$ if and only if $\boldsymbol{A}$ has full row rank, A consistent system $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ has a unique solution if and only if $\boldsymbol{A}$ has full column rank, ...

Today

- Idempotent matrix, projection and orthogonal projection
- Method of least squares
- Announcement
  - Change of TA office hours
  - Pre-lecture notes
  - Questions on homework 1?
  - Make use of comments on course webpages

Idempotent matrix

Assume $\boldsymbol{A} \in \mathbb{R}^{n\times n}$.

- A matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is *idempotent* if and only if $\boldsymbol{A}^2 = \boldsymbol{A}$.
- Any idempotent matrix $\boldsymbol{A}$ is a generalized inverse of itself.
- The only idempotent matrix of full rank is $\boldsymbol{I}$.

  Proof. Since $\boldsymbol{A}$ has full rank, the inverse $\boldsymbol{A}^{-1}$ exists. Then $\boldsymbol{A} = \boldsymbol{A}^{-1}\boldsymbol{A}\boldsymbol{A} = \boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}$. □

  Interpretation: all idempotent matrices are singular except the identity matrix.

- $\boldsymbol{A}$ is idempotent if and only if $\boldsymbol{A}^T$ is idempotent if and only if $\boldsymbol{I}_n - \boldsymbol{A}$ is idempotent.
- If $\boldsymbol{A}^2 = k\boldsymbol{A}$ for some nonzero scalar k, then $\text{tr}(\boldsymbol{A}) = k\,\text{rank}(\boldsymbol{A})$.

  Proof. Suppose $\text{rank}(\boldsymbol{A}) = r$. Choose a set of r linearly independent columns and form matrix $\boldsymbol{B} \in \mathbb{R}^{n\times r}$, which has full column rank. There exists matrix $\boldsymbol{C} \in \mathbb{R}^{r\times n}$ such that $\boldsymbol{A} = \boldsymbol{BC}$. $\boldsymbol{C}$ must have full row rank, otherwise $\boldsymbol{A}$ cannot have rank r. Then

  $$\boldsymbol{B}(\boldsymbol{CB})\boldsymbol{C} = \boldsymbol{A}^2 = k\boldsymbol{A} = k\boldsymbol{BC} = \boldsymbol{B}(k\boldsymbol{I}_r)\boldsymbol{C}.$$

  Since $\boldsymbol{B}$ has full column rank, we must have $(\boldsymbol{CB})\boldsymbol{C} = (k\boldsymbol{I}_r)\boldsymbol{C}$ (why?). Since $\boldsymbol{C}$ has full row rank, we must have $\boldsymbol{CB} = k\boldsymbol{I}_r$. Then $\text{tr}(\boldsymbol{A}) = \text{tr}(\boldsymbol{BC}) = \text{tr}(\boldsymbol{CB}) = \text{tr}(k\boldsymbol{I}_r) = kr = k\,\text{rank}(\boldsymbol{A})$. □
- For any idempotent matrix $\boldsymbol{A}$, $\text{tr}(\boldsymbol{A}) = \text{rank}(\boldsymbol{A})$.
- For any idempotent matrix $\boldsymbol{A}$,

  $$\text{rank}(\boldsymbol{I}_n - \boldsymbol{A}) = \text{tr}(\boldsymbol{I}_n - \boldsymbol{A}) = n - \text{rank}(\boldsymbol{A}).$$

  So $\text{rank}(\boldsymbol{A}) + \text{rank}(\boldsymbol{I} - \boldsymbol{A}) = n$ for any idempotent matrix $\boldsymbol{A}$.
- For a *general* matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, the matrices $\boldsymbol{A}^-\boldsymbol{A}$ and $\boldsymbol{AA}^-$ are idempotent and

  $$\begin{aligned} \text{rank}(\boldsymbol{A}) = \text{rank}(\boldsymbol{A}^-\boldsymbol{A}) &= \text{rank}(\boldsymbol{AA}^-) = \text{tr}(\boldsymbol{A}^-\boldsymbol{A}) = \text{tr}(\boldsymbol{AA}^-) \\ \text{rank}(\boldsymbol{I}_n - \boldsymbol{A}^-\boldsymbol{A}) &= \text{tr}(\boldsymbol{I}_n - \boldsymbol{A}^-\boldsymbol{A}) = n - \text{rank}(\boldsymbol{A}) \\ \text{rank}(\boldsymbol{I}_m - \boldsymbol{AA}^-) &= \text{tr}(\boldsymbol{I}_m - \boldsymbol{AA}^-) = m - \text{rank}(\boldsymbol{A}). \end{aligned}$$

- $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ and $\boldsymbol{B} \in \mathbb{R}^{n\times m}$. Then the following statements are equivalent.
  1. $\boldsymbol{B}$ is a generalized inverse of $\boldsymbol{A}$.
  2. $\boldsymbol{AB}$ is idempotent and $\text{rank}(\boldsymbol{AB}) = \text{rank}(\boldsymbol{A})$.
  3. $\boldsymbol{BA}$ is idempotent and $\text{rank}(\boldsymbol{BA}) = \text{rank}(\boldsymbol{A})$.

  Proof. Previous result shows 1 implies 2. We need to show 2 implies 1. If $\text{rank}(\boldsymbol{AB}) = \text{rank}(\boldsymbol{A})$, then $\mathcal{C}(\boldsymbol{A}) = \mathcal{C}(\boldsymbol{AB})$ and $\boldsymbol{A} = \boldsymbol{ABT}$ for some matrix $\boldsymbol{T}$. Thus by idempotency of $\boldsymbol{AB}$, $\boldsymbol{ABA} = \boldsymbol{ABABT} = \boldsymbol{ABT} = \boldsymbol{A}$. Equivalence between 1 and 3 is shown in a similar way. □

Projection

- A matrix $\boldsymbol{P} \in \mathbb{R}^{n\times n}$ is a *projection* onto a vector space $\mathcal{V}$ if and only if (a) $\boldsymbol{P}$ is idempotent, (b) $\boldsymbol{Px} \in \mathcal{V}$ for any $\boldsymbol{x} \in \mathbb{R}^n$, and (c) $\boldsymbol{Pz} = \boldsymbol{z}$ for any $\boldsymbol{z} \in \mathcal{V}$.

- Any idempotent matrix $\boldsymbol{P}$ is a projection onto its own column space $\mathcal{C}(\boldsymbol{P})$.

  Proof. Property (a) is free. Property (b) is trivial since $\boldsymbol{Px} \in \mathcal{C}(\boldsymbol{P})$ for any $\boldsymbol{x}$. For property (c), note $\boldsymbol{PP} = \boldsymbol{P}$ says $\boldsymbol{Pp}_i = \boldsymbol{p}_i$ for each column $\boldsymbol{p}_i$ of $\boldsymbol{P}$. Thus $\boldsymbol{Pz} = \boldsymbol{z}$ for any $\boldsymbol{z} \in \mathcal{C}(\boldsymbol{P})$. □

- $\boldsymbol{AA}^-$ is *a* projection onto the column space $\mathcal{C}(\boldsymbol{A})$.
  This gives a recipe for finding projections onto the column space of a matrix. Recall in last class we showed $\mathcal{C}(\boldsymbol{AA}^-) = \mathcal{C}(\boldsymbol{A})$. Therefore the proof is trivial.

  Direct proof. (a) Idempotent: $\boldsymbol{AA}^-\boldsymbol{AA}^- = \boldsymbol{AA}^-$ by definition of generalized inverse. (b) $\boldsymbol{AA}^-\boldsymbol{v} = \boldsymbol{A}(\boldsymbol{A}^-\boldsymbol{v}) \in \mathcal{C}(\boldsymbol{A})$ for any $\boldsymbol{v}$. (c) Let $\boldsymbol{z} \in \mathcal{C}(\boldsymbol{A})$, then $\boldsymbol{z} = \boldsymbol{Ac}$ for some $\boldsymbol{c}$. Thus $\boldsymbol{AA}^-\boldsymbol{z} = \boldsymbol{AA}^-\boldsymbol{Ac} = \boldsymbol{Ac} = \boldsymbol{z}$. □

- $\boldsymbol{I}_n - \boldsymbol{A}^-\boldsymbol{A}$ is *a* projection onto the null space $\mathcal{N}(\boldsymbol{A})$.
  This gives a recipe for finding projection onto the null space of a matrix. Recall in last class we showed $\mathcal{C}(\boldsymbol{I}_n - \boldsymbol{A}^-\boldsymbol{A}) = \mathcal{N}(\boldsymbol{A})$. Therefore the proof is trivial.

  Direct proof. (a) $\boldsymbol{I}_n - \boldsymbol{A}^-\boldsymbol{A}$ is idempotent because $\boldsymbol{A}^-\boldsymbol{A}$ is idempotent. (b) For any $\boldsymbol{x}$, $\boldsymbol{A}(\boldsymbol{I}_n - \boldsymbol{A}^-\boldsymbol{A})\boldsymbol{x} = (\boldsymbol{A} - \boldsymbol{A})\boldsymbol{x} = \boldsymbol{0}$. That is $\boldsymbol{x} \in \mathcal{N}(\boldsymbol{A})$. (c) For any $\boldsymbol{z} \in \mathcal{N}(\boldsymbol{A})$, $(\boldsymbol{I}_n - \boldsymbol{A}^-\boldsymbol{A})\boldsymbol{z} = \boldsymbol{z} - \boldsymbol{A}^-\boldsymbol{Az} = \boldsymbol{z} - \boldsymbol{0} = \boldsymbol{z}$. □

- In general, the projections onto a vector space are not unique.

- JM Example A.7. Let $\boldsymbol{A} = \begin{pmatrix} 1 \\ 2 \end{pmatrix} \in \mathbb{R}^{2\times 1}$. We first find generalized inverse $\boldsymbol{G} = (u, v) \in \mathbb{R}^{1\times 2}$ of $\boldsymbol{A}$. Definition of generalized inverse requires $\boldsymbol{AGA} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}(u, v)\begin{pmatrix} 1 \\ 2 \end{pmatrix} = \boldsymbol{A} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$, i.e., $v = (1-u)/2$. Thus $\boldsymbol{G}_u = (u, (1-u)/2)$ is a generalized inverse of $\boldsymbol{A}$ for any value of u. Thus

$$\boldsymbol{AG}_u = \begin{pmatrix} 1 \\ 2 \end{pmatrix}(u, (1-u)/2) = \begin{pmatrix} u & (1-u)/2 \\ 2u & 1-u \end{pmatrix}$$

is a projection onto $\mathcal{C}(\boldsymbol{A})$ for any u. Taking $u = 0$ gives a projection

$$\boldsymbol{A}\boldsymbol{G}_0 = \begin{pmatrix} 0 & 1/2 \\ 0 & 1 \end{pmatrix}.$$

$\boldsymbol{A}\boldsymbol{G}_0\boldsymbol{x} = \begin{pmatrix} x_2/2 \\ x_2 \end{pmatrix}$ for any $\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$. That is $\boldsymbol{A}\boldsymbol{G}_0$ projects in $\mathbb{R}^2$ points vertically to the line $\mathcal{C}(\boldsymbol{A})$. Taking $u = 1$ gives

$$\boldsymbol{A}\boldsymbol{G}_1 = \begin{pmatrix} 1 & 0 \\ 2 & 0 \end{pmatrix},$$

which projects points in $\mathbb{R}^2$ horizontally to the line. What if we require the projection $\boldsymbol{A}\boldsymbol{G}_u$ to be symmetric? Then $2u = (1-u)/2$ suggests $u = 1/5$ and

$$\boldsymbol{A}\boldsymbol{G}_{1/5} = \begin{pmatrix} 1/5 & 2/5 \\ 2/5 & 4/5 \end{pmatrix},$$

which projects points in $\mathbb{R}^2$ to the closest point on the line. This is an instance of *orthogonal projection*.

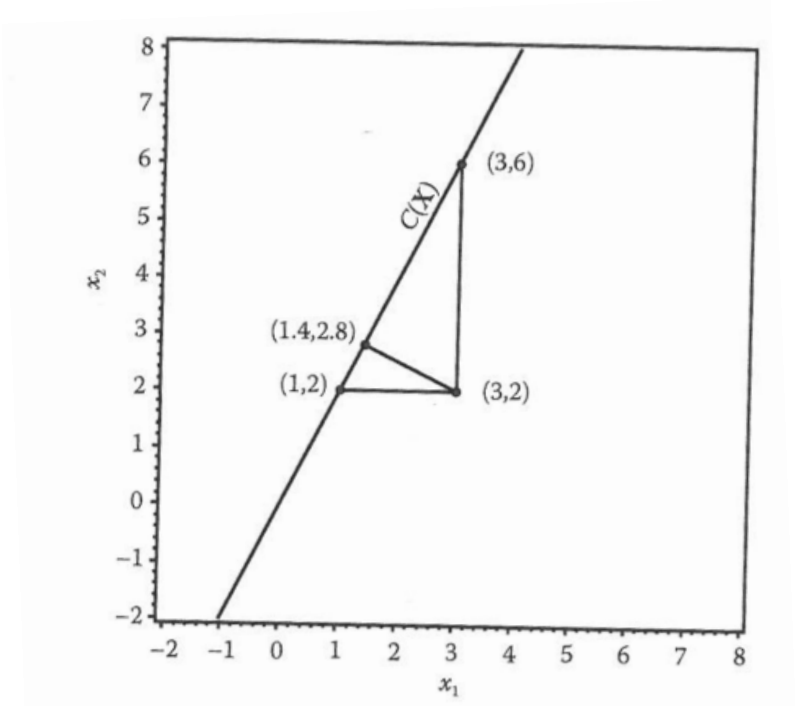


Figure A.1: Three projections onto a column space.

Orthogonal projection

- A symmetric, idempotent matrix $\boldsymbol{P}$ that projects onto a vector space $\mathcal{V}$ is unique.

  Proof. Let $\boldsymbol{P}$ and $\boldsymbol{Q}$ be any two symmetric, idempotent matrices that project to the same vector space. $\boldsymbol{Q}$ projects onto $\mathcal{C}(\boldsymbol{Q})$, thus $\boldsymbol{P}$ projects onto $\mathcal{C}(\boldsymbol{Q})$ too. Therefore $\boldsymbol{P}\boldsymbol{q}_i = \boldsymbol{q}_i$ for each column $\boldsymbol{q}_i$ of $\boldsymbol{Q}$. That is $\boldsymbol{P}\boldsymbol{Q} = \boldsymbol{Q}$. Similarly $\boldsymbol{Q}\boldsymbol{P} = \boldsymbol{P}$. Thus

  $$\begin{aligned} & \langle \boldsymbol{P}-\boldsymbol{Q}, \boldsymbol{P}-\boldsymbol{Q} \rangle = \operatorname{tr}((\boldsymbol{P}-\boldsymbol{Q})^T(\boldsymbol{P}-\boldsymbol{Q})) \\ = \;& \operatorname{tr}(\boldsymbol{P}^T\boldsymbol{P} - \boldsymbol{P}^T\boldsymbol{Q} - \boldsymbol{Q}^T\boldsymbol{P} + \boldsymbol{Q}^T\boldsymbol{Q}) \\ = \;& \operatorname{tr}(\boldsymbol{P}) - \operatorname{tr}(\boldsymbol{P}\boldsymbol{Q}) - \operatorname{tr}(\boldsymbol{Q}\boldsymbol{P}) + \operatorname{tr}(\boldsymbol{Q}) \\ = \;& \operatorname{tr}(\boldsymbol{P}) - \operatorname{tr}(\boldsymbol{Q}) - \operatorname{tr}(\boldsymbol{P}) + \operatorname{tr}(\boldsymbol{Q}) \\ = \;& 0. \end{aligned}$$

  $\boldsymbol{P}$ and $\boldsymbol{Q}$ must be equal. □

- Any symmetric, idempotent matrix is called an *orthogonal projection*. The term orthogonal comes from the fact that for any vector $\boldsymbol{y}$, the residual $\boldsymbol{y} - \boldsymbol{P}\boldsymbol{y}$ is orthogonal to the vector space $\boldsymbol{P}$ is projecting onto. That is, for any $\boldsymbol{v} \in \mathcal{V}$, $\langle \boldsymbol{y} - \boldsymbol{P}\boldsymbol{y}, \boldsymbol{v} \rangle = \boldsymbol{y}^T(\boldsymbol{I}_n - \boldsymbol{P}^T)\boldsymbol{v} = \boldsymbol{y}^T(\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{v} = \boldsymbol{y}^T(\boldsymbol{v} - \boldsymbol{v}) = 0$.

- Many books use the term "projection" in the sense of orthogonal projection.

- If a symmetric, idempotent matrix $\boldsymbol{P}$ projects onto $\mathcal{V}$, then $\boldsymbol{I} - \boldsymbol{P}$ projects onto the orthogonal complement $\mathcal{V}^\perp$.

  Proof. Since $\boldsymbol{P}$ projects onto $\mathcal{C}(\boldsymbol{P})$, we need to show $\boldsymbol{I} - \boldsymbol{P}$ projects onto $\mathcal{C}(\boldsymbol{P})^\perp = \mathcal{N}(\boldsymbol{P}^T) = \mathcal{N}(\boldsymbol{P})$. But $\boldsymbol{P}$ is a generalized inverse of itself and thus $\boldsymbol{I} - \boldsymbol{P} = \boldsymbol{I} - \boldsymbol{P}\boldsymbol{P}$ projects to the $\mathcal{C}(\boldsymbol{I} - \boldsymbol{P}\boldsymbol{P}) = \mathcal{N}(\boldsymbol{P})$. Symmetry is trivial. □

Method of least squares

Given $\boldsymbol{y} \in \mathbb{R}^n$ and $\boldsymbol{X} \in \mathbb{R}^{n \times p}$, we want to find a vector $\boldsymbol{b} \in \mathbb{R}^p$ such that $\boldsymbol{X}\boldsymbol{b}$ approximates $\boldsymbol{y}$ well. In other words, we want to approximate $\boldsymbol{y} \in \mathbb{R}^n$ by a linear combination of vectors $\boldsymbol{x}_j \in \mathbb{R}^n$, $j = 1, \ldots, p$. Note the method of least squares only concerns approximation and has nothing to do with randomness and estimation.

- Suppose the linear system $\boldsymbol{X}\boldsymbol{b} = \boldsymbol{y}$ is consistent, we just need to find a solution to the linear system, which takes the general form $\boldsymbol{X}^{-}\boldsymbol{y}+(\boldsymbol{I}_p-\boldsymbol{X}^{-}\boldsymbol{X})\boldsymbol{q}$, $\boldsymbol{q} \in \mathbb{R}^p$. Recall that $\boldsymbol{I}_p - \boldsymbol{X}^{-}\boldsymbol{X}$ is a projection onto the null space $\mathcal{N}(\boldsymbol{X})$.

- Suppose the linear system is inconsistent. The *method of least squares* (due to Gauss) seeks $\boldsymbol{b}$ that minimizes the Euclidean norm of the residual vector

$$\boldsymbol{Q}(\boldsymbol{b}) = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\|_2^2 = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}) = \boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{b}^T\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{b}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{b}. \quad (1)$$

- Normal equation. To find the minimum, we take derivative and set the gradient to $\mathbf{0}$

$$\nabla Q(\boldsymbol{b}) = -2\boldsymbol{X}^T\boldsymbol{y} + 2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{b} = \mathbf{0}_p.$$

  This leads to the *normal equation*

$$\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{b} = \boldsymbol{X}^T\boldsymbol{y}. \quad (2)$$

- Is there a solution to the normal equation?
  Normal equation is always consistent and thus admits at least one solution $\hat{\boldsymbol{b}}$. (HW1: show that $\mathcal{C}(\boldsymbol{X}^T) = \mathcal{C}(\boldsymbol{X}^T\boldsymbol{X})$.)

- Is the solution to the normal equation the minimizer to the least squares criterion?
  Any solution $\hat{\boldsymbol{b}}$ to the normal equation (2) minimizes the least squares criterion (1).
  Optimization argument: Any stationarity point (points with zero gradient vector) of a convex function is a global minimum. Now the least squares criterion is convex because the Hessian $\nabla^2 Q(\boldsymbol{b}) = \boldsymbol{X}^T\boldsymbol{X}$ is positive semidefinite. Therefore any solution to the normal equation is a stationarity point and thus a global minimum.
  Direct argument: Let $\hat{\boldsymbol{b}}$ be a solution to the normal equation. For arbitrary

$\boldsymbol{b} \in \mathbb{R}^p$,

$$\begin{aligned} Q(\boldsymbol{b}) - Q(\hat{\boldsymbol{b}}) &= -2(\boldsymbol{b}-\hat{\boldsymbol{b}})^T\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{b}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{b} - \hat{\boldsymbol{b}}^T\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{b}} \\ &= -2(\boldsymbol{b}-\hat{\boldsymbol{b}})^T\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{b}} + \boldsymbol{b}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{b} - \hat{\boldsymbol{b}}^T\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{b}} \\ &= \hat{\boldsymbol{b}}^T\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{b}} + \boldsymbol{b}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{b} - 2\boldsymbol{b}^T\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{b}} \\ &= (\boldsymbol{b}-\hat{\boldsymbol{b}})^T\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{b}-\hat{\boldsymbol{b}}) \\ &= \|\boldsymbol{X}(\boldsymbol{b}-\hat{\boldsymbol{b}})\|_2^2 \\ &\geq 0. \end{aligned}$$

- The direct argument also reveals that the *fitted values* $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{b}}$ is invariant to the choice of the solution to the normal equation.

- Now we know the normal equation is always consistent and we want to find solution(s). In general the solution can be represented as

$$\hat{\boldsymbol{b}}(\boldsymbol{q}) = (\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{y} + [\boldsymbol{I}_p - (\boldsymbol{X}^T\boldsymbol{X})^-(\boldsymbol{X}^T\boldsymbol{X})]\boldsymbol{q}, \tag{3}$$

  where $\boldsymbol{q} \in \mathbb{R}^p$ is arbitrary. One specific solution is

$$\hat{\boldsymbol{b}} = (\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{y}$$

  with corresponding fitted values

$$\hat{\boldsymbol{y}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{y}.$$

- When is the least squares solution unique?
  The least squares solution is unique if and only if $\boldsymbol{X}$ has full column rank. The solution is given by $\hat{\boldsymbol{b}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$.

  Proof. The solution to normal equation is unique if and only if $\boldsymbol{X}^T\boldsymbol{X}$ has full (column) rank. Therefore $\text{rank}(\boldsymbol{X}) = \text{rank}(\boldsymbol{X}^T\boldsymbol{X}) = p$. □

- $(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T$ is a generalized inverse of $\boldsymbol{X}$.

  Proof. By definition of generalized inverse, $\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{X}^T\boldsymbol{X}$. Re-arranging gives

$$\boldsymbol{X}^T\boldsymbol{X}[(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{X} - \boldsymbol{I}_p] = \boldsymbol{0}_{p\times p}.$$

Each column of the matrix $\boldsymbol{X}[(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{X} - \boldsymbol{I}_p]$ is in $\mathcal{N}(\boldsymbol{X}^T)$ and also in $\mathcal{C}(\boldsymbol{X})$. Since $\mathcal{C}(\boldsymbol{X}) \cap \mathcal{N}(\boldsymbol{X}^T) = \{\boldsymbol{0}\}$, that column must be $\boldsymbol{0}$. Therefore $\boldsymbol{X}[(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{X} - \boldsymbol{I}_p] = \boldsymbol{0}_{n\times p}$. That is $\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{X}$. □

- The least squares solution has same format (3) regardless the consistency of the system $\boldsymbol{X}\boldsymbol{b} = \boldsymbol{y}$, since we can replace $(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T$ is a version of $\boldsymbol{X}^-$ from the preceding result.

4 Lecture 4: Sep 4

Announcement

- HW1 is due today @ 11:59pm; TA is available answering questions in the afternoon session
- Read JM chapter 3
- Instructor out of town next week; no classes.

Last time

- Idempotent matrix
- Projection
  - any idempotent matrix $\boldsymbol{P}$ projects onto $\mathcal{C}(\boldsymbol{P})$
  - $\boldsymbol{A}\boldsymbol{A}^-$ projects onto $\mathcal{C}(\boldsymbol{A})$
  - $\boldsymbol{I} - \boldsymbol{A}^-\boldsymbol{A}$ projects onto $\mathcal{N}(\boldsymbol{A})$
- Orthogonal projection (symmetric idempotent matrix): uniqueness and orthogonality; we will see $\boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{A})^-\boldsymbol{A}^T$ is the orthogonal projection onto $\mathcal{C}(\boldsymbol{A})$
- Method of least squares:
  - least squares criterion: $Q(\boldsymbol{b}) = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\|_2^2$
  - normal equation: $\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{b} = \boldsymbol{X}^T\boldsymbol{y}$
  - normal equation is always consistent
  - any solution to the normal equation is a least squares solution (that minimizes the least squares criterion)
  - least squares solution takes the general form

    $$\hat{\boldsymbol{b}} = (\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{y} + (\boldsymbol{I}_p - (\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{X})\boldsymbol{q},$$

    where $\boldsymbol{q} \in \mathbb{R}^p$ is arbitrary
  - the least squares solution is unique if and only if $\boldsymbol{X}$ has full column rank
  - $(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T$ is a generalized inverse of $\boldsymbol{X}$; the least squares solution is the same whether the system $\boldsymbol{X}\boldsymbol{b} = \boldsymbol{y}$ is consistent or not

Today

- Geometry of least squares solution
- Gram-Schmidt orthogonalization
- Reparameterizations
- linear mean model
- random vectors
- estimable functions

Geometry of the least squares solution

- $\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T$ is *the* orthogonal projection onto $\mathcal{C}(\boldsymbol{X})$.

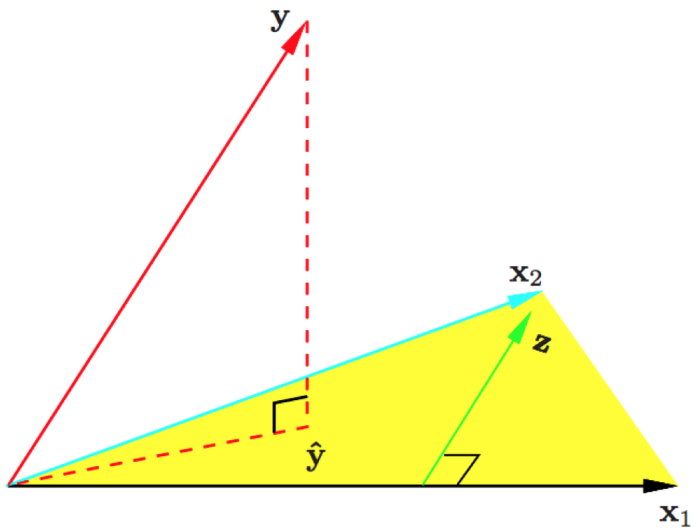
  Proof. We showed that $(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T$ is a generalized inverse of $\boldsymbol{X}$ in HW1. Therefore $\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T$ is a projection onto $\mathcal{C}(\boldsymbol{X})$. We only need to show the symmetry, which follows from the fact that transpose of $(\boldsymbol{X}^T\boldsymbol{X})^-$ is a generalized inverse of $(\boldsymbol{X}^T\boldsymbol{X})^T = \boldsymbol{X}^T\boldsymbol{X}$. □

- Since orthogonal projection is unique, $\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T$ is invariant to the choice of the generalized inverse $(\boldsymbol{X}^T\boldsymbol{X})^-$ and thus can be denoted by $\boldsymbol{P}_{\boldsymbol{X}}$.
- Whichever least squares solution $\hat{\boldsymbol{b}}$ we use, the fitted values $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{b}}$ is the same since

$$\begin{aligned}\hat{\boldsymbol{y}} &= \boldsymbol{X}\hat{\boldsymbol{b}} \\ &= \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{X}(\boldsymbol{I}_p - (\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{X})\boldsymbol{q} \\ &= \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{y} \\ &= \boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{y}\end{aligned}$$

  and the orthogonal projection is unique.

- Geometry: The fitted value from the least squares solution $\hat{\boldsymbol{y}} = \boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{y}$ is the orthogonal projection of the response vector $\boldsymbol{y}$ onto the column space $\mathcal{C}(\boldsymbol{X})$.



- $\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}}$ is the orthogonal projection onto $\mathcal{N}(\boldsymbol{X}^T)$.
- Decomposition of $\boldsymbol{y}$:

$$\boldsymbol{y} = \boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{y} + (\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{y} = \hat{\boldsymbol{y}} + \hat{\boldsymbol{e}},$$

  where $\hat{\boldsymbol{y}} \perp \hat{\boldsymbol{e}}$ and

$$\|\boldsymbol{y}\|_2^2 = \|\hat{\boldsymbol{y}}\|_2^2 + \|\hat{\boldsymbol{e}}\|_2^2.$$

  The Pythagorean theorem follows from the orthogonality between $\mathcal{C}(\boldsymbol{X})$ and $\mathcal{N}(\boldsymbol{X}^T)$ since $\|\boldsymbol{y}\|_2^2 = \boldsymbol{y}^T\boldsymbol{y} = (\hat{\boldsymbol{y}} + \hat{\boldsymbol{e}})^T(\hat{\boldsymbol{y}} + \hat{\boldsymbol{e}}) = \hat{\boldsymbol{y}}^T\hat{\boldsymbol{y}} + \hat{\boldsymbol{y}}^T\hat{\boldsymbol{e}} + \hat{\boldsymbol{e}}^T\hat{\boldsymbol{y}} + \hat{\boldsymbol{e}}^T\hat{\boldsymbol{e}} = \|\hat{\boldsymbol{y}}\|_2^2 + \|\hat{\boldsymbol{e}}\|_2^2$.

- Example: simple linear regression (predict y_i from intercept and one predictor: $y_i \approx b_0 + x_i b_1$).

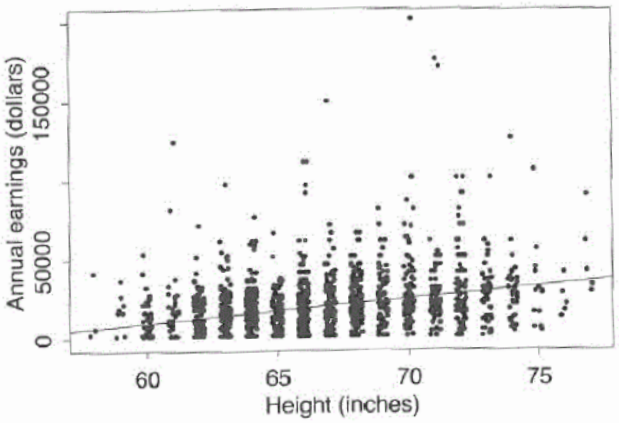


Fig. 4.3 Earnings vs. height for a random sample of adult Americans in 1990. The heights have been jittered slightly so that the points do not overlap.

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}.$$

Assume $(x_1, \ldots, x_n)$ is not a constant vector. The Gramian matrix is

$$\boldsymbol{X}^T\boldsymbol{X} = \begin{pmatrix} n & \sum_i x_i \\ \sum_i x_i & \sum_i x_i^2 \end{pmatrix}$$

and its inverse is

$$(\boldsymbol{X}^T\boldsymbol{X})^{-1} = \frac{1}{n\sum_i (x_i - \bar{x})^2} \begin{pmatrix} \sum_i x_i^2 & -\sum_i x_i \\ -\sum_i x_i & n \end{pmatrix}.$$

The (unique) least squares solution is

$$\begin{aligned} \hat{\boldsymbol{b}} &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \\ &= \frac{1}{n\sum_i (x_i - \bar{x})^2} \begin{pmatrix} \sum_i x_i^2 & -\sum_i x_i \\ -\sum_i x_i & n \end{pmatrix} \begin{pmatrix} \mathbf{1}_n^T \\ \boldsymbol{x}^T \end{pmatrix} \boldsymbol{y} \\ &= \frac{1}{n\sum_i (x_i - \bar{x})^2} \begin{pmatrix} \sum_i x_i^2 & -\sum_i x_i \\ -\sum_i x_i & n \end{pmatrix} \begin{pmatrix} \sum_i y_i \\ \sum_i x_i y_i \end{pmatrix} \\ &= \frac{1}{n\sum_i (x_i - \bar{x})^2} \begin{pmatrix} (\sum_i x_i^2)(\sum_i y_i) - (\sum_i x_i)(\sum_i x_i y_i) \\ n\sum_i x_i y_i - (\sum_i x_i)(\sum_i y_i) \end{pmatrix} \\ &= \begin{pmatrix} \bar{y} - \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}\bar{x} \\ \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} \end{pmatrix}. \end{aligned}$$

The fitted values are

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{b}} = (\bar{y} - \hat{b}_1\bar{x})\mathbf{1}_n + \hat{b}_1\boldsymbol{x} = \bar{y}\mathbf{1}_n - \hat{b}_1(\boldsymbol{x} - \bar{x}\mathbf{1}_n).$$

That is $\hat{y}_i = \bar{y} + \hat{b}_1(x_i - \bar{x})$. The residuals are $\hat{e}_i = (y_i - \bar{y}) - \hat{b}_1(x_i - \bar{x})$. And

$$\begin{aligned}
\langle \hat{\boldsymbol{y}}, \hat{\boldsymbol{e}} \rangle &= \sum_i \hat{y}_i \hat{e}_i \\
&= \sum_i \Big[\bar{y} + \hat{b}_1(x_i - \bar{x})\Big][(y_i - \bar{y}) - \hat{b}_1(x_i - \bar{x})\Big] \\
&= \hat{b}_1 \left[\sum_i (x_i - \bar{x})(y_i - \bar{y}) - \hat{b}_1 \sum_i (x_i - \bar{x})^2\right] \\
&= 0.
\end{aligned}$$

Gram-Schmidt and QR decomposition

- Let $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^n$ be two non-zero, linearly independent vectors. Then $\boldsymbol{P}_{\mathcal{C}\{\boldsymbol{x}_1\}} = \boldsymbol{x}_1(\boldsymbol{x}_1^T\boldsymbol{x}_1)^{-1}\boldsymbol{x}_1^T = \boldsymbol{x}_1\boldsymbol{x}_1^T/\|\boldsymbol{x}_1\|_2^2$ and

$$\boldsymbol{x}_2 - \boldsymbol{P}_{\boldsymbol{x}_1}\boldsymbol{x}_2 = \boldsymbol{x}_2 - \frac{\langle \boldsymbol{x}_2, \boldsymbol{x}_1 \rangle}{\|\boldsymbol{x}_1\|_2^2}\boldsymbol{x}_1$$

is orthogonal to $\boldsymbol{x}_1$. This is the Gram-Schmidt orthogonalization of two vectors.

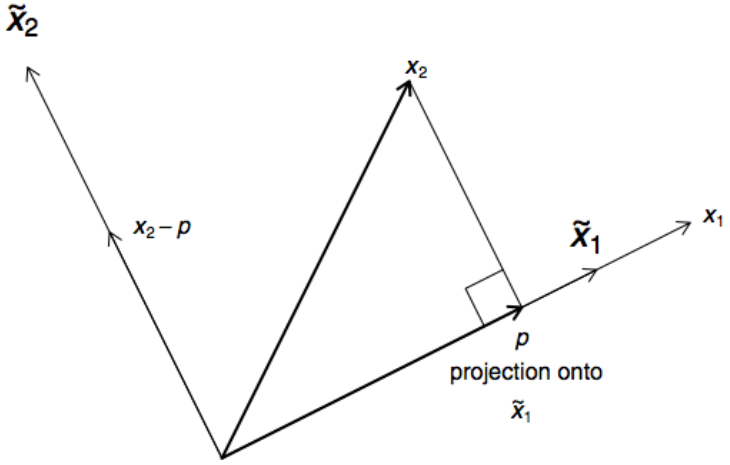


Fig. 2.2. Orthogonalization of x_1 and x_2

- Gram-Schmidt algorithm orthonormalizes a set of non-zero, *linearly independent* vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$. Initialize $\boldsymbol{q}_1 = \boldsymbol{x}_1/\|\boldsymbol{x}_1\|_2$; then for $j = 2, \ldots, p$,

$$\begin{aligned} \boldsymbol{v}_j &= \boldsymbol{x}_j - \boldsymbol{P}_{\mathcal{C}\{\boldsymbol{q}_1,\ldots,\boldsymbol{q}_{j-1}\}}\boldsymbol{x}_j = \boldsymbol{x}_j - \sum_{k=1}^{j-1}\langle \boldsymbol{x}_j, \boldsymbol{q}_k\rangle \boldsymbol{q}_k \\ \boldsymbol{q}_j &= \boldsymbol{v}_j/\|\boldsymbol{v}_j\|_2 \end{aligned}$$

- For $j = 1, \ldots, q$, $\mathcal{C}\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_j\} = \mathcal{C}\{\boldsymbol{q}_1, \ldots, \boldsymbol{q}_j\}$ and $\boldsymbol{q}_j \perp \mathcal{C}\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{j-1}\}$.
- Collectively, we have $\boldsymbol{X} = \boldsymbol{Q}\boldsymbol{R}$, where
  - $\boldsymbol{Q} \in \mathbb{R}^{n\times p}$ has orthonormal columns $\boldsymbol{q}_j$ and thus $\boldsymbol{Q}^T\boldsymbol{Q} = \boldsymbol{I}_p$.
  - $\boldsymbol{R} = \boldsymbol{Q}^T\boldsymbol{X} \in \mathbb{R}^{p\times p}$ has entries $r_{kj} = \langle \boldsymbol{q}_k, \boldsymbol{x}_j\rangle$, which are available from the algorithm. Note $r_{kj} = 0$ for $k > j$. Thus $\boldsymbol{R}$ is upper triangular.

  This is called the QR decomposition of $\boldsymbol{X}$.
- The original normal equation $\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{b} = \boldsymbol{X}^T\boldsymbol{y}$ becomes $\boldsymbol{R}^T\boldsymbol{R}\boldsymbol{b} = \boldsymbol{R}^T\boldsymbol{Q}\boldsymbol{y}$, which is easy to solve (why?). Therefore QR decomposition by Gram-Schmidt offers a practical algorithm for solving the least squares problem.
- $\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{R}^T\boldsymbol{Q}^T\boldsymbol{Q}\boldsymbol{R} = \boldsymbol{R}^T\boldsymbol{R}$ is the *Cholesky decomposition* of $\boldsymbol{X}^T\boldsymbol{X}$.

Reparameterization

- Example: We want to predict weight y_i by height x_i (and intercept). Intuitively it should not matter whether we record weight in pounds or kilograms.

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \quad \boldsymbol{W} = \begin{pmatrix} 1 & w_1 \\ 1 & w_2 \\ \vdots & \vdots \\ 1 & w_n \end{pmatrix},$$

  where $w_i = 0.4536x_i$ (1 pound = 0.4536 kilogram). Note

$$\boldsymbol{W} = \boldsymbol{X}\begin{pmatrix} 1 & 0 \\ 0 & 0.4536 \end{pmatrix}, \quad \boldsymbol{X} = \boldsymbol{W}\begin{pmatrix} 1 & 0 \\ 0 & 2.2046 \end{pmatrix}.$$

- Example: We want to predict salary y_i by gender x_i (and intercept). Intuitively it should not matter whether we record gender as $x_i = 1_{\{\text{male}\}}$ or $w_i = 1_{\{\text{female}\}} = 1 - x_i$ or even keeping both x_i and w_i as predictors.

$$\boldsymbol{X}_1 = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \quad \boldsymbol{X}_2 = \begin{pmatrix} 1 & 1-x_1 \\ 1 & 1-x_2 \\ \vdots & \vdots \\ 1 & 1-x_n \end{pmatrix}, \quad \boldsymbol{X}_3 = \begin{pmatrix} 1 & x_1 & 1-x_1 \\ 1 & x_2 & 1-x_2 \\ \vdots & \vdots & \\ 1 & x_n & 1-x_n \end{pmatrix}.$$

  Note

$$\boldsymbol{X}_2 = \boldsymbol{X}_1 \begin{pmatrix} 1 & 1 \\ 0 & -1 \end{pmatrix}, \quad \boldsymbol{X}_3 = \boldsymbol{X}_1 \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & -1 \end{pmatrix}, ...$$

- Two linear models $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{e}$, where $\boldsymbol{X} \in \mathbb{R}^{n \times p}$, and $\boldsymbol{y} = \boldsymbol{W}\boldsymbol{c} + \boldsymbol{e}$, where $\boldsymbol{W} \in \mathbb{R}^{n \times q}$, are equivalent or *reparameterizations* of each other if $\mathcal{C}(\boldsymbol{X}) = \mathcal{C}(\boldsymbol{W})$.

- If $\mathcal{C}(\boldsymbol{X}) = \mathcal{C}(\boldsymbol{W})$, then $\boldsymbol{P}_{\boldsymbol{X}} = \boldsymbol{P}_{\boldsymbol{W}}$.

  Proof. $\boldsymbol{P}_{\boldsymbol{X}}$ is an orthogonal projection onto $\mathcal{C}(\boldsymbol{X})$. $\boldsymbol{P}_{\boldsymbol{W}}$ is an orthogonal projection onto $\mathcal{C}(\boldsymbol{W})$. But we know orthogonal projection onto a vector space is unique. □

- The fitted values $\hat{\boldsymbol{y}}$ and residuals $\hat{\boldsymbol{e}}$ of two equivalent linear models (parameterizations) are same. Hence $\boldsymbol{P}_{\boldsymbol{X}} = \boldsymbol{P}_{\boldsymbol{W}}$.

  Proof. $\hat{\boldsymbol{y}} = \boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{y}$ and $\hat{\boldsymbol{e}} = (\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{y}$. Since $\boldsymbol{P}_{\boldsymbol{X}} = \boldsymbol{P}_{\boldsymbol{W}}$, we see the fitted values and residuals must be same. □

- Translation between solutions of equivalent models.
  If $\mathcal{C}(\boldsymbol{X}) = \mathcal{C}(\boldsymbol{W})$, then there exist $\boldsymbol{T} \in \mathbb{R}^{p \times q}$ and $\boldsymbol{S} \in \mathbb{R}^{q \times p}$ such that $\boldsymbol{W} = \boldsymbol{X}\boldsymbol{T}$ and $\boldsymbol{X} = \boldsymbol{W}\boldsymbol{S}$.

  - If $\hat{\boldsymbol{c}}$ is a solution to the normal equation $\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{c} = \boldsymbol{W}^T\boldsymbol{y}$, then $\hat{\boldsymbol{b}} = \boldsymbol{T}\hat{\boldsymbol{c}}$ is a solution to the normal equation $\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{b} = \boldsymbol{X}^T\boldsymbol{y}$.

    Proof. $\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{b}} = \boldsymbol{X}^T\boldsymbol{X}\boldsymbol{T}\hat{\boldsymbol{c}} = \boldsymbol{X}^T\boldsymbol{W}\hat{\boldsymbol{c}} = \boldsymbol{X}^T\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{y} = \boldsymbol{X}^T\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{y} = (\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{X})^T\boldsymbol{y} = \boldsymbol{X}^T\boldsymbol{y}$ □

  - Similarly, if $\hat{\boldsymbol{b}}$ is a solution to the normal equation $\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{b} = \boldsymbol{X}^T\boldsymbol{y}$, then $\hat{\boldsymbol{c}} = \boldsymbol{S}\hat{\boldsymbol{b}}$ is a solution to the normal equation $\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{c} = \boldsymbol{W}^T\boldsymbol{y}$.

Linear mean model

So far we have considered the method of least squares, which concerns the approximation of a vector $\boldsymbol{y} \in \mathbb{R}^n$ by p vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p \in \mathbb{R}^n$. Next we consider the linear mean model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{e},$$

where the random errors $\boldsymbol{e}$ are assumed to have mean $\mathrm{E}(\boldsymbol{e}) = \boldsymbol{0}$.

Random vectors

A brief review of random vectors.

- Let $\boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \in \mathbb{R}^n$ be a vector of n random variables with mean $\mathrm{E}(y_i) = \mu_i$, variance $\mathrm{Var}(y_i) = \mathrm{E}(y_i - \mathrm{E}y_i)^2 = \sigma_i^2$, and covariance $\mathrm{Cov}(y_i, y_j) = \mathrm{E}[(y_i - \mathrm{E}y_i)(y_j - \mathrm{E}y_j)] = \sigma_{ij} = \sigma_{ji}$. Collectively, we write

$$\mathrm{E}(\boldsymbol{y}) = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix} = \boldsymbol{\mu}, \quad \mathrm{Cov}(\boldsymbol{y}) = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{12} & \sigma_{22} & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nn} \end{pmatrix}.$$

- Let $\boldsymbol{y} \in \mathbb{R}^n$ be a random vector, then

$$\mathrm{Cov}(\boldsymbol{y}) = \mathrm{E}(\boldsymbol{y}\boldsymbol{y}^T) - (\mathrm{E}\boldsymbol{y})(\mathrm{E}\boldsymbol{y})^T.$$

  Special case: $\mathrm{Var}(y) = \mathrm{E}y^2 - (\mathrm{E}y)^2$.

- If $\boldsymbol{y} \in \mathbb{R}^n$ is a random vector, $\boldsymbol{A} \in \mathbb{R}^{r \times n}$ is a constant matrix, and $\boldsymbol{b} \in \mathbb{R}^r$ is a constant vector, then

$$\begin{aligned} \mathrm{E}(\boldsymbol{A}\boldsymbol{y} + \boldsymbol{b}) &= \boldsymbol{A}\mathrm{E}(\boldsymbol{y}) + \boldsymbol{b} \\ \mathrm{Cov}(\boldsymbol{A}\boldsymbol{y} + \boldsymbol{b}) &= \boldsymbol{A}\mathrm{Cov}(\boldsymbol{y})\boldsymbol{A}^T. \end{aligned}$$

  Special case: $\mathrm{E}(ay + b) = a\mathrm{E}(y) + b$ and $\mathrm{Var}(ay + b) = a^2\mathrm{Var}(y)$ for random variable y.

- For example, under the linear mean model, $\mathrm{E}\boldsymbol{y} = \mathrm{E}(\boldsymbol{X}\boldsymbol{b} + \boldsymbol{e}) = \boldsymbol{X}\boldsymbol{b} + \mathrm{E}(\boldsymbol{e}) = \boldsymbol{X}\boldsymbol{b}$.

- The covariance between two random vectors $\boldsymbol{x} \in \mathbb{R}^m$ and $\boldsymbol{y} \in \mathbb{R}^n$ is defined as

$$\mathrm{Cov}(\boldsymbol{x}, \boldsymbol{y}) = \mathrm{E}[(\boldsymbol{x} - \mathrm{E}\boldsymbol{x})(\boldsymbol{y} - \mathrm{E}\boldsymbol{y})^T] \in \mathbf{R}^{m \times n}.$$

  Note $\mathrm{Cov}(\boldsymbol{y}, \boldsymbol{y}) = \mathrm{Cov}(\boldsymbol{y})$.

- Let $\boldsymbol{x} \in \mathbb{R}^m$ and $\boldsymbol{y} \in \mathbb{R}^n$ be two random vectors and $\boldsymbol{A} \in \mathbb{R}^{p \times m}$ and $\boldsymbol{B} \in \mathbb{R}^{q \times n}$ be two constant matrices. Then

$$\mathrm{Cov}(\boldsymbol{A}\boldsymbol{x}, \boldsymbol{B}\boldsymbol{y}) = \boldsymbol{A}\mathrm{Cov}(\boldsymbol{x}, \boldsymbol{y})\boldsymbol{B}^T.$$

  Special case: $\mathrm{Cov}(ax, by) = ab\mathrm{Cov}(x, y)$ for random variables x, y.

- Let $\boldsymbol{x} \in \mathbb{R}^m, \boldsymbol{y} \in \mathbb{R}^n$ be two random vectors and $\boldsymbol{A} \in \mathbb{R}^{p \times m}$, $\boldsymbol{B} \in \mathbb{R}^{p \times n}$ are two constant matrices. Then

$$\begin{aligned} & \mathrm{Cov}(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{y}) \\ = \; & \boldsymbol{A}\mathrm{Cov}(\boldsymbol{x})\boldsymbol{A}^T + \boldsymbol{A}\mathrm{Cov}(\boldsymbol{x}, \boldsymbol{y})\boldsymbol{B}^T + \boldsymbol{B}\mathrm{Cov}(\boldsymbol{y}, \boldsymbol{x})\boldsymbol{A}^T + \boldsymbol{B}\mathrm{Cov}(\boldsymbol{y})\boldsymbol{B}^T. \end{aligned}$$

  Special case: $\mathrm{Var}(ax + by) = a^2\mathrm{Var}(x) + 2ab\mathrm{Cov}(x, y) + b^2\mathrm{Var}(y)$ for random variables x, y.

- Expectation of quadratic form. Let $\boldsymbol{x} \in \mathbb{R}^n$ be a random vector with mean $\mathrm{E}(\boldsymbol{x}) = \boldsymbol{\mu}$ and covariance $\mathrm{Cov}(\boldsymbol{x}) = \boldsymbol{\Omega}$. $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ is a constant matrix. Then

$$\mathrm{E}(\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}) = \mathrm{tr}(\boldsymbol{A}\boldsymbol{\Omega}) + \boldsymbol{\mu}^T\boldsymbol{A}\boldsymbol{\mu}.$$

  Proof. $\mathrm{E}(\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}) = \mathrm{Etr}(\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}) = \mathrm{Etr}(\boldsymbol{A}\boldsymbol{x}\boldsymbol{x}^T) = \mathrm{trE}(\boldsymbol{A}\boldsymbol{x}\boldsymbol{x}^T) = \mathrm{tr}\boldsymbol{A}\mathrm{E}(\boldsymbol{x}\boldsymbol{x}^T) = \mathrm{tr}(\boldsymbol{A}(\boldsymbol{\Omega} + \boldsymbol{\mu}\boldsymbol{\mu}^T)) = \mathrm{tr}(\boldsymbol{A}\boldsymbol{\Omega}) + \boldsymbol{\mu}^T\boldsymbol{A}\boldsymbol{\mu}$. □

  Special case: $\mathrm{E}(ax^2) = a\mathrm{Var}(x) + a[\mathrm{E}(x)]^2$ for a random variable x.

5 Lecture 5: Sep 16

Announcement

- HW1 returned.
- HW2 is posted and it is due next Wed, Sep 25.
  `http://hua-zhou.github.io/teaching/st552-2013fall/ST552-2013-HW2.pdf`
  Recitations in afternoon sessions of Sep 18 and Sep 25.
- Read JM chapter 3.

Last time

- Geometry of least squares solution: $\boldsymbol{P_X} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T$, fitted values $\hat{\boldsymbol{y}} = \boldsymbol{P_X}\boldsymbol{y}$, residuals $\hat{\boldsymbol{e}} = (\boldsymbol{I}_n - \boldsymbol{P_X})\boldsymbol{y}$, $\hat{\boldsymbol{y}} \perp \hat{\boldsymbol{e}}$, $\|\boldsymbol{y}\|_2^2 = \|\hat{\boldsymbol{y}}\|_2^2 + \|\hat{\boldsymbol{e}}\|_2^2$
- Gram-Schmidt orthogonalization and QR decomposition: GS as a series of linear regressions, $\boldsymbol{X} = \boldsymbol{Q}\boldsymbol{R}$ as a practical way to solve normal equation, see HW2 for the case of non-full rank $\boldsymbol{X}$
- Reparameterizations (equivalent models): $\mathcal{C}(\boldsymbol{X}) = \mathcal{C}(\boldsymbol{W})$, $\boldsymbol{P_X} = \boldsymbol{P_W}$, translation between solutions of equivalent models
- Linear mean model: $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{e}$, where $\mathbf{E}\boldsymbol{e} = \mathbf{0}$
- Random vectors

Today

- Positive (semi)definite matrix
- Estimable functions

Positive (semi)definite matrix

Assume $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is symmetric.

- A real symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is *positive semi-definite* (or *nonnegative definite*, or p.s.d.) if $\boldsymbol{x}^\top\boldsymbol{A}\boldsymbol{x} \geq 0$ for all $\boldsymbol{x}$. Notation $\boldsymbol{A} \succeq \mathbf{0}_{n\times n}$.

- E.g., the *Gramian matrix* $\boldsymbol{X}^\intercal\boldsymbol{X}$ or $\boldsymbol{X}\boldsymbol{X}^T$.
- The notation $\boldsymbol{A} \geq \boldsymbol{0}_{n\times n}$ means $a_{ij} \geq 0$ for all i, j.
- If inequality is strict for all $\boldsymbol{x} \neq \boldsymbol{0}$, then $\boldsymbol{A}$ is *positive definite*. $\boldsymbol{A} \succ \boldsymbol{0}_{n\times n}$.
- The notation $\boldsymbol{A} > \boldsymbol{0}_{n\times n}$ means $a_{ij} > 0$ for all i, j.
- We write $\boldsymbol{A} \succeq \boldsymbol{B}$ means $\boldsymbol{A} - \boldsymbol{B} \succeq \boldsymbol{0}_{n\times n}$.
- If $\boldsymbol{A} \succeq \boldsymbol{B}$, then $\det(\boldsymbol{A}) \geq \det(\boldsymbol{B})$ with equality if and only if $\boldsymbol{A} = \boldsymbol{B}$.
- Cholesky decomposition. Each positive semidefinite matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ can be factorized as $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{L}^T$ for some lower triangular matrix $\boldsymbol{L} \in \mathbb{R}^{n\times n}$ with nonnegative diagonal entries.
- $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is positive semidefinite if and only if $\boldsymbol{A}$ is a covariance matrix of a random vector.

  Proof. "If part": let $\boldsymbol{A} = \mathrm{Cov}(\boldsymbol{x})$ for some random vector $\boldsymbol{x}$. Then for any constant $\boldsymbol{c}$ of same length as $\boldsymbol{x}$, $\boldsymbol{c}^T\boldsymbol{A}\boldsymbol{c} = \boldsymbol{c}^T\mathrm{Cov}(\boldsymbol{x})\boldsymbol{c} = \mathrm{Var}(\boldsymbol{c}^T\boldsymbol{x}) \geq 0$. "Only if part": let $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{L}^T$ be the Cholesky decomposition and $\boldsymbol{x}$ a vector of iid standard normals. Then $\boldsymbol{L}\boldsymbol{x}$ has covariance matrix $\boldsymbol{L}\mathrm{Cov}(\boldsymbol{x})\boldsymbol{L}^T = \boldsymbol{L}\boldsymbol{I}_n\boldsymbol{L}^T = \boldsymbol{A}$. □

Estimable function

Assume the linear mean model: $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{e}$, $\mathrm{E}(\boldsymbol{e}) = \boldsymbol{0}$. One main interest is estimation of the underlying parameter $\boldsymbol{b}$. Can $\boldsymbol{b}$ be estimated or what functions of $\boldsymbol{b}$ can be estimated?

- A parametric function $\boldsymbol{\Lambda}\boldsymbol{b}$, $\boldsymbol{\Lambda} \in \mathbb{R}^{m\times p}$, is said to be (linearly) *estimable* if there exists an *affinely unbiased estimator* of $\boldsymbol{\Lambda}\boldsymbol{b}$ for all $\boldsymbol{b} \in \mathbb{R}^p$. That is there exist constants $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ and $\boldsymbol{c} \in \mathbb{R}^m$ such that $\mathrm{E}(\boldsymbol{A}\boldsymbol{y} + \boldsymbol{c}) = \boldsymbol{\Lambda}\boldsymbol{b}$ for all $\boldsymbol{b}$.
- Theorem: Assuming the linear mean model, the parametric function $\boldsymbol{\Lambda}\boldsymbol{b}$ is (linearly) estimable if and only if $\mathcal{C}(\boldsymbol{\Lambda}^T) \subset \mathcal{C}(\boldsymbol{X}^T)$, or equivalently $\mathcal{N}(\boldsymbol{X}) \subset \mathcal{N}(\boldsymbol{\Lambda})$.
  "$\boldsymbol{\Lambda}\boldsymbol{b}$ is estimable $\Leftrightarrow$ the row space of $\boldsymbol{\Lambda}$ is contained in the row space of $\boldsymbol{X}$ $\Leftrightarrow$ the null space of $\boldsymbol{X}$ is contained in the null space of $\boldsymbol{\Lambda}$."

Proof. Let $\boldsymbol{A}\boldsymbol{y} + \boldsymbol{c}$ be an affine estimator of $\boldsymbol{\Lambda}\boldsymbol{b}$. Unbiasedness requires

$$\mathrm{E}(\boldsymbol{A}\boldsymbol{y} + \boldsymbol{c}) = \boldsymbol{A}\mathrm{E}(\boldsymbol{y}) + \boldsymbol{c} = \boldsymbol{A}\boldsymbol{X}\boldsymbol{b} + \boldsymbol{c} = \boldsymbol{\Lambda}\boldsymbol{b}$$

for all $\boldsymbol{b} \in \mathbb{R}^p$. Taking the special value $\boldsymbol{b} = \boldsymbol{0}$ shows that $\boldsymbol{c} = \boldsymbol{0}$. Thus $(\boldsymbol{A}\boldsymbol{X} - \boldsymbol{\Lambda})\boldsymbol{b} = \boldsymbol{0}$ for all $\boldsymbol{b}$. Now taking special values $\boldsymbol{b} = \boldsymbol{e}_i$ shows that the columns of the matrix $\boldsymbol{A}\boldsymbol{X} - \boldsymbol{\Lambda}$ are all zeros. This implies $\boldsymbol{A}\boldsymbol{X} = \boldsymbol{\Lambda}$. Therefore the matrix $\boldsymbol{A}$ exists if and only if rows of $\boldsymbol{\Lambda}$ are linear combinations of the rows of $\boldsymbol{X}$, that is, if and only if $\mathcal{C}(\boldsymbol{\Lambda}^T) \subset \mathcal{C}(\boldsymbol{X}^T)$. □

- Corollary: $\boldsymbol{X}\boldsymbol{b}$ is estimable.
  "Expected value of any observation $\mathrm{E}(y_i)$ and their linear combinations are estimable."

- Corollary: If $\boldsymbol{X}$ has full column rank, then any linear combinations of $\boldsymbol{b}$ are estimable.

- If $\boldsymbol{\Lambda}\boldsymbol{b}$ is (linearly) estimable, then its *least squares estimator* $\boldsymbol{\Lambda}\hat{\boldsymbol{b}}$ is invariant to the choice of the least squares solution $\hat{\boldsymbol{b}}$.

  Proof. Let $\hat{\boldsymbol{b}}_1, \hat{\boldsymbol{b}}_2$ be two least squares solutions. Then $\hat{\boldsymbol{b}}_1 - \hat{\boldsymbol{b}}_2 \in \mathcal{N}(\boldsymbol{X}^T\boldsymbol{X}) = \mathcal{N}(\boldsymbol{X}) \subset \mathcal{N}(\boldsymbol{\Lambda})$. Hence $\boldsymbol{\Lambda}(\hat{\boldsymbol{b}}_1 - \hat{\boldsymbol{b}}_2) = \boldsymbol{0}$, that is $\boldsymbol{\Lambda}\hat{\boldsymbol{b}}_1 = \boldsymbol{\Lambda}\hat{\boldsymbol{b}}_2$. □

- The least squares estimator $\boldsymbol{\Lambda}\hat{\boldsymbol{b}}$ is a linearly unbiased estimator of $\boldsymbol{\Lambda}\boldsymbol{b}$.

  Proof. The least squares solution takes the general form

  $$\hat{\boldsymbol{b}} = (\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{y} + [\boldsymbol{I}_p - (\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{X}]\boldsymbol{q}$$

  where $\boldsymbol{q} \in \mathbb{R}^p$ is arbitrary. Thus the least squares estimator

  $$\begin{aligned} \boldsymbol{\Lambda}\hat{\boldsymbol{b}} &= \boldsymbol{\Lambda}(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{\Lambda}[\boldsymbol{I}_p - (\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{X}]\boldsymbol{q} \\ &= \boldsymbol{\Lambda}(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{y} \end{aligned}$$

  is a linear function of $\boldsymbol{y}$. Now

  $$\begin{aligned} \mathrm{E}(\boldsymbol{\Lambda}\hat{\boldsymbol{b}}) &= \boldsymbol{\Lambda}(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T\mathrm{E}(\boldsymbol{y}) \\ &= \boldsymbol{\Lambda}(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{b} \\ &= \boldsymbol{\Lambda}\boldsymbol{b}, \end{aligned}$$

since $\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^-$ is a projection onto $\mathcal{C}(\boldsymbol{X}^T\boldsymbol{X}) = \mathcal{C}(\boldsymbol{X}^T)$ and $\mathcal{C}(\boldsymbol{\Lambda}^T) \subset \mathcal{C}(\boldsymbol{X}^T)$. Therefore the least squares estimator is unbiased. □

- General recipes for showing estimability of $\boldsymbol{\Lambda b}$:

  1. Identify each row of $\boldsymbol{\Lambda}$ as a linear combination of rows of $\boldsymbol{X}$.
  2. Find a projection $\boldsymbol{P}$ onto $\mathcal{C}(\boldsymbol{X}^T)$, such as $\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^-$ or $\boldsymbol{P}_{\boldsymbol{X}^T} = \boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T)^-\boldsymbol{X}$. $\boldsymbol{\Lambda b}$ is estimable if and only if $\boldsymbol{P}\boldsymbol{\Lambda}^T = \boldsymbol{\Lambda}^T$.
  3. Find a basis $\boldsymbol{B} = (\boldsymbol{b}_1, \ldots, \boldsymbol{b}_{p-r}) \in \mathbb{R}^{p\times(p-r)}$ of $\mathcal{N}(\boldsymbol{X})$. $\boldsymbol{\Lambda b}$ is estimable if and only if $\boldsymbol{\Lambda B} = \boldsymbol{0}_{m\times(p-r)}$.

One-way ANOVA analysis

- Example: We want to study whether the tips (percentage of meal cost) at a restaurant depends on the gender of the service person (waiter or waitress)? Data may be presented as

Waiter	Waitress
0.19, 0.15, 0.10	0.15, 0.14, 0.20

  Responses: tip. Factors: gender of service person (waiter, waitress).

- Model: $y_{ij} = \mu + \alpha_i + e_{ij}$, $i = 1, \ldots, a$ (a groups), $j = 1, \ldots, n_i$ (i-th group has n_i observations).

$$\mathrm{E}(\boldsymbol{y}) = \boldsymbol{X}\boldsymbol{b} = \begin{pmatrix} \boldsymbol{1}_{n_1} & \boldsymbol{1}_{n_1} & & & \\ \boldsymbol{1}_{n_2} & & \boldsymbol{1}_{n_2} & & \\ \vdots & & & \ddots & \\ \boldsymbol{1}_{n_a} & & & & \boldsymbol{1}_{n_a} \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_a \end{pmatrix}.$$

  In total we have $n = \sum_{i=1}^a n_i$ observations and $p = a + 1$ parameters.

- $\boldsymbol{X}$ has rank $r = a$ and thus $\dim(\mathcal{N}(\boldsymbol{X})) = 1$. An obvious basis for $\mathcal{N}(\boldsymbol{X})$ is

$$\begin{pmatrix} 1 \\ -\boldsymbol{1}_a \end{pmatrix}.$$

Therefore $\boldsymbol{\lambda}^T\boldsymbol{b}$ is estimable if and only if $\boldsymbol{\lambda}^T \begin{pmatrix} 1 \\ -\mathbf{1}_a \end{pmatrix} = \lambda_0 - \sum_{i=1}^a \lambda_i = 0$.

- Examples of estimable functions:
  - grand mean: $\mu + \bar{\alpha}_\cdot$, where $\bar{\alpha}_\cdot = a^{-1}\sum_{i=1}^a \alpha_i$
  - cell means: $\mu + \alpha_i$
  - *contrast*: $\sum_{i=1}^a d_i\alpha_i$ where $\sum_{i=1}^a d_i = 0$

  Examples of non-estimable function:
  - individual parameters: μ, α_i
  - $\alpha_i + \alpha_j$

- To find the least squares solution, we form the normal equation

$$\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{b} = \begin{pmatrix} n & n_1 & n_2 & \cdots & n_a \\ n_1 & n_1 & & & \\ n_2 & & n_2 & & \\ \vdots & & & \ddots & \\ n_a & & & & n_a \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_a \end{pmatrix} = \begin{pmatrix} n\bar{y}_{\cdot\cdot} \\ n_1\bar{y}_{1\cdot} \\ n_2\bar{y}_{2\cdot} \\ \vdots \\ n_a\bar{y}_{a\cdot} \end{pmatrix} = \boldsymbol{X}^T\boldsymbol{y}.$$

  A generalized inverse of $\boldsymbol{X}^T\boldsymbol{X}$ is (see previous notes for finding generalized inverses)

$$(\boldsymbol{X}^T\boldsymbol{X})^- = \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 \\ 0 & n_1^{-1} & & & \\ 0 & & n_2^{-1} & & \\ \vdots & & & \ddots & \\ 0 & & & & n_a^{-1} \end{pmatrix}.$$

  Therefore the least squares solution takes the general form

$$\hat{\boldsymbol{b}} \;=\; (\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{y} + [\boldsymbol{I}_p - (\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{X}]\boldsymbol{q} = \begin{pmatrix} 0 \\ \bar{y}_{1\cdot} \\ \bar{y}_{2\cdot} \\ \vdots \\ \bar{y}_{a\cdot} \end{pmatrix} + z \begin{pmatrix} 1 \\ -1 \\ -1 \\ \vdots \\ -1 \end{pmatrix}$$

and the least squares estimator of an estimable function $\boldsymbol{\lambda}^T\boldsymbol{b}$ is

$$\boldsymbol{\lambda}^T \begin{pmatrix} 0 \\ \bar{y}_{1\cdot} \\ \bar{y}_{2\cdot} \\ \vdots \\ \bar{y}_{a\cdot} \end{pmatrix} = \sum_{i=1}^{a} \lambda_i \bar{y}_{i\cdot}.$$

For example, least squares estimators of $\mu + \alpha_i$ and $\sum_{i=1}^{a} d_i\alpha_i$ are $\bar{y}_{i\cdot}$ and $\sum_{i=1}^{a} d_i\bar{y}_{i\cdot}$ respectively.

6 Lecture 6: Sep 18

Announcement

- HW2 recitation this afternoon
- Read JM chapter 4

Last time

- Positive (semi)definite matrix
- Estimability: $\boldsymbol{\Lambda b}$ is estimable if and only if $\mathcal{C}(\boldsymbol{\Lambda}^T) \subset \mathcal{C}(\boldsymbol{X}^T)$, or equivalently $\mathcal{N}(\boldsymbol{X}) \subset \mathcal{N}(\boldsymbol{\Lambda})$, or equivalently $\boldsymbol{\Lambda}^T \perp \mathcal{N}(\boldsymbol{X})$
- The least squares estimator $\boldsymbol{\Lambda}\hat{\boldsymbol{b}}$ is a linearly unbiased estimate for an estimable function $\boldsymbol{\Lambda b}$ and invariant to the choice of least squares solution $\hat{\boldsymbol{b}}$
- One-way ANOVA

Today

- Two-way ANOVA without and with interaction
- Estimability under reparameterizations

Two-way ANOVA without interaction

- Example: We want to study whether the tips (percentage of meal cost) at a restaurant depends on the gender of the service person (waitress or waiter) and the gender of paying customer? Data may be presented as

	Waiter	Waitress
Male customer	0.19, 0.15, 0.10	0.15, 0.14, 0.20
Female customer	0.16, 0.18, 0.14	0.13, 0.15, 0.19

  Responses: tip. Factors: gender of service person (waiter, waitress), gender of paying customer (male, female).

- Model $y_{ijk} = \mu + \alpha_i + \beta_j + e_{ijk}$, $i = 1, \ldots, a$ (a levels in factor 1), $j = 1, \ldots, b$ (b levels in factor 2), and $k = 1, \ldots, n_{ij}$ (n_{ij} observations in the (i, j)-th cell). In total we have $n = \sum_{i,j} n_{ij}$ observations and $p = a + b + 1$ parameters. For simplicity, we consider the case without replicates, i.e., $n_{ij} = 1$. Note adding more replicates to each cell does *not* change the rank of $\boldsymbol{X}$.

$$\mathrm{E}(\boldsymbol{y}) = \boldsymbol{X}\boldsymbol{b} = \begin{pmatrix} \mathbf{1}_b & \mathbf{1}_b & & & \boldsymbol{I}_b \\ \mathbf{1}_b & & \mathbf{1}_b & & \boldsymbol{I}_b \\ \vdots & & & \ddots & \\ \mathbf{1}_b & & & \mathbf{1}_b & \boldsymbol{I}_b \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \vdots \\ \alpha_a \\ \beta_1 \\ \vdots \\ \beta_b \end{pmatrix}.$$

- $\boldsymbol{X} \in \mathbb{R}^{ab \times (1+a+b)}$ has rank $r = a + b - 1$ and $\dim(\mathcal{N}(\boldsymbol{X})) = 2$. An obvious basis for $\mathcal{N}(\boldsymbol{X})$ is

$$\left\{ \begin{pmatrix} 1 \\ -\mathbf{1}_a \\ \mathbf{0}_b \end{pmatrix}, \quad \begin{pmatrix} 1 \\ \mathbf{0}_a \\ -\mathbf{1}_b \end{pmatrix} \right\}.$$

  Therefore $\boldsymbol{\lambda}^T \boldsymbol{b}$ is estimable if and only if $\boldsymbol{\lambda}^T \begin{pmatrix} 1 \\ -\mathbf{1}_a \\ \mathbf{0}_b \end{pmatrix} = \lambda_0 - \sum_{i=1}^a \lambda_i = 0$ *and* $\boldsymbol{\lambda}^T \begin{pmatrix} 1 \\ \mathbf{0}_a \\ -\mathbf{1}_b \end{pmatrix} = \lambda_0 - \sum_{j=1}^b \lambda_{a+j} = 0$.

- Examples of estimable functions:
  - grand mean: $\mu + \bar{\alpha}_\cdot + \bar{\beta}_\cdot$, where $\bar{\alpha}_\cdot = a^{-1} \sum_{i=1}^a \alpha_i$ and $\bar{\beta}_\cdot = b^{-1} \sum_{j=1}^b \beta_j$
  - cell means: $\mu + \alpha_i + \beta_j$
  - *contrast* $\sum_{i=1}^a d_i \alpha_i$ where $\sum_{i=1}^a d_i = 0$
  - *contrast* $\sum_{j=1}^b f_j \beta_j$ where $\sum_{j=1}^b f_j = 0$
- Examples of non-estimable functions:

  - individual parameters: μ, α_i, β_j
  - marginal means: $\mu + \alpha_i$, $\mu + \beta_j$
  - $\alpha_i - \beta_j$

- To find the least squares estimators for these estimable functions, see HW2.

Two-way ANOVA with interaction

- Example: Responses: tip. Factors: gender of service person (waiter, waitress), gender of paying customer (male, female), and their interaction.

- Model $y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}$, $i = 1, \ldots, a$ (a levels in factor 1), $j = 1, \ldots, b$ (b levels in factor 2), and $k = 1, \ldots, n_{ij}$ (n_{ij} observations in the (i, j)-th cell). In total we have $n = \sum_{i,j} n_{ij}$ observations and $p = 1 + a + b + ab$ parameters.

- We consider the simple case without replicates, i.e., $n_{ij} = 1$. Note adding more replicates to each cell does *not* change the rank of $\boldsymbol{X}$.

$$\mathrm{E}(\boldsymbol{y}) = \boldsymbol{X}\boldsymbol{b} = \begin{pmatrix} \mathbf{1}_b & \mathbf{1}_b & & & & \boldsymbol{I}_b & \boldsymbol{I}_b & & & \\ \mathbf{1}_b & & \mathbf{1}_b & & & \boldsymbol{I}_b & & \boldsymbol{I}_b & & \\ \vdots & & & \ddots & & \vdots & & & \ddots & \\ \mathbf{1}_b & & & & \mathbf{1}_b & \boldsymbol{I}_b & & & & \boldsymbol{I}_b \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \vdots \\ \alpha_a \\ \beta_1 \\ \vdots \\ \beta_b \\ \gamma_{11} \\ \gamma_{12} \\ \vdots \\ \gamma_{a,(b-1)} \\ \gamma_{ab} \end{pmatrix}.$$

- $\boldsymbol{X} \in \mathbb{R}^{ab \times (1+a+b+ab)}$ has rank $r = ab$ and $\dim(\mathcal{N}(\boldsymbol{X})) = 1 + a + b$. A basis for $\mathcal{N}(\boldsymbol{X})$ is

$$\left\{ \begin{pmatrix} -1 \\ \mathbf{1}_a \\ \mathbf{0}_b \\ \mathbf{0}_a \otimes \mathbf{0}_b \end{pmatrix}, \quad \begin{pmatrix} 0 \\ -\boldsymbol{e}_i \\ \mathbf{0}_b \\ \boldsymbol{e}_i \otimes \mathbf{1}_b \end{pmatrix}, i = 1, \ldots, a, \begin{pmatrix} 0 \\ \mathbf{0}_a \\ -\boldsymbol{e}_j \\ \mathbf{1}_a \otimes \boldsymbol{e}_j \end{pmatrix}, j = 1, \ldots, b \right\}.$$

Therefore $\boldsymbol{\lambda}^T\boldsymbol{b}$ is estimable if and only if $\boldsymbol{\lambda}$ is orthogonal to all these basis vectors.

- Examples of estimable functions:
  - grand mean: $\mu + \bar{\alpha}_\cdot + \bar{\beta}_\cdot + \bar{\gamma}_{\cdot\cdot}$
  - cell means: $\mu + \alpha_i + \beta_j + \gamma_{ij}$
  - main effect differences: $(\alpha_i + \bar{\gamma}_{i\cdot}) - (\alpha_{i'} + \bar{\gamma}_{i'\cdot})$, $(\beta_j + \bar{\gamma}_{\cdot j}) - (\beta_{j'} + \bar{\gamma}_{\cdot j'})$
  - Interaction effect: $\gamma_{ij} - \gamma_{ij'} - \gamma_{i'j} + \gamma_{i'j'}$
- Examples of non-estimable functions:
  - individual parameters: μ, α_i, β_j, γ_{ij}
  - marginal means: $\mu + \alpha_i + \bar{\gamma}_{i\cdot}$, where $\bar{\gamma}_{i\cdot} = b^{-1}\sum_{j=1}^{b}\gamma_{ij}$
  - marginal means: $\mu + \beta_j + \bar{\gamma}_{\cdot j}$, where $\bar{\gamma}_{\cdot j} = a^{-1}\sum_{i=1}^{a}\gamma_{ij}$

Estimability under reparameterizations

- Recall that two linear models $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{e}$ and $\boldsymbol{y} = \boldsymbol{W}\boldsymbol{c} + \boldsymbol{e}$ are equivalent (*reparameterizations*) if $\mathcal{C}(\boldsymbol{X}) = \mathcal{C}(\boldsymbol{W})$. There exist transformations $\boldsymbol{S}$ and $\boldsymbol{T}$ such that

$$\boldsymbol{X} = \boldsymbol{W}\boldsymbol{S} \quad \text{and} \quad \boldsymbol{W} = \boldsymbol{X}\boldsymbol{T}.$$

Thus

$$\begin{aligned} \mathrm{E}(\boldsymbol{y}) &= \boldsymbol{X}\boldsymbol{b} = \boldsymbol{W}\boldsymbol{S}\boldsymbol{b} \\ &= \boldsymbol{W}\boldsymbol{c} = \boldsymbol{X}\boldsymbol{T}\boldsymbol{c}. \end{aligned}$$

- We already know how to translate least square solutions between equivalent models:
  - If $\hat{\boldsymbol{c}}$ solves $\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{c} = \boldsymbol{W}^T\boldsymbol{y}$, then $\hat{\boldsymbol{b}} = \boldsymbol{T}\hat{\boldsymbol{c}}$ solves $\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{b} = \boldsymbol{X}^T\boldsymbol{y}$.
  - If $\hat{\boldsymbol{b}}$ solves $\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{b} = \boldsymbol{X}^T\boldsymbol{y}$, then $\hat{\boldsymbol{c}} = \boldsymbol{S}\hat{\boldsymbol{b}}$ solves $\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{c} = \boldsymbol{W}^T\boldsymbol{y}$.
- Estimability under reparameterization

– If $\boldsymbol{\Lambda c}$ is estimable under model $\boldsymbol{y} = \boldsymbol{W c} + \boldsymbol{e}$, then $\boldsymbol{\Lambda S b}$ is estimable under model $\boldsymbol{y} = \boldsymbol{X b} + \boldsymbol{e}$.

Proof. Since $\boldsymbol{\Lambda c}$ is estimable, $\mathcal{C}(\boldsymbol{\Lambda}^T) \subset \mathcal{C}(\boldsymbol{W}^T)$ and there exists a matrix $\boldsymbol{A}$ such that $\boldsymbol{\Lambda}^T = \boldsymbol{W}^T\boldsymbol{A}$. Thus $\boldsymbol{S}^T\boldsymbol{\Lambda}^T = \boldsymbol{S}^T\boldsymbol{W}^T\boldsymbol{A} = \boldsymbol{X}^T\boldsymbol{A}$. In other words, $\mathcal{C}(\boldsymbol{S}^T\boldsymbol{\Lambda}^T) \subset \mathcal{C}(\boldsymbol{X}^T)$. Therefore $\boldsymbol{\Lambda S b}$ is estimable. □

– Similarly, if $\boldsymbol{\Lambda b}$ is estimable under model $\boldsymbol{y} = \boldsymbol{X b} + \boldsymbol{e}$, then $\boldsymbol{\Lambda T c}$ is estimable under model $\boldsymbol{y} = \boldsymbol{W c} + \boldsymbol{e}$.

- Different parameterizations of one-way ANOVA analysis

  – An over-parameterized model.

$$\mathrm{E}(\boldsymbol{y}) = \boldsymbol{X b} = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{1}_{n_1} & & & \\ \mathbf{1}_{n_2} & & \mathbf{1}_{n_2} & & \\ \vdots & & & \ddots & \\ \mathbf{1}_{n_a} & & & & \mathbf{1}_{n_a} \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_a \end{pmatrix}.$$

  Individual parameters in $\boldsymbol{b}$ are *not* estimable since $\boldsymbol{X}$ does not have full column rank. Certain functions are estimable, such as the grand mean $\mu + \bar{\alpha}$, cell means $\mu + \alpha_i$, and contrasts $\sum_{i=1}^a d_i\alpha_i$ ($\sum_{i=1}^a d_i = 0$).

  – A full rank parameterization (deleting the first column of $\boldsymbol{X}$): *cell means model*

$$\mathrm{E}(\boldsymbol{y}) = \boldsymbol{Z\mu} = \begin{pmatrix} \mathbf{1}_{n_1} & & & \\ & \mathbf{1}_{n_2} & & \\ & & \ddots & \\ & & & \mathbf{1}_{n_a} \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_a \end{pmatrix}.$$

  Note

$$\boldsymbol{X} = \boldsymbol{Z} \begin{pmatrix} 1 & 1 & & \\ \vdots & & \ddots & \\ 1 & & & 1 \end{pmatrix} \text{ and } \boldsymbol{\mu} = \begin{pmatrix} \mu + \alpha_1 \\ \vdots \\ \mu + \alpha_a \end{pmatrix}.$$

  Each element of $\boldsymbol{\mu}$ (cell means) is estimable since $\boldsymbol{Z}$ has full column rank.

– A full rank parameterization (deleting the last column of $\boldsymbol{X}$): *reference cell model*

$$\mathrm{E}(\boldsymbol{y}) = \boldsymbol{W}\boldsymbol{c} = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{1}_{n_1} & & & \\ \mathbf{1}_{n_2} & & \mathbf{1}_{n_2} & & \\ \vdots & & & \ddots & \\ \mathbf{1}_{n_{a-1}} & & & & \mathbf{1}_{n_{a-1}} \\ \mathbf{1}_{n_a} & & & & \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_a \end{pmatrix}.$$

Note

$$\boldsymbol{X} = \boldsymbol{W} \begin{pmatrix} 1 & & & 1 \\ & 1 & & -1 \\ & & \ddots & \vdots \\ & & 1 & -1 \end{pmatrix} \text{ and } \boldsymbol{c} = \begin{pmatrix} \mu + \alpha_a \\ \alpha_1 - \alpha_a \\ \vdots \\ \alpha_{a-1} - \alpha_a \end{pmatrix}.$$

Each element of $\boldsymbol{c}$ is estimable since $\boldsymbol{W}$ has full column rank.

– A full rank parameterization: *difference from the mean model*

$$\mathrm{E}(\boldsymbol{y}) = \boldsymbol{U}\boldsymbol{\delta} = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{1}_{n_1} & & & \\ \mathbf{1}_{n_2} & & \mathbf{1}_{n_2} & & \\ \vdots & & & \ddots & \\ \mathbf{1}_{n_{a-1}} & & & & \mathbf{1}_{n_{a-1}} \\ \mathbf{1}_{n_a} & -\mathbf{1}_{n_a} & -\mathbf{1}_{n_a} & \cdots & -\mathbf{1}_{n_a} \end{pmatrix} \begin{pmatrix} \delta_1 \\ \delta_2 \\ \vdots \\ \delta_a \end{pmatrix}.$$

Note

$$\boldsymbol{X} = \boldsymbol{U} \begin{pmatrix} 1 & a^{-1} & a^{-1} & & a^{-1} \\ & 1 - a^{-1} & -a^{-1} & & -a^{-1} \\ & -a^{-1} & 1 - a^{-1} & & -a^{-1} \\ & \vdots & & \ddots & \\ & -a^{-1} & -a^{-1} & & 1 - a^{-1} \end{pmatrix} \text{ and } \boldsymbol{\delta} = \begin{pmatrix} \mu + \bar{\alpha} \\ \alpha_1 - \bar{\alpha} \\ \alpha_2 - \bar{\alpha} \\ \vdots \\ \alpha_a - \bar{\alpha} \end{pmatrix}.$$

Each element of $\boldsymbol{\delta}$ is estimable since $\boldsymbol{U}$ has full column rank.

Review: correlation and independence

- Two random variables $x, y \in \mathbb{R}$ are *uncorrelated* if $\mathrm{Cov}(x, y) = 0$.

- Two random vectors $\boldsymbol{x} \in \mathbb{R}^m$, $\boldsymbol{y} \in \mathbb{R}^n$ are *uncorrelated* if $\mathrm{Cov}(\boldsymbol{x}, \boldsymbol{y}) = \mathbf{0}_{m \times n}$.

- If $\boldsymbol{x}$ and $\boldsymbol{y}$ are uncorrelated random vectors, then

$$\mathrm{Cov}(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{y}) = \boldsymbol{A}\mathrm{Cov}(\boldsymbol{x})\boldsymbol{A}^T + \boldsymbol{B}\mathrm{Cov}(\boldsymbol{y})\boldsymbol{B}^T.$$

  Special case: If x, y are uncorrelated random variables, then $\mathrm{Cov}(ax + by) = a^2\mathrm{Var}(x) + b^2\mathrm{Var}(b)$.

- We say random vectors $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are (mutually) independent if their joint density factorizes as product of marginal densities

$$f_{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = \prod_{i=1}^{n} f_{\boldsymbol{X}_i}(\boldsymbol{x}_i)$$

  for all values of $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$.

- Mutual independence implies *pairwise independence*. The converse is not true in general. (HW3)

- If $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are (mutually) independent, then for any functions $g_1, \ldots, g_n$

$$\mathrm{E}\prod_{i=1}^{n} g_i(\boldsymbol{X}_i) = \prod_{i=1}^{n} \mathrm{E}g_i(\boldsymbol{X}_i).$$

- If two random vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ are independent, then they are uncorrelated. The converse is not true in general. (HW3)

  Proof. Corollary of the preceding result. □

7 Lecture 7: Sep 23

Announcement

- Reminder: HW2 due this Wed
- HW3 posted and due Oct 2
  `http://hua-zhou.github.io/teaching/st552-2013fall/ST552-2013-HW3.pdf`
- HW 2/3 recitation this Wed afternoon

Last time

- Estimability of two-way ANOVA without and with interaction
- Estimability under reparameterizations
- 4 commonly used parameterizations of one-way ANOVA
- Review: independence and correlation
- Gauss-Markov model: introduction

Today

- Gauss-Markov theorem
- Least squares estimator of σ^2
- Underfitting and overfitting
- Aitken model

Gauss-Markov model and Gauss-Markov theorem (JM 4.2)

- Assumptions of Gauss-Markov model:

  $$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{e},$$

  where $\mathrm{E}(\boldsymbol{e}) = \boldsymbol{0}$ and $\mathrm{Cov}(\boldsymbol{e}) = \sigma^2 \boldsymbol{I}_n$. In words, errors have zero mean, constant variance (homoskedasiticity), and are uncorrelated. Equivalently, $\mathrm{E}(\boldsymbol{y}) = \boldsymbol{X}\boldsymbol{b}$ and $\mathrm{Cov}(\boldsymbol{y}) = \sigma^2 \boldsymbol{I}_n$.

- Under linear mean model, we showed that the *least squares estimator* $\boldsymbol{\Lambda}\hat{\boldsymbol{b}}$, where $\hat{\boldsymbol{b}}$ is any least squares solution to the normal equation, is a linearly unbiased estimator of an estimable function $\boldsymbol{\Lambda b}$ ($\mathcal{C}(\boldsymbol{\Lambda}^T) \subset \mathcal{C}(\boldsymbol{X}^T)$).

- Under the Gauss-Markov assumptions, we can compute the covariance matrix of the least squares estimator

$$\begin{aligned}
\mathrm{Cov}(\boldsymbol{\Lambda}\hat{\boldsymbol{b}}) &= \mathrm{Cov}(\boldsymbol{\Lambda}(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{y}) \\
&= \boldsymbol{\Lambda}(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T\mathrm{Cov}(\boldsymbol{y})\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{\Lambda}^T \\
&= \sigma^2\boldsymbol{\Lambda}(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{\Lambda}^T \\
&= \sigma^2\boldsymbol{\Lambda}(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{\Lambda}^T.
\end{aligned}$$

  The last equality is because $\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^-$ is a projection onto $\mathcal{C}(\boldsymbol{X}^T\boldsymbol{X}) = \mathcal{C}(\boldsymbol{X}^T) \supset \mathcal{C}(\boldsymbol{\Lambda}^T)$.

- The celebrated Gauss-Markov theorem states that the least squares estimator $\boldsymbol{\Lambda}\hat{\boldsymbol{b}}$ has the smallest variance among all linear unbiased estimators of $\boldsymbol{\Lambda b}$. Why smaller variance is better?

- Gauss-Markov Theorem (vector version). Under Gauss-Markov assumptions, if $\boldsymbol{\Lambda b}$ is estimable, then the least squares estimator $\boldsymbol{\Lambda}\hat{\boldsymbol{b}}$ is the *best (minimum variance) affine unbiased estimator* (MVAUE). That is

$$\mathrm{Cov}(\hat{\boldsymbol{\theta}}) \succeq \mathrm{Cov}(\boldsymbol{\Lambda}\hat{\boldsymbol{b}})$$

  for any affine unbiased estimator $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\Lambda b}$.

  Proof. Let $\hat{\boldsymbol{\theta}} = \boldsymbol{C}\boldsymbol{y} + \boldsymbol{d}$. Unbiasedness requires $\mathrm{E}(\hat{\boldsymbol{\theta}}) = \boldsymbol{C}\boldsymbol{X}\boldsymbol{b} + \boldsymbol{d} = \boldsymbol{\Lambda b}$ for all $\boldsymbol{b}$. Taking $\boldsymbol{b} = \boldsymbol{0}$ shows that $\boldsymbol{d} = \boldsymbol{0}$. Thus $\hat{\boldsymbol{\theta}}$ must be a linear estimator.

$$\begin{aligned}
& \mathrm{Cov}(\hat{\boldsymbol{\theta}}) \\
= \ & \mathrm{Cov}(\boldsymbol{C}\boldsymbol{y}) \\
= \ & \mathrm{Cov}(\boldsymbol{\Lambda}\hat{\boldsymbol{b}} + \boldsymbol{C}\boldsymbol{y} - \boldsymbol{\Lambda}\hat{\boldsymbol{b}}) \\
= \ & \mathrm{Cov}(\boldsymbol{\Lambda}\hat{\boldsymbol{b}}) + \mathrm{Cov}(\boldsymbol{C}\boldsymbol{y} - \boldsymbol{\Lambda}\hat{\boldsymbol{b}}) + \mathrm{Cov}(\boldsymbol{\Lambda}\hat{\boldsymbol{b}}, \boldsymbol{C}\boldsymbol{y} - \boldsymbol{\Lambda}\hat{\boldsymbol{b}}) + \mathrm{Cov}(\boldsymbol{C}\boldsymbol{y} - \boldsymbol{\Lambda}\hat{\boldsymbol{b}}, \boldsymbol{\Lambda}\hat{\boldsymbol{b}}).
\end{aligned}$$

The covariance terms vanish because

$$\begin{aligned}
& \mathrm{Cov}(\boldsymbol{\Lambda}\hat{\boldsymbol{b}}, \boldsymbol{C}\boldsymbol{y} - \boldsymbol{\Lambda}\hat{\boldsymbol{b}}) \\
= \; & \mathrm{E}(\boldsymbol{\Lambda}\hat{\boldsymbol{b}} - \boldsymbol{\Lambda}\boldsymbol{b})(\boldsymbol{C}\boldsymbol{y} - \boldsymbol{\Lambda}\hat{\boldsymbol{b}})^T \\
= \; & \mathrm{E}(\boldsymbol{\Lambda}\hat{\boldsymbol{b}}\boldsymbol{y}^T\boldsymbol{C}^T) - \mathrm{E}(\boldsymbol{\Lambda}\hat{\boldsymbol{b}}\hat{\boldsymbol{b}}^T\boldsymbol{\Lambda}^T) - \mathrm{E}(\boldsymbol{\Lambda}\boldsymbol{b}\boldsymbol{y}^T\boldsymbol{C}^T) + \mathrm{E}(\boldsymbol{\Lambda}\boldsymbol{b}\hat{\boldsymbol{b}}^T\boldsymbol{\Lambda}^T) \\
= \; & \mathrm{E}(\boldsymbol{\Lambda}\hat{\boldsymbol{b}}\boldsymbol{y}^T\boldsymbol{C}^T) - \mathrm{E}(\boldsymbol{\Lambda}\hat{\boldsymbol{b}}\hat{\boldsymbol{b}}^T\boldsymbol{\Lambda}^T) - \boldsymbol{\Lambda}\boldsymbol{b}\boldsymbol{b}^T\boldsymbol{\Lambda}^T + \boldsymbol{\Lambda}\boldsymbol{b}\boldsymbol{b}^T\boldsymbol{\Lambda}^T \\
= \; & \mathrm{E}(\boldsymbol{\Lambda}\hat{\boldsymbol{b}}\boldsymbol{y}^T\boldsymbol{C}^T) - \mathrm{E}(\boldsymbol{\Lambda}\hat{\boldsymbol{b}}\hat{\boldsymbol{b}}^T\boldsymbol{\Lambda}^T) \\
= \; & \mathrm{E}[\boldsymbol{\Lambda}(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{y}\boldsymbol{y}^T\boldsymbol{C}^T] - \mathrm{E}[\boldsymbol{\Lambda}(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{y}\boldsymbol{y}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{\Lambda}^T] \\
= \; & \boldsymbol{\Lambda}(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T\mathrm{E}(\boldsymbol{y}\boldsymbol{y}^T)\boldsymbol{C}^T - \boldsymbol{\Lambda}(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T\mathrm{E}(\boldsymbol{y}\boldsymbol{y}^T)\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{\Lambda}^T \\
= \; & \boldsymbol{\Lambda}(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T\mathrm{E}(\boldsymbol{y}\boldsymbol{y}^T)[\boldsymbol{C}^T - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{\Lambda}^T] \\
= \; & \boldsymbol{\Lambda}(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T(\sigma^2\boldsymbol{I} + \boldsymbol{X}\boldsymbol{b}\boldsymbol{b}^T\boldsymbol{X}^T)[\boldsymbol{C}^T - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{\Lambda}^T] \\
= \; & [\sigma^2\boldsymbol{\Lambda}(\boldsymbol{X}^T\boldsymbol{X})^- + \boldsymbol{\Lambda}(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{b}\boldsymbol{b}^T][\boldsymbol{X}^T\boldsymbol{C}^T - \boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{\Lambda}^T] \\
= \; & [\sigma^2\boldsymbol{\Lambda}(\boldsymbol{X}^T\boldsymbol{X})^- + \boldsymbol{\Lambda}(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{b}\boldsymbol{b}^T][\boldsymbol{X}^T\boldsymbol{C}^T - \boldsymbol{\Lambda}^T](\text{why?}) \\
= \; & [\sigma^2\boldsymbol{\Lambda}(\boldsymbol{X}^T\boldsymbol{X})^- + \boldsymbol{\Lambda}(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{b}\boldsymbol{b}^T]\boldsymbol{0}_{p\times m}(\text{why?}) \\
= \; & \boldsymbol{0}_{m\times m}.
\end{aligned}$$

Therefore $\mathrm{Cov}(\hat{\boldsymbol{\theta}}) - \mathrm{Cov}(\boldsymbol{\Lambda}\hat{\boldsymbol{b}}) = \mathrm{Cov}(\boldsymbol{C}\boldsymbol{y} - \boldsymbol{\Lambda}\hat{\boldsymbol{b}}) \succeq \boldsymbol{0}_{m\times m}$. □

- En route, we showed that any affine unbiased estimator of $\boldsymbol{\Lambda}\boldsymbol{b}$ must be a linear estimator. Therefore we also say that the least squares estimator is BLUE (best linear unbiased estimator).

- En route, we also showed that $\mathrm{Cov}(\boldsymbol{\Lambda}\hat{\boldsymbol{b}}, \boldsymbol{C}\boldsymbol{y} - \boldsymbol{\Lambda}\hat{\boldsymbol{b}}) = \boldsymbol{0}_{m\times m}$ for any unbiased estimator $\boldsymbol{C}\boldsymbol{y}$ of $\boldsymbol{\Lambda}\hat{\boldsymbol{b}}$. Therefore the least squares estimator $\boldsymbol{\Lambda}\hat{\boldsymbol{b}}$ is uncorrelated with any (linearly) unbiased estimators of $\boldsymbol{0}_m$.

- When $\boldsymbol{X}$ has full column rank, then the least squares solution

$$\hat{\boldsymbol{b}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

  is BLUE for estimating $\boldsymbol{b}$.

- A drawback of the criterion of best (minimum variance) affine unbiased estimator (MVAUE) is that it is not operational. We cannot minimize a matrix. However, we can minimize a scalar function of matrix: trace, determinant, largest eigenvalue and so on. The trace criterion is the most convenient.

- The *affine minimum-trace unbiased estimator* (MTAUE) of an estimable function $\boldsymbol{\Lambda b}$ is an affine unbiased estimator of $\boldsymbol{\Lambda b}$, say $\widehat{\boldsymbol{\Lambda b}}$, such that,

$$\mathrm{tr}(\mathrm{Cov}(\hat{\boldsymbol{\theta}})) \geq \mathrm{tr}(\mathrm{Cov}(\widehat{\boldsymbol{\Lambda b}}))$$

  for all affine unbiased estimator $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\Lambda b}$.

- The best (minimum variance) affine unbiased estimator (MVAUE) is also an affine minimum-trace unbiased estimator (MTAUE). The converse is not true in general (HW3). If we can find an MTAUE and show that it is unique, then it must be the MVAUE unless it does not exist.

- A mechanic way to *derive* MVAUE is

  1. Let $\hat{\boldsymbol{\theta}} = \boldsymbol{Ay} + \boldsymbol{c}$ be an affine estimator of $\boldsymbol{\Lambda b}$
  2. Unbiasedness imposes $\boldsymbol{AX} = \boldsymbol{\Lambda}$ and $\boldsymbol{c} = \boldsymbol{0}$
  3. Minimize $\mathrm{tr}(\mathrm{Cov}(\hat{\boldsymbol{\theta}})) = \sigma^2 \mathrm{tr}(\boldsymbol{AA}^T)$ subject to $\boldsymbol{AX} = \boldsymbol{\Lambda}$ to determine $\boldsymbol{A}$. This yields MTAUE.
  4. Show that MTAUE is MVAUE by showing that MTAUE is unique and MVAUE exists or by direct argument

Least squares estimator of σ^2 (JM 4.3)

- Since σ^2 is a quadratic concept, we consider estimation of σ^2 by a quadratic function of $\boldsymbol{y}$. That is, a function of form

$$\boldsymbol{y}^T \boldsymbol{Ay}.$$

  Any estimator of this form is called a *quadratic estimator*. If in addition $\boldsymbol{A}$ is positive (semi)definite, then it is called a *quadratic and positive* estimator.

- An estimator $\hat{\sigma}^2$ of σ^2 is *unbiased* if

$$\mathrm{E}(\hat{\sigma}^2) = \sigma^2$$

  for all $\boldsymbol{b} \in \mathbb{R}^p$ and $\sigma^2 > 0$.

- Under Gauss-Markov assumptions ($\mathrm{E}(\boldsymbol{y}) = \boldsymbol{X}\boldsymbol{b}$ and $\mathrm{Cov}(\boldsymbol{e}) = \sigma^2\boldsymbol{I}$), the quadratic estimator

$$\hat{\sigma}^2 = \frac{1}{n-r}\boldsymbol{y}^T(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{y} = \frac{\|\boldsymbol{y} - \hat{\boldsymbol{y}}\|_2^2}{n-r} = \frac{\|\hat{\boldsymbol{e}}\|_2^2}{n-r} = \frac{\mathrm{SSE}}{n-r},$$

where $r = \mathrm{rank}(\boldsymbol{X})$, is unbiased.

Proof.

$$\begin{aligned}
\mathrm{E}(\hat{\sigma}^2) &= \frac{1}{n-r}\mathrm{E}[\boldsymbol{y}^T(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{y}] \\
&= \frac{1}{n-r}\mathrm{tr}[(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}})\mathrm{Cov}(\boldsymbol{y})] + \frac{1}{n-r}\mathrm{E}(\boldsymbol{y})^T(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}})\mathrm{E}(\boldsymbol{y}) \\
&= \frac{\sigma^2}{n-r}\mathrm{tr}(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}}) + \frac{1}{n-r}(\boldsymbol{X}\boldsymbol{b})^T(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}})(\boldsymbol{X}\boldsymbol{b}) \\
&= \frac{\sigma^2}{n-r}\mathrm{tr}(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}}) \\
&= \frac{\sigma^2}{n-r}[\mathrm{rank}(\boldsymbol{I}) - \mathrm{rank}(\boldsymbol{P}_{\boldsymbol{X}})] \\
&= \sigma^2.
\end{aligned}$$

□

Remark: $\hat{\sigma}^2 = \frac{1}{n-r}\boldsymbol{y}^T(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{y}$ is called the *least squares estimator of* σ^2.

- To show that the least squares estimator of σ^2 is best (minimum variance), we need further assumptions on the third and fourth moment of $\boldsymbol{e}$. We will do this later under the *normal* linear regression model $\boldsymbol{y} \sim N(\boldsymbol{X}\boldsymbol{b}, \sigma^2\boldsymbol{I})$.

Underfitting and misspecification (JM 4.4.1)

- Underfitting (also called misspecification) means omitting predictors in the true model.

- Let $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{\eta} + \boldsymbol{e}$ represent the true model, where $\boldsymbol{\eta}$ includes the omitted variables and their coefficients.

- The bias of a least squares estimater $\boldsymbol{\Lambda b}$ for an estimable function $\boldsymbol{\Lambda b}$ is

$$\begin{aligned}
\mathrm{E}(\boldsymbol{\Lambda}\hat{\boldsymbol{b}}) - \boldsymbol{\Lambda b} &= \boldsymbol{\Lambda}(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T\mathrm{E}(\boldsymbol{y}) - \boldsymbol{\Lambda b} \\
&= \boldsymbol{\Lambda}(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T(\boldsymbol{Xb}+\boldsymbol{\eta}) - \boldsymbol{\Lambda b} \\
&= \boldsymbol{\Lambda b} + \boldsymbol{\Lambda}(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{\eta} - \boldsymbol{\Lambda b} \\
&= \boldsymbol{\Lambda}(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{\eta} \\
&= \boldsymbol{A}^T\boldsymbol{P_X}\boldsymbol{\eta}
\end{aligned}$$

  where $\boldsymbol{\Lambda}^T = \boldsymbol{X}^T\boldsymbol{A}$. $\boldsymbol{A}$ exists since $\mathcal{C}(\boldsymbol{\Lambda}^T) \subset \mathcal{C}(\boldsymbol{X}^T)$ due to estimability.

  If $\boldsymbol{\eta} \perp \mathcal{C}(\boldsymbol{X})$, then the bias is zero.

- The bias of the least squares estimator of σ^2 is

$$\begin{aligned}
& (n-r)^{-1}\mathrm{E}[\boldsymbol{y}^T(\boldsymbol{I}-\boldsymbol{P_X})\boldsymbol{y}] - \sigma^2 \\
= \; & (n-r)^{-1}\mathrm{tr}[(\boldsymbol{I}-\boldsymbol{P_X})\sigma^2\boldsymbol{I}] + (n-r)^{-1}(\boldsymbol{Xb}+\boldsymbol{\eta})^T(\boldsymbol{I}-\boldsymbol{P_X})(\boldsymbol{Xb}+\boldsymbol{\eta}) - \sigma^2 \\
= \; & \sigma^2 + (n-r)^{-1}(\boldsymbol{Xb}+\boldsymbol{\eta})^T(\boldsymbol{I}-\boldsymbol{P_X})(\boldsymbol{Xb}+\boldsymbol{\eta}) - \sigma^2 \\
= \; & (n-r)^{-1}(\boldsymbol{Xb}+\boldsymbol{\eta})^T(\boldsymbol{I}-\boldsymbol{P_X})(\boldsymbol{Xb}+\boldsymbol{\eta}) \\
= \; & (n-r)^{-1}\boldsymbol{\eta}^T(\boldsymbol{I}-\boldsymbol{P_X})\boldsymbol{\eta}.
\end{aligned}$$

  If $\boldsymbol{\eta} \in \mathcal{C}(\boldsymbol{X})$, then the bias is 0.

- A decomposition of the misspecification $\boldsymbol{\eta}$:

$$\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{\eta} + \boldsymbol{e} = \boldsymbol{Xb} + \boldsymbol{P_X}\boldsymbol{\eta} + (\boldsymbol{I}-\boldsymbol{P_X})\boldsymbol{\eta} + \boldsymbol{e}.$$

  The piece $\boldsymbol{P_X}\boldsymbol{\eta}$ affects estimation of $\boldsymbol{\Lambda b}$; the other piece $(\boldsymbol{I}-\boldsymbol{P_X})\boldsymbol{\eta}$ affects estimation of σ^2.

- In presence of underfitting (misspecification), unbiasedness in estimating both $\boldsymbol{\Lambda b}$ and σ^2 can be achieved only when $\boldsymbol{\eta} = \boldsymbol{0}_n$.

Overfitting and multicollinearity (JM 4.4.2)

- Overfitting means including unnecessary predictors in the model.

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{b}_1 + \boldsymbol{X}_2\boldsymbol{b}_2 + \boldsymbol{e},$$

  where the second group of predictors are redundant. That is $\boldsymbol{b}_2 = \boldsymbol{0}$.

- For simplicity, assume both $\boldsymbol{X}_1$ and $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2)$ have full column rank.
- Using only $\boldsymbol{X}_1$, the least squares estimator of $\boldsymbol{b}_1$ is

$$\tilde{\boldsymbol{b}}_1 = (\boldsymbol{X}_1^T \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^T \boldsymbol{y}$$

with

$$\begin{aligned} \mathrm{E}(\tilde{\boldsymbol{b}}_1) &= \boldsymbol{b}_1 \\ \mathrm{Cov}(\tilde{\boldsymbol{b}}_1) &= \sigma^2 (\boldsymbol{X}_1^T \boldsymbol{X}_1)^{-1}. \end{aligned}$$

- Using both $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ (overfitting), the least squares estimator of $\begin{pmatrix} \boldsymbol{b}_1 \\ \boldsymbol{b}_2 \end{pmatrix}$ is

$$\begin{aligned} \begin{pmatrix} \hat{\boldsymbol{b}}_1 \\ \hat{\boldsymbol{b}}_2 \end{pmatrix} &= \left[\begin{pmatrix} \boldsymbol{X}_1^T \\ \boldsymbol{X}_2^T \end{pmatrix} \begin{pmatrix} \boldsymbol{X}_1 & \boldsymbol{X}_2 \end{pmatrix} \right]^{-1} \begin{pmatrix} \boldsymbol{X}_1^T \\ \boldsymbol{X}_2^T \end{pmatrix} \boldsymbol{y} \\ &= \begin{pmatrix} \boldsymbol{X}_1^T \boldsymbol{X}_1 & \boldsymbol{X}_1^T \boldsymbol{X}_2 \\ \boldsymbol{X}_2^T \boldsymbol{X}_1 & \boldsymbol{X}_2^T \boldsymbol{X}_2 \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{X}_1^T \boldsymbol{y} \\ \boldsymbol{X}_2^T \boldsymbol{y} \end{pmatrix} \end{aligned}$$

with

$$\begin{aligned} \mathrm{E} \begin{pmatrix} \hat{\boldsymbol{b}}_1 \\ \hat{\boldsymbol{b}}_2 \end{pmatrix} &= \begin{pmatrix} \boldsymbol{b}_1 \\ \boldsymbol{0} \end{pmatrix} \\ \mathrm{Cov} \begin{pmatrix} \hat{\boldsymbol{b}}_1 \\ \hat{\boldsymbol{b}}_2 \end{pmatrix} &= \sigma^2 \begin{pmatrix} \boldsymbol{X}_1^T \boldsymbol{X}_1 & \boldsymbol{X}_1^T \boldsymbol{X}_2 \\ \boldsymbol{X}_2^T \boldsymbol{X}_1 & \boldsymbol{X}_2^T \boldsymbol{X}_2 \end{pmatrix}^{-1}. \end{aligned}$$

The corresponding block for $\mathrm{Cov}(\hat{\boldsymbol{b}}_1)$ is (see HW3)

$$\begin{aligned} \mathrm{Cov}(\hat{\boldsymbol{b}}_1) &= \sigma^2 (\boldsymbol{X}_1^T \boldsymbol{X}_1)^{-1} \\ &\quad + \sigma^2 (\boldsymbol{X}_1^T \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^T \boldsymbol{X}_2 [\boldsymbol{X}_2^T (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}_1}) \boldsymbol{X}_2]^{-1} \boldsymbol{X}_2^T \boldsymbol{X}_1 (\boldsymbol{X}_1^T \boldsymbol{X}_1)^{-1}. \end{aligned}$$

- Therefore overfitting does not change the unbiasedness of least squares estimators. It inflates their variance.
- Also overfitting does not change the unbiasedness of the least squares estimator of σ^2 since

$$\begin{aligned} \mathrm{E}(\hat{\sigma}_1^2) &= \frac{1}{n - r_1} \mathrm{E}[\boldsymbol{y}^T (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}_1}) \boldsymbol{y}] = \sigma^2 \\ \mathrm{E}(\hat{\sigma}^2) &= \frac{1}{n - r} \mathrm{E}[\boldsymbol{y}^T (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}}) \boldsymbol{y}] = \sigma^2, \end{aligned}$$

where $r_1 = \mathrm{rank}(\boldsymbol{X}_1)$ and $r = \mathrm{rank}(\boldsymbol{X})$.

- Under orthogonality between two sets of predictors $\boldsymbol{X}_1^T \boldsymbol{X}_2 = \mathbf{0}_{p_1 \times p_2}$, we have $\tilde{\boldsymbol{b}}_1 = \hat{\boldsymbol{b}}_1$ and $\text{Cov}(\tilde{\boldsymbol{b}}_1) = \text{Cov}(\hat{\boldsymbol{b}}_1)$.
- When $\boldsymbol{X}_2$ gets close to $\boldsymbol{X}_1$ (*multicollinearity*), the second term of $\text{Cov}(\hat{\boldsymbol{b}}_1)$ (extra variance) explodes.

8 Lecture 8: Sep 25

Announcement

- HW2 due today
- HW3 due next Wed Oct 2
- HW 2/3 recitation this afternoon
- HW4 will be posted this week (last HW before Midterm 1)

Last time

- Gauss-Markov model ($\mathrm{E}(\boldsymbol{y}) = \boldsymbol{X}\boldsymbol{b}$ and $\mathrm{Cov}(\boldsymbol{y}) = \sigma^2\boldsymbol{I}$) and Gauss-Markov theorem (least squares estimators $\boldsymbol{\Lambda}\hat{\boldsymbol{b}}$ are MVAUE)
- Least squares estimator of σ^2, $\hat{\sigma}^2 = (n-r)^{-1}\boldsymbol{y}^T(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{y}$, is unbiased
- Underfitting (causing bias) and overfitting (inflating variance)
- Aitken model: introduction

Today

- Aitken model

Aitken model (JM 4.5)

- In the *Aitken model*, we relax the variance assumption $\mathrm{Cov}(\boldsymbol{e}) = \sigma^2\boldsymbol{I}$ to $\mathrm{Cov}(\boldsymbol{e}) = \sigma^2\boldsymbol{V}$, where $\boldsymbol{V}$ is a fixed, positive semidefinite matrix.
- Example: Heteroskedasticity model $\boldsymbol{V} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_n^2)$. For instance, $\sigma_i^2 = \sigma^2 x_i^2$, $x_i \neq 0$.

- Example: Equicorrelation model

$$\mathrm{Cov}(\boldsymbol{y}) = \mathrm{Cov}(\boldsymbol{e}) = \sigma^2\boldsymbol{I} + \tau^2\mathbf{1}_n\mathbf{1}_n^T = \begin{pmatrix} \sigma^2+\tau^2 & \tau^2 & \tau^2 & \cdots & \tau^2 \\ \tau^2 & \sigma^2+\tau^2 & & \cdots & \tau^2 \\ \tau^2 & \tau^2 & \sigma^2+\tau^2 & & \\ \vdots & & & \ddots & \\ \tau^2 & \tau^2 & & & \sigma^2+\tau^2 \end{pmatrix}$$

- Example: AR(1) (autoregressive model) $e_i = \rho e_{i-1} + a_i$ where $\mathrm{Cov}(\boldsymbol{a}) = \sigma_a^2\boldsymbol{I}$.

$$\mathrm{Cov}(\boldsymbol{y}) = \mathrm{Cov}(\boldsymbol{e}) = \frac{\sigma_a^2}{1-\rho^2}\begin{pmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{n-1} \\ \rho & 1 & \rho & \cdots & \rho^{n-2} \\ \rho^2 & \rho & 1 & & \\ \vdots & & & \ddots & \\ \rho^{n-1} & \rho^{n-2} & & & 1 \end{pmatrix} = \frac{\sigma_a^2}{1-\rho^2}\left(\rho^{|i-j|}\right)_{i,j}.$$

Estimability under Aitken model

- Under the Aitken model ($\mathrm{E}(\boldsymbol{y}) = \boldsymbol{X}\boldsymbol{b}, \mathrm{Cov}(\boldsymbol{y}) = \sigma^2\boldsymbol{V}$), a linear function $\boldsymbol{\Lambda}\boldsymbol{b}$ is estimable if and only if $\mathcal{C}(\boldsymbol{\Lambda}^T) \subset \mathcal{C}(\boldsymbol{X}^T)$.

  Remark: no assumption about singularity of $\boldsymbol{V}$ is needed.

  Proof. The previous proof for the linear mean model case ($\mathrm{E}(\boldsymbol{y}) = \boldsymbol{X}\boldsymbol{b}$) applies here since no assumption about the second moment were used. (Aitken model is just a special linear mean model.) □

Review: method of Lagrangian multipliers for optimization with equality constraints (JM A.2)

$\min_U f(\boldsymbol{x})$, $U \subset \mathbb{R}^n$, subject to constraints $g_i(\boldsymbol{x}) = 0$ for $i = 1, \ldots, m$. $g : \mathbb{R}^n \mapsto \mathbb{R}^m$.

- Lagrange multiplier theory. *Lagrangian* function $L(\boldsymbol{x}, \boldsymbol{\lambda}) = f(\boldsymbol{x}) + \boldsymbol{\lambda}^\intercal g(\boldsymbol{x}) = f(\boldsymbol{x}) + \sum_{i=1}^m \lambda_i g_i(\boldsymbol{x})$. Strategy for finding equality constrained minimum. Find the stationary point $(\boldsymbol{x}^*, \boldsymbol{\lambda}^*)$ of the Lagrangian, $\nabla_{\boldsymbol{x}} L(\boldsymbol{x}, \boldsymbol{\lambda}) = \mathbf{0}_n$ for all $\boldsymbol{\lambda} \in \mathbb{R}^n$ together with $g(\boldsymbol{x}) = \mathbf{0}_m$.

- (Necessary conditions) Assume conditions (i) $g(\boldsymbol{y}) = \mathbf{0}_m$, (2) f and g are differentiable in some n-ball $B(\boldsymbol{y})$, (iii) $Dg(\boldsymbol{y}) \in \mathbb{R}^{m\times n}$ is continuous at $\boldsymbol{y}$, (iv) $Dg(\boldsymbol{y})$ has full row rank, (v) $f(\boldsymbol{x}) \geq f(\boldsymbol{y})$ for any $\boldsymbol{x} \in B(\boldsymbol{y})$ satisfying $g(\boldsymbol{x}) = \mathbf{0}_m$ ($\boldsymbol{y}$ a local minimum subject to constraints). Then there exists $\boldsymbol{\lambda} \in \mathbb{R}^m$ satisfying $\nabla f(\boldsymbol{y}) + Dg(\boldsymbol{y})^\top \boldsymbol{\lambda} = \mathbf{0}_n$, i.e., $(\boldsymbol{y}, \boldsymbol{\lambda})$ is a stationarity point of the Lagrangian $L(\boldsymbol{x}, \boldsymbol{\lambda})$. In other words, there exists $\boldsymbol{\lambda} \in \mathbb{R}^m$, such that $\nabla L(\boldsymbol{y}, \boldsymbol{\lambda}) = \mathbf{0}_{m+n}$.

- (Sufficient conditions) (i) f twice differentiable at $\boldsymbol{y}$, (ii) g twice differentiable at $\boldsymbol{y}$, (iii) the Jacobian matrix $Dg(\boldsymbol{y}) \in \mathbb{R}^{m\times n}$ has full row rank m, (iv) it is a stationarity point of the Lagrangian at a given $\boldsymbol{\lambda} \in \mathbb{R}^m$, (v) $\boldsymbol{u}^\top d^2 f(\boldsymbol{y}) \boldsymbol{u} > 0$ for all $\boldsymbol{u} \neq \mathbf{0}_n$ satisfying $Dg(\boldsymbol{y})\boldsymbol{u} = \mathbf{0}_m$. Then $\boldsymbol{y}$ is a strict local minimum under constraint $g(\boldsymbol{y}) = \mathbf{0}_m$.

- Check condition (v). Condition (v) is equivalent to the "bordered determinantal criterion"

$$(-1)^m \det \begin{pmatrix} \mathbf{0}_{m\times m} & \boldsymbol{B}_r \\ \boldsymbol{B}_r^\top & \boldsymbol{A}_{rr} \end{pmatrix} > 0$$

  for $r = m+1, \ldots, n$, where $\boldsymbol{A}_{rr}$ is the top left r-by-r block of $d^2 f(\boldsymbol{y}) + \sum_{i=1}^m \lambda_i d^2 g_i(\boldsymbol{y})$, $\boldsymbol{B}_r \in \mathbb{R}^{m\times r}$ is the first r columns of $Dg(\boldsymbol{y})$.

- (Sufficient condition for a global minimum) Lagrangian first order condition + convexity of the Lagrangian on U for some $(\boldsymbol{y}, \boldsymbol{\lambda})$.

- Example: Linearly constrained least squares solution. $\min \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2$ subject to linear constrained $\boldsymbol{V}\boldsymbol{\beta} = \boldsymbol{d}$. Form the Lagrangian

$$L(\boldsymbol{\beta}, \boldsymbol{\lambda}) = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \boldsymbol{\lambda}^\top(\boldsymbol{V}\boldsymbol{\beta} - \boldsymbol{d}).$$

  Stationary condition says

$$\begin{aligned} \boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{X}^\top\boldsymbol{y} + \boldsymbol{V}^\top\boldsymbol{\lambda} &= \mathbf{0}_p \\ \boldsymbol{V}\boldsymbol{\beta} &= \boldsymbol{d} \end{aligned}$$

  or equivalently

$$\begin{pmatrix} \boldsymbol{X}^\top\boldsymbol{X} & \boldsymbol{V}^\top \\ \boldsymbol{V} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\lambda} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}^\top\boldsymbol{y} \\ \boldsymbol{d} \end{pmatrix}.$$

Review: derivatives of trace functions

Let $\boldsymbol{X} \in \mathbb{R}^{m \times n}$. $\boldsymbol{A}$ and $\boldsymbol{B}$ are constant matrices of compatible dimensions. Then

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{X}} \mathrm{tr}(\boldsymbol{A}\boldsymbol{X}) &= \boldsymbol{A}^T \\ \frac{\partial}{\partial \boldsymbol{X}} \mathrm{tr}(\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^T\boldsymbol{B}) &= \boldsymbol{B}^T\boldsymbol{X}\boldsymbol{A}^T + \boldsymbol{B}\boldsymbol{X}\boldsymbol{A} \\ \frac{\partial}{\partial \boldsymbol{X}} \mathrm{tr}(\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}\boldsymbol{B}) &= \boldsymbol{B}^T\boldsymbol{X}^T\boldsymbol{A}^T + \boldsymbol{A}^T\boldsymbol{X}^T\boldsymbol{B}^T. \end{aligned}$$

Proof: check elementwise.

Aitken theorem and generalized least squares (no constraints, non-singular $\boldsymbol{V}$)

- (Aitken theorem) Assume Aitken model ($\mathrm{E}(\boldsymbol{y}) = \boldsymbol{X}\boldsymbol{b}$, $\mathrm{Cov}(\boldsymbol{y}) = \sigma^2\boldsymbol{V}$ and $\boldsymbol{V} \succ \boldsymbol{0}_{n\times n}$. The best (minimum variance) affine unbiased estimator (MVAUE) of an estimable function $\boldsymbol{\Lambda}\boldsymbol{b}$ is

  $$\widehat{\boldsymbol{\Lambda}\boldsymbol{b}} = \boldsymbol{\Lambda}(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{y}$$

  with variance matrix

  $$\mathrm{Cov}(\boldsymbol{\Lambda}\boldsymbol{b}) = \sigma^2\boldsymbol{\Lambda}(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^-\boldsymbol{\Lambda}^T.$$

  Proof. Since we are told the answer. We can go ahead checking unbiasedness and minimum variance property directly. But let's proceed in a more constructive way. First derive the estimator as a minimum trace affine unbiased estimator (MTAUE) and then show that it is MVAUE.

  Let $\widehat{\boldsymbol{\Lambda}\boldsymbol{b}} = \boldsymbol{A}\boldsymbol{y} + \boldsymbol{c}$ be an affine estimator for $\boldsymbol{\Lambda}\boldsymbol{b}$. Unbiasedness requires

  $$\mathrm{E}(\widehat{\boldsymbol{\Lambda}\boldsymbol{b}}) = \boldsymbol{A}\boldsymbol{X}\boldsymbol{b} + \boldsymbol{c} = \boldsymbol{\Lambda}\boldsymbol{b} \text{ for all } \boldsymbol{b}.$$

  This implies $\boldsymbol{c} = \boldsymbol{0}$ and $\boldsymbol{A}\boldsymbol{X} = \boldsymbol{\Lambda}$. So $\widehat{\boldsymbol{\Lambda}\boldsymbol{b}}$ is linear and its variance is

  $$\mathrm{Cov}(\widehat{\boldsymbol{\Lambda}\boldsymbol{b}}) = \mathrm{Cov}(\boldsymbol{A}\boldsymbol{y}) = \sigma^2\boldsymbol{A}\boldsymbol{V}\boldsymbol{A}^T.$$

  To find a MTAUE, we consider the minimization problem

  $$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\mathrm{tr}(\boldsymbol{A}\boldsymbol{V}\boldsymbol{A}^T) \\ \text{subject to} \quad & \boldsymbol{A}\boldsymbol{X} = \boldsymbol{\Lambda}. \end{aligned}$$

The relevant Lagrangian function is

$$\psi(\boldsymbol{A}, \boldsymbol{L}) = \frac{1}{2}\mathrm{tr}(\boldsymbol{A}\boldsymbol{V}\boldsymbol{A}^T) - \mathrm{tr}(\boldsymbol{L}^T(\boldsymbol{A}\boldsymbol{X} - \boldsymbol{\Lambda}))$$

where $\boldsymbol{L} \in \mathbb{R}^{m\times p}$ is a matrix of Lagrange multipliers. Differentiating ψ with respect to $\boldsymbol{A}$ yields

$$\begin{aligned} \boldsymbol{A}\boldsymbol{V} &= \boldsymbol{L}\boldsymbol{X}^T \\ \boldsymbol{A}\boldsymbol{X} &= \boldsymbol{\Lambda}. \end{aligned}$$

From the first equation, $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{X}^T\boldsymbol{V}^{-1}$. Substitution into the second equation we have

$$\boldsymbol{L}\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X} = \boldsymbol{\Lambda}.$$

This system is always consistent (why? show $\mathcal{C}(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}) = \mathcal{C}(\boldsymbol{X}^T) \supset \mathcal{C}(\boldsymbol{\Lambda}^T)$ in HW4) so $\boldsymbol{\Lambda}(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^-$ is *a* solution to $\boldsymbol{L}$ (may not be unique). Therefore

$$\boldsymbol{A} = \boldsymbol{\Lambda}(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{V}^{-1}$$

is *a* solution to the constrained minimization problem. Actually it is invariant to the choice of generalized inverse $(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^-$ (HW4). Hence it is unique. Therefore the estimator

$$\widehat{\boldsymbol{\Lambda}\boldsymbol{b}} = \boldsymbol{\Lambda}(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{y}$$

with variance

$$\begin{aligned} \mathrm{Cov}(\widehat{\boldsymbol{\Lambda}\boldsymbol{b}}) &= \sigma^2\boldsymbol{\Lambda}(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{V}\boldsymbol{V}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^-\boldsymbol{\Lambda}^T \\ &= \sigma^2\boldsymbol{\Lambda}(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^-\boldsymbol{\Lambda}^T \end{aligned}$$

is the unique MTAUE. If MVAUE exists, this is it.

Now we have a good candidate, namely the MTAUE. It becomes a routine to check that it is actually the MVAUE. Consider an arbitrary affine unbiased estimator

$$[\boldsymbol{\Lambda}(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{V}^{-1} + \boldsymbol{C}]\boldsymbol{y} + \boldsymbol{d}.$$

Unbiasedness imposes $\boldsymbol{C}\boldsymbol{X} = \boldsymbol{0}_{m\times p}$ and $\boldsymbol{d} = \boldsymbol{0}_m$. Its covariance is

$$\begin{aligned} & \sigma^2[\boldsymbol{\Lambda}(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{V}^{-1} + \boldsymbol{C}]\boldsymbol{V}[\boldsymbol{\Lambda}(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{V}^{-1} + \boldsymbol{C}]^T \\ = \; & \sigma^2\boldsymbol{\Lambda}(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^-\boldsymbol{\Lambda}^T + \sigma^2\boldsymbol{C}\boldsymbol{V}\boldsymbol{C}^T \\ \succeq \; & \mathrm{Cov}(\widehat{\boldsymbol{\Lambda}\boldsymbol{b}}). \end{aligned}$$

Therefore the MTAUE is indeed the MVAUE. □

- $\widehat{\boldsymbol{\Lambda}\boldsymbol{b}} = \boldsymbol{\Lambda}(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{y}$ is called the *generalized least squares estimator* of an estimable function $\boldsymbol{\Lambda}\boldsymbol{b}$.
- The same good candidate can also be obtained by transforming the Aitken model back to the Gauss-Markov model (as done in the textbook).

Aitken model with linear constraints ($\boldsymbol{V} \succ \boldsymbol{0}_{n\times n}$)

- In many applications, researchers want to fit linear model with some constraints on the parameter $\boldsymbol{b}$. We investigate the estimability issue and try to find the MVAUE assuming that $\boldsymbol{V}$ is non-singular.
- (Estimability under linear constraints) Consider the Aitken model ($\mathrm{E}(\boldsymbol{y}) = \boldsymbol{X}\boldsymbol{b}$ and $\mathrm{Cov}(\boldsymbol{y}) = \sigma^2\boldsymbol{V}$) with linear constraints $\boldsymbol{R}\boldsymbol{b} = \boldsymbol{r}$. A linear function $\boldsymbol{\Lambda}\boldsymbol{b}$ is estimable if and only if $\mathcal{C}(\boldsymbol{\Lambda}^T) \subset \mathcal{C}((\boldsymbol{X}^T, \boldsymbol{R}^T))$.

  Remark 1: Apparently there are *more* estimable functions with constraints than without constraints.
  Remark 2: No assumption on the singularity of $\boldsymbol{V}$ is needed here.

  Proof. Aitken model with linear constraints is equivalent to the following expanded unconstrained Aitken model

$$\begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{r} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X} \\ \boldsymbol{R} \end{pmatrix} \boldsymbol{b} + \begin{pmatrix} \boldsymbol{e} \\ \boldsymbol{0} \end{pmatrix},$$

  where the error term has first two moments

$$\mathrm{E}\begin{pmatrix} \boldsymbol{e} \\ \boldsymbol{0} \end{pmatrix} = \boldsymbol{0}, \quad \mathrm{Cov}\begin{pmatrix} \boldsymbol{e} \\ \boldsymbol{0} \end{pmatrix} = \sigma^2 \begin{pmatrix} \boldsymbol{V} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix}.$$

  Thus the estimability result follows from that for this expanded Aitken model without constraints. □

- (MVAUE under linear constraints) Consider the Aitken model ($\mathrm{E}(\boldsymbol{y}) = \boldsymbol{X}\boldsymbol{b}$ and $\mathrm{Cov}(\boldsymbol{y}) = \sigma^2\boldsymbol{V}$) with linear constraints $\boldsymbol{R}\boldsymbol{b} = \boldsymbol{r}$ and assume $\boldsymbol{V} \succ \boldsymbol{0}_{n\times n}$. The best (minimum variance) affine unbiased estimator (MVAUE) of an estimable function $\boldsymbol{\Lambda}\boldsymbol{b}$ is

$$\widehat{\boldsymbol{\Lambda}\boldsymbol{b}} = \boldsymbol{\Lambda}\boldsymbol{G}^{-}\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{y} + \boldsymbol{\Lambda}\boldsymbol{G}^{-}\boldsymbol{R}^T(\boldsymbol{R}\boldsymbol{G}^{-}\boldsymbol{R}^T)^{-}(\boldsymbol{r} - \boldsymbol{R}\boldsymbol{G}^{-}\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{y})$$

where

$$\boldsymbol{G} = \boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X} + \boldsymbol{R}^T\boldsymbol{R}.$$

It has variance matrix

$$\mathrm{Cov}(\widehat{\boldsymbol{\Lambda}\boldsymbol{b}}) = \sigma^2\boldsymbol{\Lambda}\boldsymbol{G}^{-}\boldsymbol{\Lambda}^T - \sigma^2\boldsymbol{\Lambda}\boldsymbol{G}^{-}\boldsymbol{R}^T(\boldsymbol{R}\boldsymbol{G}^{-}\boldsymbol{R}^T)^{-}\boldsymbol{R}\boldsymbol{G}^{-}\boldsymbol{\Lambda}^T.$$

Remark 1: Notice that the MVAUE is indeed affine and not linear anymore under linear constraints.
Remark 2: Notice the variance is reduced in presence of constraints.

Proof. We again proceed in the constructive way: first find MTAUE and then show it is MVAUE.

Let $\widehat{\boldsymbol{\Lambda}\boldsymbol{b}} = \boldsymbol{A}\boldsymbol{y} + \boldsymbol{c}$ be an affine estimator of $\boldsymbol{\Lambda}\boldsymbol{b}$. Unbiasedness imposes

$$\mathrm{E}(\widehat{\boldsymbol{\Lambda}\boldsymbol{b}}) = \boldsymbol{A}\boldsymbol{X}\boldsymbol{b} + \boldsymbol{c} = \boldsymbol{\Lambda}\boldsymbol{b} \quad \text{for all } \boldsymbol{b} \text{ satisfying } \boldsymbol{R}\boldsymbol{b} = \boldsymbol{r}.$$

Substituting the general solution $\boldsymbol{b}^* = \boldsymbol{R}^{-}\boldsymbol{r} + (\boldsymbol{I} - \boldsymbol{R}^{-}\boldsymbol{R})\boldsymbol{q}$ shows

$$(\boldsymbol{\Lambda} - \boldsymbol{A}\boldsymbol{X})[\boldsymbol{R}^{-}\boldsymbol{r} + (\boldsymbol{I} - \boldsymbol{R}^{-}\boldsymbol{R})\boldsymbol{q}] = \boldsymbol{c} \quad \text{for all } \boldsymbol{q},$$

which implies

$$\begin{aligned} (\boldsymbol{\Lambda} - \boldsymbol{A}\boldsymbol{X})\boldsymbol{R}^{-}\boldsymbol{r} &= \boldsymbol{c} \\ (\boldsymbol{\Lambda} - \boldsymbol{A}\boldsymbol{X})(\boldsymbol{I} - \boldsymbol{R}^{-}\boldsymbol{R}) &= \boldsymbol{0}_{m\times p}. \end{aligned}$$

The second equation says $\mathcal{C}((\boldsymbol{\Lambda} - \boldsymbol{A}\boldsymbol{X})^T) \subset \mathcal{C}(\boldsymbol{R}^T)$. There exists a matrix $\boldsymbol{B}$ such that $\boldsymbol{\Lambda} - \boldsymbol{A}\boldsymbol{X} = \boldsymbol{B}\boldsymbol{R}$ and $\boldsymbol{c} = \boldsymbol{B}\boldsymbol{R}\boldsymbol{R}^{-}\boldsymbol{r} = \boldsymbol{B}\boldsymbol{r}$. Therefore the affine estimator takes the form

$$\widehat{\boldsymbol{\Lambda}\boldsymbol{b}} = \boldsymbol{A}\boldsymbol{y} + \boldsymbol{B}\boldsymbol{r}$$

and the unbiasedness condition boils down to $\boldsymbol{AX} + \boldsymbol{BR} = \boldsymbol{\Lambda}$.

To find a MTAUE, we consider the minimization problem

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}\text{tr}(\boldsymbol{AVA}^T) \\ \text{subject to} & \boldsymbol{AX} + \boldsymbol{BR} = \boldsymbol{\Lambda}. \end{array}$$

The relevant Lagrangian function is

$$\psi(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{L}) = \frac{1}{2}\text{tr}\boldsymbol{AVA}^T - \text{tr}\boldsymbol{L}^T(\boldsymbol{AX} + \boldsymbol{BR} - \boldsymbol{\Lambda}),$$

where $\boldsymbol{L}$ is a matrix of Lagrangian multipliers. Differentiating with respect to $\boldsymbol{A}$ and $\boldsymbol{B}$ gives the first order conditions

$$\begin{aligned} \boldsymbol{AV} &= \boldsymbol{LX}^T \\ \boldsymbol{RL}^T &= \boldsymbol{0} \\ \boldsymbol{AX} + \boldsymbol{BR} &= \boldsymbol{\Lambda}. \end{aligned}$$

Step 1: Express $\boldsymbol{L}$ in terms of $\boldsymbol{A}$. From the first equation, $\boldsymbol{A} = \boldsymbol{LX}^T\boldsymbol{V}^{-1}$. Thus

$$\boldsymbol{AX} = \boldsymbol{LX}^T\boldsymbol{V}^{-1}\boldsymbol{X} = \boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X} + \boldsymbol{R}^T\boldsymbol{R}) = \boldsymbol{LG}$$

by the second equation, where $\boldsymbol{G} = \boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X} + \boldsymbol{R}^T\boldsymbol{R}$ is p.s.d. It can be shown that $\mathcal{C}(\boldsymbol{G}) \supset \mathcal{C}(\boldsymbol{X}^T) \cup \mathcal{C}(\boldsymbol{R}^T)$ (HW4). Therefore $\boldsymbol{AX} = \boldsymbol{LG}$ as an equation in $\boldsymbol{L}$ is always consistent (since $\boldsymbol{GG}^-\boldsymbol{X}^T\boldsymbol{A}^T = \boldsymbol{X}^T\boldsymbol{A}^T$) and $\boldsymbol{AXG}^-$ is one solution to $\boldsymbol{L}$.

Step 2: Solve for $\boldsymbol{B}$. Post-multiply both sides of the third equation by $\boldsymbol{G}^-\boldsymbol{R}^T$ we get $\boldsymbol{BRG}^-\boldsymbol{R}^T = \boldsymbol{\Lambda G}^-\boldsymbol{R}^T$. As a linear equation in $\boldsymbol{B}$, it is always consistent if we can check

$$(\boldsymbol{RG}^-\boldsymbol{R}^T)(\boldsymbol{RG}^-\boldsymbol{R}^T)^-\boldsymbol{RG}^-\boldsymbol{\Lambda}^T = \boldsymbol{RG}^-\boldsymbol{\Lambda}^T.$$

We show this by writing $\boldsymbol{\Lambda}^T = (\boldsymbol{X}^T, \boldsymbol{R}^T)\boldsymbol{T}$ for some transformation matrix $\boldsymbol{T}$. Such $\boldsymbol{T}$ exists because of the estimability of $\boldsymbol{\Lambda b}$. Then it only remains to show that $\mathcal{C}(\boldsymbol{RG}^-\boldsymbol{R}^T) \supset \mathcal{C}(\boldsymbol{RG}^-\boldsymbol{X}^T)$ (HW4).

Therefore the linear system is always consistent and one specific solution to $\boldsymbol{B}$ is $\boldsymbol{\Lambda G}^-\boldsymbol{R}^T(\boldsymbol{RG}^-\boldsymbol{R}^T)^-$.

Step 3: Aseemble pieces to get the minimizing $\boldsymbol{A}$

$$\begin{aligned}
\boldsymbol{A} &= \boldsymbol{L}\boldsymbol{X}^T\boldsymbol{V}^{-1} = \boldsymbol{A}\boldsymbol{X}\boldsymbol{G}^-\boldsymbol{X}^T\boldsymbol{V}^{-1} \\
&= (\boldsymbol{\Lambda} - \boldsymbol{B}\boldsymbol{R})\boldsymbol{G}^-\boldsymbol{X}^T\boldsymbol{V}^{-1} \\
&= \boldsymbol{\Lambda}\boldsymbol{G}^-\boldsymbol{X}^T\boldsymbol{V}^{-1} - \boldsymbol{B}\boldsymbol{R}\boldsymbol{G}^-\boldsymbol{X}^T\boldsymbol{V}^{-1} \\
&= \boldsymbol{\Lambda}\boldsymbol{G}^-\boldsymbol{X}^T\boldsymbol{V}^{-1} - \boldsymbol{\Lambda}\boldsymbol{G}^-\boldsymbol{R}^T(\boldsymbol{R}\boldsymbol{G}^-\boldsymbol{R}^T)^-\boldsymbol{R}\boldsymbol{G}^-\boldsymbol{X}^T\boldsymbol{V}^{-1}.
\end{aligned}$$

Therefore the MTAUE is

$$\begin{aligned}
\widehat{\boldsymbol{\Lambda}\boldsymbol{b}} &= \boldsymbol{A}\boldsymbol{y} + \boldsymbol{B}\boldsymbol{r} \\
&= \boldsymbol{\Lambda}\boldsymbol{G}^-\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{y} + \boldsymbol{\Lambda}\boldsymbol{G}^-\boldsymbol{R}^T(\boldsymbol{R}\boldsymbol{G}^-\boldsymbol{R}^T)^-(\boldsymbol{r} - \boldsymbol{R}\boldsymbol{G}^-\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{y}),
\end{aligned}$$

which serves a good candidate for MVAUE. The last two (routine) steps are to derive the variance formula for this MTAUE and show that it is indeed the MVAUE. They are left as exercises in HW4. □

9 Lecture 9: Sep 30

Announcement

- HW2 returned (77 ± 15). For comparison: HW1 (89 ± 4).
- HW3 due this Wed Oct 2
- HW 3/4 recitation this Wednesday afternoon
- HW4 is posted and due Oct 9 (HW4 questions are covered in Midterm 1)
- Plan for this week: finish Aitken model and Midterm 1 review
- Please send your specific questions (HWs, lecture notes, textbook, ...) to me (`hua_zhou@ncsu.edu`) by this Saturday; so I can go over during classes (Oct 2 or Oct 7)

Last time

- Aitken model (no constraints, non-singular $\boldsymbol{V}$):
  MVAUE is $\widehat{\boldsymbol{\Lambda b}} = \boldsymbol{\Lambda}(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{y}$ with variance $\sigma^2\boldsymbol{\Lambda}(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^-\boldsymbol{\Lambda}^T$
- Aitken model (linear constraints $\boldsymbol{Rb} = \boldsymbol{r}$, non-singular $\boldsymbol{V}$):
  Estimability: $\mathcal{C}(\boldsymbol{\Lambda}^T) \subset \mathcal{C}((\boldsymbol{X}^T, \boldsymbol{R}^T))$, MVAUE (via MTAUE).

Today

- Aitken model: imposing constraints for a unique solution
- Aitken model with singular $\boldsymbol{V}$

Aitken model with linear constraints: special case $\mathcal{C}(\boldsymbol{R}^T) \cap \mathcal{C}(\boldsymbol{X}^T) = \{\boldsymbol{0}\}$

- As we saw, incorporating constraints increases the class the linear functions $\boldsymbol{\Lambda b}$ that are estimatible. When $\begin{pmatrix} \boldsymbol{X} \\ \boldsymbol{R} \end{pmatrix}$ has full column rank p, we can estimate every single parameter in $\boldsymbol{b}$!

- In many applications, we purposely add constraints which are linearly independent of rows of $\boldsymbol{X}$. That is $\mathcal{C}(\boldsymbol{R}^T) \cap \mathcal{C}(\boldsymbol{X}^T) = \{\boldsymbol{0}\}$. In this case, the MVAUE takes a simpler form.

- Estimability: $\mathcal{C}(\boldsymbol{\Lambda}^T) \subset \mathcal{C}((\boldsymbol{X}^T, \boldsymbol{R}^T))$. Note $\text{rank}((\boldsymbol{X}^T, \boldsymbol{R}^T)) = \text{rank}(\boldsymbol{X}^T) + \text{rank}(\boldsymbol{R}^T)$ when $\mathcal{C}(\boldsymbol{R}^T) \cap \mathcal{C}(\boldsymbol{X}^T) = \{\boldsymbol{0}\}$.

- For example, in one-way ANOVA

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{1}_{n_1} & \boldsymbol{1}_{n_1} & & & \\ \boldsymbol{1}_{n_2} & & \boldsymbol{1}_{n_2} & & \\ \vdots & & & \ddots & \\ \boldsymbol{1}_{n_a} & & & & \boldsymbol{1}_{n_a} \end{pmatrix},$$

  where $\text{rank}(\boldsymbol{X}) = a$. Three commonly used constraints are

  1. $\alpha_a = 0$
  2. $\sum_{i=1}^a \alpha_i = 0$
  3. $\sum_{i=1}^a n_i \alpha_i = 0$.

  Each of them will make every single parameter in $(\mu, \alpha_1, \ldots, \alpha_a)$ estimable.

- (MVAUE) Consider the Aitken model ($\text{E}(\boldsymbol{y}) = \boldsymbol{X}\boldsymbol{b}$ and $\text{Cov}(\boldsymbol{y}) = \sigma^2\boldsymbol{V}$) with linear constraints $\boldsymbol{R}\boldsymbol{b} = \boldsymbol{r}$ where $\mathcal{C}(\boldsymbol{R}^T) \cap \mathcal{C}(\boldsymbol{X}^T) = \{\boldsymbol{0}\}$. Assume $\boldsymbol{V} \succ \boldsymbol{0}_{n\times n}$. The best (minimum variance) affine unbiased estimator (MVAUE) of an estimable function $\boldsymbol{\Lambda}\boldsymbol{b}$ is

$$\widehat{\boldsymbol{\Lambda}\boldsymbol{b}} = \boldsymbol{\Lambda}\boldsymbol{G}^-(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{y} + \boldsymbol{R}^T\boldsymbol{r}),$$

  where

$$\boldsymbol{G} = \boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X} + \boldsymbol{R}^T\boldsymbol{R}.$$

  It has variance matrix

$$\text{Cov}(\widehat{\boldsymbol{\Lambda}\boldsymbol{b}}) = \sigma^2\boldsymbol{\Lambda}\boldsymbol{G}^-\boldsymbol{\Lambda}^T - \sigma^2\boldsymbol{\Lambda}\boldsymbol{G}^-\boldsymbol{R}^T\boldsymbol{R}\boldsymbol{G}^-\boldsymbol{\Lambda}^T.$$

  Remark: note $\mathcal{C}(\boldsymbol{R}^T) \cap \mathcal{C}(\boldsymbol{X}^T) = \{\boldsymbol{0}\}$ is equivalent to $\text{rank}(\boldsymbol{X}^T, \boldsymbol{R}^T) = \text{rank}(\boldsymbol{X}^T) + \text{rank}(\boldsymbol{R}^T)$.

  Proof. Do HW4 Q3 and you will find some cancellations occur in the previous more general result. □

Aitken model with linear constraints and a singular $\boldsymbol{V}$

Now we are in a position to consider the most general case: Aiken model with linear constraints $\boldsymbol{Rb} = \boldsymbol{r}$ and a possibly singular $\boldsymbol{V}$.

- First we notice that, with a singular $\boldsymbol{V}$, not all values of $\boldsymbol{y}$ are possible. For example, $y_i = \mu + e_i$, $i = 1, 2, 3, 4$, where $\mathrm{Var}(e_1) = \mathrm{Var}(e_2) = 0$ and $\mathrm{Var}(e_i) > 0$ for $i = 3, 4$, is a linear model with $\mathrm{rank}(\boldsymbol{V}) = 2$. Any $\boldsymbol{y}$ with $y_1 \neq y_2$ is inconsistent with this model.

- We can write the constrained linear model as an unconstrained one

$$\boldsymbol{y}_e = \boldsymbol{X}_e\boldsymbol{b} + \boldsymbol{u}, \quad \mathrm{E}(\boldsymbol{u}) = \boldsymbol{0}, \quad , \mathrm{Cov}(\boldsymbol{u}) = \sigma^2\boldsymbol{V}_e,$$

  where

$$\boldsymbol{y}_e = \begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{r} \end{pmatrix}, \quad \boldsymbol{X}_e = \begin{pmatrix} \boldsymbol{X} \\ \boldsymbol{R} \end{pmatrix}, \quad \boldsymbol{V}_e = \begin{pmatrix} \boldsymbol{V} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix}.$$

- (Consistency) Aitken model ($\mathrm{E}(\boldsymbol{y}) = \boldsymbol{Xb}$, $\mathrm{Cov}(\boldsymbol{y}) = \sigma^2\boldsymbol{V}$) with linear constraints $\boldsymbol{Rb} = \boldsymbol{r}$ is consistent if and only if

$$\begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{r} \end{pmatrix} \in \mathcal{C}\begin{pmatrix} \boldsymbol{X} & \boldsymbol{V} \\ \boldsymbol{R} & \boldsymbol{0} \end{pmatrix}.$$

  Proof. Note the model $\boldsymbol{y}_e = \boldsymbol{X}_e\boldsymbol{b} + \boldsymbol{u}$ is equivalent to the model $\boldsymbol{y}_e = \boldsymbol{X}_e\boldsymbol{b} + \boldsymbol{V}_e\boldsymbol{V}_e^-\boldsymbol{u}$ since

$$\mathrm{E}(\boldsymbol{u}) = \boldsymbol{V}_e\boldsymbol{V}_e^-\mathrm{E}(\boldsymbol{u}) = \boldsymbol{0}_n, \quad \mathrm{Cov}(\boldsymbol{u}) = \sigma^2\boldsymbol{V}_e\boldsymbol{V}_e^-\boldsymbol{V}_e\boldsymbol{V}_e^-\boldsymbol{V}_e = \sigma^2\boldsymbol{V}_e.$$

  Therefore the linear model is consistent if and only if $\boldsymbol{y}_e \in \mathcal{C}((\boldsymbol{X}_e, \boldsymbol{V}_e\boldsymbol{V}_e^-)) = \mathcal{C}((\boldsymbol{X}_e, \boldsymbol{V}_e)) = \mathcal{C}\left(\begin{pmatrix} \boldsymbol{X} & \boldsymbol{V} \\ \boldsymbol{R} & \boldsymbol{0} \end{pmatrix}\right)$. □

- (Estimability) $\boldsymbol{\Lambda b}$ is estimable if and only if

$$\mathcal{C}(\boldsymbol{\Lambda}^T) \subset \mathcal{C}((\boldsymbol{X}^T, \boldsymbol{R}^T)).$$

- (MVAUE) Consider the Aitken model ($\mathrm{E}(\boldsymbol{y}) = \boldsymbol{X}\boldsymbol{b}$, $\mathrm{Cov}(\boldsymbol{y}) = \sigma^2\boldsymbol{V}$) with linear constraints $\boldsymbol{R}\boldsymbol{b} = \boldsymbol{r}$. The best (minimum) variance affine unbiased estimator (MVAUE) of an estimable function $\boldsymbol{\Lambda}\boldsymbol{b}$ is

$$\widehat{\boldsymbol{\Lambda}\boldsymbol{b}} = \boldsymbol{\Lambda}(\boldsymbol{X}_e^T\boldsymbol{V}_0^-\boldsymbol{X}_e)^-\boldsymbol{X}_e^T\boldsymbol{V}_0^-\boldsymbol{y}_e,$$

where $\boldsymbol{V}_0 = \boldsymbol{V}_e + \boldsymbol{X}_e\boldsymbol{X}_e^T$, with variance

$$\mathrm{Cov}(\widehat{\boldsymbol{\Lambda}\boldsymbol{b}}) = \sigma^2\boldsymbol{\Lambda}[(\boldsymbol{X}_e^T\boldsymbol{V}_0^-\boldsymbol{X}_e)^- - \boldsymbol{I}_p]\boldsymbol{\Lambda}^T.$$

Remark: Since $\boldsymbol{y}_e$ involves both $\boldsymbol{y}$ and $\boldsymbol{r}$, the MVAUE is affine in general.

Proof. Let's try to derive the MTAUE first, and then show that the MTAUE is MVAUE. Let $\boldsymbol{A}\boldsymbol{y}_e + \boldsymbol{c}$ be an affine unbiased estimator of an estimable function $\boldsymbol{\Lambda}\boldsymbol{b}$. Unbiasedness requires

$$\boldsymbol{A}\boldsymbol{X}_e\boldsymbol{b} + \boldsymbol{c} = \boldsymbol{\Lambda}\boldsymbol{b} \quad \text{for all } \boldsymbol{b} \in \mathbb{R}^p,$$

which implies

$$\boldsymbol{A}\boldsymbol{X}_e = \boldsymbol{\Lambda} \text{ and } \boldsymbol{c} = \boldsymbol{0}.$$

We seek the MTAUE by solving

$$\begin{array}{rl} \text{minimize} & \frac{1}{2}\mathrm{tr}\boldsymbol{A}\boldsymbol{V}_e\boldsymbol{A}^T \\ \text{subject to} & \boldsymbol{A}\boldsymbol{X}_e = \boldsymbol{\Lambda}. \end{array}$$

The relevant Lagrangian is

$$\psi(\boldsymbol{A}, \boldsymbol{L}) = \frac{1}{2}\mathrm{tr}\boldsymbol{A}\boldsymbol{V}_e\boldsymbol{A}^T - \mathrm{tr}(\boldsymbol{L}^T(\boldsymbol{A}\boldsymbol{X}_e - \boldsymbol{\Lambda})),$$

where $\boldsymbol{L}$ is a matrix of Lagrangian multipliers. Differentiating with respect to $\boldsymbol{A}$ yields the first order conditions

$$\begin{aligned} \boldsymbol{A}\boldsymbol{V}_e &= \boldsymbol{L}\boldsymbol{X}_e^T \\ \boldsymbol{A}\boldsymbol{X}_e &= \boldsymbol{\Lambda}, \end{aligned}$$

or, as a matrix equation,

$$\begin{pmatrix} \boldsymbol{V}_e & \boldsymbol{X}_e \\ \boldsymbol{X}_e^T & \boldsymbol{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{A}^T \\ -\boldsymbol{L}^T \end{pmatrix} = \begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{\Lambda}^T \end{pmatrix}.$$

By HW4 Q5(d), a generalized inverse of the matrix on left side is

$$\begin{pmatrix} \boldsymbol{V}_e & \boldsymbol{X}_e \\ \boldsymbol{X}_e^T & \boldsymbol{0} \end{pmatrix}^- = \begin{pmatrix} \boldsymbol{V}_0^- - \boldsymbol{V}_0^- \boldsymbol{X}_e \boldsymbol{C}^- \boldsymbol{X}_e^T \boldsymbol{V}_0^- & \boldsymbol{V}_0^- \boldsymbol{X}_e \boldsymbol{C}^- \\ \boldsymbol{C}^- \boldsymbol{X}_e^T \boldsymbol{V}_0^- & -\boldsymbol{C}^- + \boldsymbol{C}\boldsymbol{C}^- \end{pmatrix},$$

where $\boldsymbol{V}_0 = \boldsymbol{V}_e + \boldsymbol{X}_e \boldsymbol{X}_e^T$ and $\boldsymbol{C} = \boldsymbol{X}_e^T \boldsymbol{V}_0^- \boldsymbol{X}_e$, and

$$\begin{pmatrix} \boldsymbol{V}_e & \boldsymbol{X}_e \\ \boldsymbol{X}_e^T & \boldsymbol{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{V}_e & \boldsymbol{X}_e \\ \boldsymbol{X}_e^T & \boldsymbol{0} \end{pmatrix}^- = \begin{pmatrix} \boldsymbol{V}_0 \boldsymbol{V}_0^- & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{C}\boldsymbol{C}^- \end{pmatrix}.$$

By estimability, $\mathcal{C}(\boldsymbol{\Lambda}^T) \subset \mathcal{C}(\boldsymbol{X}_e^T) = \mathcal{C}(\boldsymbol{C})$ (In HW4 Q5(c), we also show $\mathcal{C}(\boldsymbol{X}_e^T \boldsymbol{V}_0^- \boldsymbol{X}_e) = \mathcal{C}(\boldsymbol{X}_e^T)$) and thus $\boldsymbol{C}\boldsymbol{C}^- \boldsymbol{\Lambda}^T = \boldsymbol{\Lambda}^T$. Therefore the linear system is consistent and the solution to $\boldsymbol{A}$ takes the general form

$$\boldsymbol{A} = \boldsymbol{\Lambda}(\boldsymbol{X}_e^T \boldsymbol{V}_0^- \boldsymbol{X}_e)^- \boldsymbol{X}_e^T \boldsymbol{V}_0^- + \boldsymbol{Q}(\boldsymbol{I} - \boldsymbol{V}_0 \boldsymbol{V}_0^-)$$

where $\boldsymbol{Q}$ is arbitrary. Therefore the MTAUE is

$$\begin{aligned} \widehat{\boldsymbol{\Lambda b}} &= \boldsymbol{A}\boldsymbol{y}_e \\ &= \boldsymbol{\Lambda}(\boldsymbol{X}_e^T \boldsymbol{V}_0^- \boldsymbol{X}_e)^- \boldsymbol{X}_e^T \boldsymbol{V}_0^- \boldsymbol{y}_e + \boldsymbol{Q}(\boldsymbol{I} - \boldsymbol{V}_0 \boldsymbol{V}_0^-)\boldsymbol{y}_e \\ &= \boldsymbol{\Lambda}(\boldsymbol{X}_e^T \boldsymbol{V}_0^- \boldsymbol{X}_e)^- \boldsymbol{X}_e^T \boldsymbol{V}_0^- \boldsymbol{y}_e. \end{aligned}$$

The last equality follows from $\mathcal{C}(\boldsymbol{V}_0) = \mathcal{C}(\boldsymbol{V}_e + \boldsymbol{X}_e \boldsymbol{X}_e^T) = \mathcal{C}((\boldsymbol{X}_e, \boldsymbol{V}_e))$ (HW4 Q5(a)) and the consistency $\boldsymbol{y}_e \in \mathcal{C}((\boldsymbol{X}_e, \boldsymbol{V}_e))$.

The variance of the MTAUE is

$$\begin{aligned} \mathrm{Cov}(\widehat{\boldsymbol{\Lambda b}}) &= \sigma^2 \boldsymbol{\Lambda}(\boldsymbol{X}_e^T \boldsymbol{V}_0^- \boldsymbol{X}_e)^- \boldsymbol{X}_e^T \boldsymbol{V}_0^- \boldsymbol{V}_e \boldsymbol{V}_0^- \boldsymbol{X}_e (\boldsymbol{X}_e^T \boldsymbol{V}_0^- \boldsymbol{X}_e)^- \boldsymbol{\Lambda}^T \\ &= \sigma^2 \boldsymbol{\Lambda}(\boldsymbol{X}_e^T \boldsymbol{V}_0^- \boldsymbol{X}_e)^- \boldsymbol{X}_e^T \boldsymbol{V}_0^- (\boldsymbol{V}_0 - \boldsymbol{X}_e \boldsymbol{X}_e^T) \boldsymbol{V}_0^- \boldsymbol{X}_e (\boldsymbol{X}_e^T \boldsymbol{V}_0^- \boldsymbol{X}_e)^- \boldsymbol{\Lambda}^T \\ &= \sigma^2 \boldsymbol{\Lambda}(\boldsymbol{X}_e^T \boldsymbol{V}_0^- \boldsymbol{X}_e)^- \boldsymbol{\Lambda}^T - \sigma^2 \boldsymbol{\Lambda}\boldsymbol{\Lambda}^T \\ &= \sigma^2 \boldsymbol{\Lambda}[(\boldsymbol{X}_e^T \boldsymbol{V}_0^- \boldsymbol{X}_e)^- - \boldsymbol{I}]\boldsymbol{\Lambda}^T. \end{aligned}$$

Last order of business is to show that the MTAUE is indeed MVAUE. Here is another useful device for proving this. If $\boldsymbol{\Lambda b}$ is estimable, then $\boldsymbol{c}^T \boldsymbol{\Lambda b}$ is estimable (why?) and $\boldsymbol{c}^T \widehat{\boldsymbol{\Lambda b}}$ is the MTAUE for estimating $\boldsymbol{c}^T \boldsymbol{\Lambda b}$. Let $\hat{\boldsymbol{\theta}}$ be another affine unbiased estimator of $\boldsymbol{\Lambda b}$. Then $\boldsymbol{c}^T \hat{\boldsymbol{\theta}}$ is an affine unbiased estimator of $\boldsymbol{c}^T \boldsymbol{\Lambda b}$. Thus

$$\mathrm{Var}(\boldsymbol{c}^T \widehat{\boldsymbol{\Lambda b}}) \leq \mathrm{Var}(\boldsymbol{c}^T \hat{\boldsymbol{\theta}}),$$

which implies that

$$\boldsymbol{c}^T \mathrm{Cov}(\widehat{\boldsymbol{\Lambda b}})\boldsymbol{c}^T \leq \boldsymbol{c}^T \mathrm{Cov}(\hat{\boldsymbol{\theta}})\boldsymbol{c}.$$

Since $\boldsymbol{c}$ is arbitrary, we have shown that $\mathrm{Cov}(\widehat{\boldsymbol{\Lambda b}}) \preceq \mathrm{Cov}(\hat{\boldsymbol{\theta}})$. □

- (Conflict with previous result?) The special case $\boldsymbol{V} \succ \mathbf{0}_{n\times n}$ does not recover our previous result (Aitken with linear constraints and non-singular $\boldsymbol{V}$) obviously. We will take an alternative approach below – transforming a singular Aitken model to a non-singular one.

  Before that we first review some basic results on determinant, eigenvalues and eigenvectors.

10 Lecture 10: Oct 2

Announcement

- HW3 due today
- HW 3/4 recitation this afternoon
- HW4 is due next Wed Oct 9 (covered in Midterm 1)
- Please send your specific questions (HWs, lecture notes, textbook, ...) to me (`hua_zhou@ncsu.edu`) by this Saturday

Last time

- Aitken model: imposing constraints for a unique solution
- Aitken model with linear constraints $\boldsymbol{Rb} = \boldsymbol{r}$ and a possibly singular $\boldsymbol{V}$

Today

- Generalized least squares

A result on the invariance to the choice of generalized inverse

We often need to show that a product $\boldsymbol{BA}^-\boldsymbol{C}$ is invariant to the choice of the generalized inverse. Let's summarize a useful result here.

- If $\mathcal{C}(\boldsymbol{B}^T) \subset \mathcal{C}(\boldsymbol{A}^T)$ and $\mathcal{C}(\boldsymbol{C}) \subset \mathcal{C}(\boldsymbol{A})$, then the product $\boldsymbol{BA}^-\boldsymbol{C}$ is invariant to the choice of the generalized inverse. In other words, it is unique.

  Proof. Since $\mathcal{C}(\boldsymbol{B}^T) \subset \mathcal{C}(\boldsymbol{A}^T)$, there exists a matrix $\boldsymbol{L}$ such that $\boldsymbol{B} = \boldsymbol{LA}$. Since $\mathcal{C}(\boldsymbol{C}) \subset \mathcal{C}(\boldsymbol{A})$, there exists a matrix $\boldsymbol{R}$ such that $\boldsymbol{C} = \boldsymbol{AR}$. Then

  $$\boldsymbol{BA}^-\boldsymbol{C} = \boldsymbol{LAA}^-\boldsymbol{AR} = \boldsymbol{LAR}.$$

  It is also invariant to the choice of the transformation matrix since $\boldsymbol{L}_1\boldsymbol{AR} - \boldsymbol{L}_2\boldsymbol{AR} = \boldsymbol{BR} - \boldsymbol{BR} = \boldsymbol{0}$ for any two candidate transformation matrices $\boldsymbol{L}_1$ and $\boldsymbol{L}_2$. Similarly it is invariant to the choice of transformation matrix $\boldsymbol{R}$ too. □

- Applications: uniqueness of $\boldsymbol{P}_{\boldsymbol{X}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T$, uniqueness of all MVAUEs we derived so far, uniqueness of the variance of all MVAUEs we derived so far, ...

Generalized least squares

- Gauss invented method of least squares as an approximation method. The least squares solution (magically) turns out to be MVAUE.

- For Aitken model, we approached MVAUE through MTAUE. Are the MVAUEs solutions to some generalized least squares criterion? The answer is affirmative. This coincidence is surprising but not trivial.

- (Generalized least squares for Aitken with linear constraints) Consider the Aitken model, $\mathrm{E}(\boldsymbol{y}) = \boldsymbol{X}\boldsymbol{b}$ and $\mathrm{Cov}(\boldsymbol{y}) = \sigma^2\boldsymbol{V}$, with a possibly singular $\boldsymbol{V}$ and linear constraints $\boldsymbol{R}\boldsymbol{b} = \boldsymbol{r}$. Assume

$$\begin{pmatrix}\boldsymbol{y}\\ \boldsymbol{r}\end{pmatrix} \in \mathcal{C}\left(\begin{pmatrix}\boldsymbol{X} & \boldsymbol{V}\\ \boldsymbol{R} & \boldsymbol{0}\end{pmatrix}\right) \quad \text{(consistency)}$$

and $\mathcal{C}(\boldsymbol{\Lambda}^T) \subset \mathcal{C}((\boldsymbol{X}^T, \boldsymbol{R}^T))$ (estimability). Then the best (minimum) variance affine unbiased estimator (MVAUE) of $\boldsymbol{\Lambda}\boldsymbol{b}$ is $\boldsymbol{\Lambda}\hat{\boldsymbol{b}}$, where $\hat{\boldsymbol{b}}$ minimizes the generalized least squares criterion

$$\begin{pmatrix}\boldsymbol{y}-\boldsymbol{X}\boldsymbol{b}\\ \boldsymbol{r}-\boldsymbol{R}\boldsymbol{b}\end{pmatrix}^T \begin{pmatrix}\boldsymbol{V}+\boldsymbol{X}\boldsymbol{X}^T & \boldsymbol{X}\boldsymbol{R}^T\\ \boldsymbol{R}\boldsymbol{X}^T & \boldsymbol{R}\boldsymbol{R}^T\end{pmatrix}^- \begin{pmatrix}\boldsymbol{y}-\boldsymbol{X}\boldsymbol{b}\\ \boldsymbol{r}-\boldsymbol{R}\boldsymbol{b}\end{pmatrix}.$$

Remark 1: The generalized least squares solution $\hat{\boldsymbol{b}}$ may not be unique.
Remark 2: One implication of this connection is that we can obtain the MVAUE by solving the generalized least squares problem. There are good algorithms for minimizing a generalized least squares criterion.

Proof. Consider the augmented linear model $\mathrm{E}(\boldsymbol{y}_e) = \boldsymbol{X}_e\boldsymbol{b}$, $\mathrm{Cov}(\boldsymbol{y}) = \sigma^2\boldsymbol{V}_e$ where

$$\boldsymbol{y}_e = \begin{pmatrix}\boldsymbol{y}\\ \boldsymbol{r}\end{pmatrix}, \quad \boldsymbol{X}_e = \begin{pmatrix}\boldsymbol{X}\\ \boldsymbol{R}\end{pmatrix}, \quad \boldsymbol{V}_e = \begin{pmatrix}\boldsymbol{V} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}\end{pmatrix}.$$

Then generalized least squares criterion is

$$(\boldsymbol{y}_e - \boldsymbol{X}_e\boldsymbol{b})^T\boldsymbol{V}_0^-(\boldsymbol{y}_e - \boldsymbol{X}_e\boldsymbol{b}),$$

where $\boldsymbol{V}_0 = \boldsymbol{V}_e + \boldsymbol{X}_e\boldsymbol{X}_e^T$. We first show that this expression in invariant to the choice of the generalized inverse $\boldsymbol{V}_0^-$. This is true because $\boldsymbol{y}_e - \boldsymbol{X}_e\boldsymbol{b} \in \mathcal{C}((\boldsymbol{X}_e, \boldsymbol{V}_e)) = \mathcal{C}(\boldsymbol{V}_0)$ by consistency. Therefore without loss of generality we can use the Moore-Penrose inverse $\boldsymbol{V}_0^+$, which is symmetric and positive semidefinite, in the criterion.

Let $\boldsymbol{V}_0^+ = \boldsymbol{L}\boldsymbol{L}^T$, where $\boldsymbol{L}$ can be the Cholesky factor or the symmetric square root. Then the generalized least squares criterion becomes

$$\|\boldsymbol{L}^T(\boldsymbol{y}_e - \boldsymbol{X}_e\boldsymbol{b})\|_2^2.$$

The least squares solution (may be non-unique) to this regular least squares problem is

$$\hat{\boldsymbol{b}} = (\boldsymbol{X}_e^T\boldsymbol{L}\boldsymbol{L}^T\boldsymbol{X}_e)^-\boldsymbol{X}_e^T\boldsymbol{L}\boldsymbol{L}^T\boldsymbol{y} = (\boldsymbol{X}_e^T\boldsymbol{V}_0^+\boldsymbol{X}_e)^-\boldsymbol{X}_e^T\boldsymbol{V}_0^+\boldsymbol{y}_e.$$

The final step is to show that

$$\boldsymbol{\Lambda}\hat{\boldsymbol{b}} = \boldsymbol{\Lambda}(\boldsymbol{X}_e^T\boldsymbol{V}_0^+\boldsymbol{X}_e)^-\boldsymbol{X}_e^T\boldsymbol{V}_0^+\boldsymbol{y}_e$$

is invariant to the choice of the generalized inverse $(\boldsymbol{X}_e^T\boldsymbol{V}_0^+\boldsymbol{X}_e)^-$ and we can replace $\boldsymbol{V}_0^+$ by any generalized inverse $\boldsymbol{V}_0^-$. We use the result summarized in the previous section.

- To show that we can replace the Moore-Penrose inverse $\boldsymbol{V}_0^+$ in $\boldsymbol{X}_e^T\boldsymbol{V}_0^+\boldsymbol{X}_e$ by any generalized inverse $\boldsymbol{V}_0^-$, we note $\mathcal{C}(\boldsymbol{V}_0) = \mathcal{C}((\boldsymbol{X}_e, \boldsymbol{V}_e)) \supset \mathcal{C}(\boldsymbol{X}_e)$.
- To show that we can replace the Moore-Penrose inverse $\boldsymbol{V}_0^+$ in $\boldsymbol{X}_e^T\boldsymbol{V}_0^+\boldsymbol{y}_e$ by any generalized inverse $\boldsymbol{V}_0^-$, we note $\mathcal{C}(\boldsymbol{V}_0) \supset \mathcal{C}(\boldsymbol{X}_e)$ and $\boldsymbol{y}_e \in \mathcal{C}((\boldsymbol{X}_e, \boldsymbol{V}_e)) = \mathcal{C}(\boldsymbol{V}_0)$ by consistency.
- To show that $\boldsymbol{\Lambda}\hat{\boldsymbol{b}}$ is invariant to the choice of $(\boldsymbol{X}_e^T\boldsymbol{V}_0^+\boldsymbol{X}_e)^-$, we note

$$\mathcal{C}(\boldsymbol{X}_e^T\boldsymbol{V}_0^+\boldsymbol{X}_e) = \mathcal{C}(\boldsymbol{X}_e^T) \supset \mathcal{C}(\boldsymbol{\Lambda}^T).$$

Thus we have shown that $\boldsymbol{\Lambda}\hat{\boldsymbol{b}}$ coincides with the MVAUE we derived earlier. □

- In absence of constraints, the generalized least squares criterion reduces to

$$(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b})^T(\boldsymbol{V} + \boldsymbol{X}\boldsymbol{X}^T)^-(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}).$$

11 Lecture 11: Oct 7

Announcement

- HW3 returned (92 ± 11)
- HW4 deadline extended to next Wed Oct 16
- Midterm 1, Oct 9 @ 11:45AM-1PM
- No afternoon session this Wed Oct 9
- No office hours today (instructor out for conference).
  Makeup office hours: Tue Oct 8 @ 1-3PM
- No office hours Wed Oct 9
- No class and office hours next Monday Oct 14 (instructor out of town)

Last time

- Generalized least squares

Today

- Q&A
- HW4

Q&A

- After all the lectures, I am getting confused and mixed up with some definitions. I wonder if you can give us a summary of what are the ways to get (BLUE, invertible, estimable, consistency, MTAUE, MVAUE, etc.).
- HW5 question1 (b), and question 3.
- I made a mistake in HW2 about writing nested design model (last question from book). I wonder if there are more difficult models we should know how to write out. I just cannot come up with other examples.

- When we prove the unbiasedness of the $\boldsymbol{\Lambda}\hat{\boldsymbol{b}}$, we plug in the least square solution of $\boldsymbol{b}$, then we have $\boldsymbol{\Lambda}(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{y}+\boldsymbol{\Lambda}[\boldsymbol{I}-(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{X}]\boldsymbol{q}$, then the second term is vanished. My thought is: $[\boldsymbol{I}-(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{X}]$ is projection onto null space of $\boldsymbol{X}^T\boldsymbol{X}$. Because $\mathcal{N}(\boldsymbol{X}^T\boldsymbol{X})=\mathcal{N}(\boldsymbol{X}^T)$, and because $\boldsymbol{\Lambda}\boldsymbol{b}$ is estimable, so $\mathcal{C}(\boldsymbol{\Lambda}^T)\subset\mathcal{C}(\boldsymbol{X}^T)\Leftrightarrow\mathcal{N}(\boldsymbol{X}^T)\subset\mathcal{N}(\boldsymbol{\Lambda}^T)$. My problem is: we can take transpose, which is $[[\boldsymbol{I}-(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{X}]\boldsymbol{\Lambda}^T]^T$, but we only have $\mathcal{N}(\boldsymbol{X}^T)\subset\mathcal{N}(\boldsymbol{\Lambda}^T)$, I cannot see projection onto $\mathcal{N}(\boldsymbol{X}^T)$ is also projection onto $\mathcal{N}(\boldsymbol{\Lambda}^T)$? (since $\mathcal{N}(\boldsymbol{\Lambda}^T)$ contains $\mathcal{N}(\boldsymbol{X}^T)$)

- Can you explain the jointly non-estimable function. And why the number of them is fixed, $p-r$?

- Reparameterization for one-way ANOVA, since we can reparemeterize X to be full rank, I am thinking if that matrix is equivalent to one-way ANOVA without interception? (maybe this is trivial)

- For the overfitting part, we have $\mathrm{Cov}(\hat{\boldsymbol{b}})=\mathrm{Cov}((\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{y})$, which is

$$\sigma^2(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{X}((\boldsymbol{X}^T\boldsymbol{X})^-)^T.$$

  my question is how do we come up with $\sigma^2(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{X})((\boldsymbol{X}^T\boldsymbol{X})^-)^T=\sigma^2(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{X})(\boldsymbol{X}^T\boldsymbol{X})^-=\sigma^2(\boldsymbol{X}^T\boldsymbol{X})^-$? Is that because the projection. It is like the property of $\boldsymbol{G}^+$ inverse?

- For the Lagrangian multiplier in linearly constrained least squares solution, what is that constrained do or mean? $\boldsymbol{V}\boldsymbol{\beta}=\boldsymbol{d}$ what is $\boldsymbol{d}$?

- How does that come out for the consistency of Aitken model with linear constraints?

- Conflict, yes, I got confused about the Aitken model with linear constraints and non-singular V / Aitken model with linear constraints and singular V. Why the first MVAUE is non linear , the second is linear?

- Can you go over how to prove column spaces are equal using the null space property?

- Homework 1: 2 (d).

- Homework 2: 1 (JM 2.11) (f), (g) and (h);

12 Lecture 12: Oct 16

Announcement

- HW4 due today.
- HW5 will be posted today and due next Fri Oct 25 (?).
- Midterm 1 returned (80 ± 14).

Today

- Finish Chapter 4: Aitken model with linear constraints $\boldsymbol{Rb} = \boldsymbol{r}$ and a possibly singular $\boldsymbol{V}$ – alternative approach
- Chapter 5: normal and related distributions

Review: determinant (JM A.6)

Assume $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ is a square matrix.

- The *determinant* of a matrix $\det(\boldsymbol{A}) = |\boldsymbol{A}| = \sum(-1)^{\phi(j_1,\ldots,j_n)} \prod_{i=1}^{n} a_{ij_i}$, where $\phi(j_1, \ldots, j_n)$ is the number of transpositions to change $(1, \ldots, n)$ to $(j_1, \ldots, j_n)$.
- Example,

$$\det \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} = a_{11}a_{22} - a_{12}a_{21}.$$

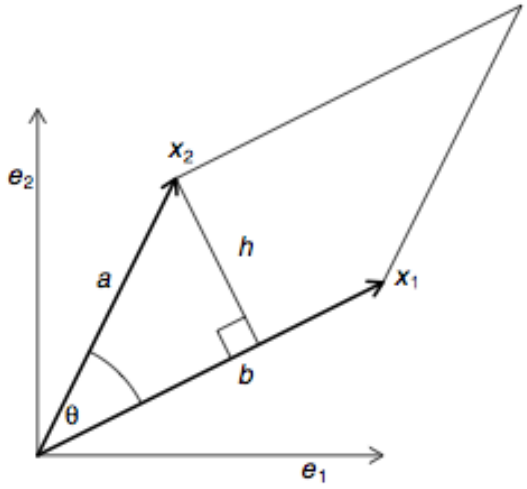


Fig. 3.1. Volume (Area) of Region Determined by x_1 and x_2

- Interpretation of the (absolute value of) determinant as the volume of the parallelogram defined by the columns of the matrix:

$$\begin{aligned}
\text{area} &= bh = \|\boldsymbol{x}_1\|\|\boldsymbol{x}_2\|\sin(\theta) \\
&= \|\boldsymbol{x}_1\|\|\boldsymbol{x}_2\|\sqrt{1-\left(\frac{\langle \boldsymbol{x}_1, \boldsymbol{x}_2\rangle}{\|\boldsymbol{x}_1\|\|\boldsymbol{x}_2\|}\right)^2} \\
&= \sqrt{\|\boldsymbol{x}_1\|^2\|\boldsymbol{x}\|^2 - (\langle \boldsymbol{x}_1, \boldsymbol{x}_2\rangle)^2} \\
&= \sqrt{(x_{11}^2+x_{12}^2)(x_{21}^2+x_{22}^2) - (x_{11}x_{21}-x_{12}x_{22})^2} \\
&= |x_{11}x_{22}-x_{12}x_{21}| \\
&= |\det(\boldsymbol{X})|.
\end{aligned}$$

- Another interpretation of the determinant is the volume changing factor when operating on a set in $\mathbb{R}^n$. $\text{volume}(f(S)) = |\det(\boldsymbol{A})|\text{volume}(f(S))$ where $f : \mathbb{R}^n \mapsto \mathbb{R}^n$ is the linear mapping defined by $\boldsymbol{A}$.

- The determinant of a lower or upper triangular matrix $\boldsymbol{A}$ is the product of the diagonal elements $\prod_{i=1}^n a_{ii}$.

- Determinant of a singular matrix is 0.

- Determinant of an orthogonal matrix is 1 (rotation) or -1 (reflection).

- $\det(\boldsymbol{A}^T) = \det(\boldsymbol{A})$.
- $\det(\boldsymbol{A}^{-1}) = 1/\det(\boldsymbol{A})$.
- $\det(c\boldsymbol{A}) = c^n \det(\boldsymbol{A})$.
- $\det(\boldsymbol{AB}) = \det(\boldsymbol{A})\det(\boldsymbol{B})$.
- For $\boldsymbol{A}$ and $\boldsymbol{D}$ square and nonsingular,

$$\det\left(\begin{pmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{pmatrix}\right) = \det(\boldsymbol{A})\det(\boldsymbol{D} - \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B}) = \det(\boldsymbol{D})\det(\boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C}).$$

Proof. Take determinant on the both sides of the matrix identity

$$\begin{pmatrix} \boldsymbol{A} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{D} - \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B} \end{pmatrix} = \begin{pmatrix} \boldsymbol{I} & \boldsymbol{0} \\ -\boldsymbol{C}\boldsymbol{A}^{-1} & \boldsymbol{I} \end{pmatrix} \begin{pmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{pmatrix} \begin{pmatrix} \boldsymbol{I} & -\boldsymbol{A}^{-1}\boldsymbol{B} \\ \boldsymbol{0} & \boldsymbol{I} \end{pmatrix}.$$

□

Review: eigenvalues and eigenvectors (JM A.6)

Assume $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ a square matrix.

- *Eigenvalues* are defined as roots of the characteristic equation $\det(\lambda \boldsymbol{I}_n - \boldsymbol{A}) = 0$.
- $\boldsymbol{A}$ is singular if and only if it has at least one 0 eigenvalue.
- If λ is an eigenvalue of $\boldsymbol{A}$, then there exist non-zero vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ such that $\boldsymbol{A}\boldsymbol{x} = \lambda\boldsymbol{x}$ and $\boldsymbol{y}^T\boldsymbol{A} = \lambda\boldsymbol{y}^T$. $\boldsymbol{x}$ and $\boldsymbol{y}$ are called the (column) *eigenvector* and *row eigenvector* of $\boldsymbol{A}$ associated with the eigenvalue λ.
- Eigenvectors associated with distinct eigenvalues are linearly independent.

  Proof. Let $\boldsymbol{A}\boldsymbol{x}_1 = \lambda_1\boldsymbol{x}_1$, $\boldsymbol{A}\boldsymbol{x}_2 = \lambda_2\boldsymbol{x}_2$, and $\lambda_1 \neq \lambda_2$. Assume that $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are linearly dependent such that $\boldsymbol{x}_2 = \alpha\boldsymbol{x}_1$. Then

$$\alpha\lambda_1\boldsymbol{x}_1 = \alpha\boldsymbol{A}\boldsymbol{x}_1 = \boldsymbol{A}\boldsymbol{x}_2 = \lambda_2\boldsymbol{x}_2 = \alpha\lambda_2\boldsymbol{x}_1.$$

  That is $\alpha(\lambda_1 - \lambda_2)\boldsymbol{x}_1 = \boldsymbol{0}$. Since $\alpha \neq 0$ and $\lambda_1 \neq \lambda_2$, we have $\boldsymbol{x}_1 = \boldsymbol{0}$, a contradiction. □

- Eigenvalues of an upper or lower triangular matrix are its diagonal entries: $\lambda_i = a_{ii}$.

- Eigenvalues of an idempotent matrix are either 0 or 1.

  Proof. If $\boldsymbol{A}\boldsymbol{A} = \boldsymbol{A}$ and $\boldsymbol{A}\boldsymbol{x} = \lambda\boldsymbol{x}$, then

  $$\lambda\boldsymbol{x} = \boldsymbol{A}\boldsymbol{x} = \boldsymbol{A}\boldsymbol{A}\boldsymbol{x} = \lambda^2\boldsymbol{x}.$$

  Hence $\lambda = \lambda^2$, i.e., $\lambda = 0$ or 1. □

- Eigenvalues of an orthogonal matrix have complex modulus 1.

- In most statistical applications, we deal with eigenvalues/eigenvectors of symmetric matrices.
  The eigenvalues and eigenvectors of a real *symmetric* matrix are real.

- Eigenvectors associated with distinct eigenvalues of a symmetry matrix are orthogonal to each other.

- Eigen-decompostion of a symmetric matrix: $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\intercal$, where

  - $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_n)$
  - columns of $\boldsymbol{U}$ are the eigenvectors which are (or can be chosen to be) mutually orthonormal

- A real symmetric matrix is positive semidefinite (positive definite) if and only if all eigenvalues are nonnegative (positive).

  Proof. If $\boldsymbol{A}$ is positive definite and $\boldsymbol{A}\boldsymbol{x} = \lambda\boldsymbol{x}$, then $\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} = \lambda\boldsymbol{x}^T\boldsymbol{x}$. Now $\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} > 0$ and $\boldsymbol{x}^T\boldsymbol{x} > 0$ imply $\lambda > 0$. □

- If $\boldsymbol{A}$ has r non-zero eigenvalues, then $\text{rank}(\boldsymbol{A}) \geq r$.

  If $\boldsymbol{A}$ is symmetric and has r non-zero eigenvalues, then $\text{rank}(\boldsymbol{A}) = r$.

  If $\boldsymbol{A}$ is idempotent and has r eigenvalues equal to 1, then $\text{rank}(\boldsymbol{A}) = \text{tr}(\boldsymbol{A}) = r$.

- If $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ is positive definite and has eigen-decomposition

$$\boldsymbol{A} = \boldsymbol{U}\text{diag}(\lambda_1, \ldots, \lambda_n)\boldsymbol{U}^T,$$

where the eigenvalues λ_i are positive. Then the inverse is

$$\boldsymbol{A}^{-1} = \boldsymbol{U}\text{diag}(\lambda_1^{-1}, \ldots, \lambda_n^{-1})\boldsymbol{U}^T.$$

- If $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ is psd with rank $r < n$ and has eigen-decomposition

$$\boldsymbol{A} = \boldsymbol{U}\text{diag}(\lambda_1, \ldots, \lambda_r, 0, \ldots, 0)\boldsymbol{U}^T,$$

where $\lambda_i > 0$, $i = 1, \ldots, r$, are positive eigenvalues. Then the Moore-Penrose inverse is

$$\boldsymbol{A}^+ = \boldsymbol{U}\text{diag}(\lambda_1^{-1}, \ldots, \lambda_r^{-1}, 0, \ldots, 0)\boldsymbol{U}^T = \boldsymbol{U}_r\text{diag}(\lambda_1^{-1}, \ldots, \lambda_r^{-1})\boldsymbol{U}_r^T,$$

where $\boldsymbol{U}_r$ contains the first r columns of $\boldsymbol{U}$. Note $\boldsymbol{A}^+$ is psd too.

- If $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ is psd, then

$$\boldsymbol{A}^{1/2} = \boldsymbol{U}\text{diag}(\lambda_1^{1/2}, \ldots, \lambda_r^{1/2}, 0, \ldots, 0)\boldsymbol{U}^T = \boldsymbol{U}_r\text{diag}(\lambda_1^{1/2}, \ldots, \lambda_r^{1/2})\boldsymbol{U}_r^T$$

is a symmetric square root of $\boldsymbol{A}$ and $\boldsymbol{A}^{1/2}$ is psd too.

- $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ a square matrix (not required to be symmetric), then $\text{tr}(\boldsymbol{A}) = \sum_i \lambda_i$ and $\det(\boldsymbol{A}) = \prod_i \lambda_i$.

Aitken model with linear constraints and a singular $\boldsymbol{V}$ (alternative approach)

Here we derive an alternative expression for the MVAUE of Aiken model, $\text{E}(\boldsymbol{y}) = \boldsymbol{X}\boldsymbol{b}$ and $\text{Cov}(\boldsymbol{y}) = \sigma^2\boldsymbol{V}$, with linear constraints $\boldsymbol{R}\boldsymbol{b} = \boldsymbol{r}$ and a possibly singular $\boldsymbol{V}$.

- First we show that a singular $\boldsymbol{V}$ essentially imposes linear constraints on $\boldsymbol{b}$.

Suppose $\boldsymbol{V}$ is positive semi-definte with rank s. From eigen-decomposition,

$$\begin{aligned}
\boldsymbol{V} &= \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^T \\
&= (\boldsymbol{S}_{n\times s}, \boldsymbol{T}_{n\times(n-s)}) \begin{pmatrix} d_1 & & & & & \\ & \ddots & & & & \\ & & d_s & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{pmatrix} \begin{pmatrix} \boldsymbol{S}^T_{n\times s} \\ \boldsymbol{T}^T_{n\times(n-s)} \end{pmatrix} \\
&= \boldsymbol{S}\mathrm{diag}(d_1, \ldots, d_s)\boldsymbol{S}^T \\
&= \boldsymbol{S}\boldsymbol{D}_s\boldsymbol{S}^T,
\end{aligned}$$

where $d_1, \ldots, d_s$ are positive eigenvalues, $\boldsymbol{S}^T\boldsymbol{S} = \boldsymbol{I}_s$, $\boldsymbol{T}^T\boldsymbol{T} = \boldsymbol{I}_{n-s}$, and $\boldsymbol{S}^T\boldsymbol{T} = \boldsymbol{0}_{s\times(n-s)}$. $\mathcal{C}(\boldsymbol{S}) = \mathcal{C}(\boldsymbol{V})$ and $\mathcal{C}(\boldsymbol{T}) = \mathcal{N}(\boldsymbol{V}^T) = \mathcal{N}(\boldsymbol{V})$. Note that

$$\boldsymbol{V}^+ = \boldsymbol{S}\boldsymbol{D}_s^{-1}\boldsymbol{S}^T,$$

is the Moore-Penrose inverse of $\boldsymbol{V}$.

Let's transform the original model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{e}$ by $\begin{pmatrix} \boldsymbol{S}^T \\ \boldsymbol{T}^T \end{pmatrix}$:

$$\begin{aligned}
\boldsymbol{S}^T\boldsymbol{y} &= \boldsymbol{S}^T\boldsymbol{X}\boldsymbol{b} + \boldsymbol{S}^T\boldsymbol{e} = \boldsymbol{S}^T\boldsymbol{X}\boldsymbol{b} + \boldsymbol{u} \\
\boldsymbol{T}^T\boldsymbol{y} &= \boldsymbol{T}^T\boldsymbol{X}\boldsymbol{b} + \boldsymbol{T}^T\boldsymbol{e} = \boldsymbol{T}^T\boldsymbol{X}\boldsymbol{b} + \boldsymbol{v},
\end{aligned}$$

where

$$\begin{aligned}
\mathrm{E}(\boldsymbol{u}) &= \boldsymbol{0}_s, & \mathrm{Cov}(\boldsymbol{u}) &= \sigma^2\boldsymbol{D}_s \\
\mathrm{E}(\boldsymbol{v}) &= \boldsymbol{0}_{n-s}, & \mathrm{Cov}(\boldsymbol{v}) &= \boldsymbol{0}_{(n-s)\times(n-s)}.
\end{aligned}$$

This shows that the singularity of $\boldsymbol{V}$ essentially introduces linear constraints to the parameters

$$\boldsymbol{T}^T\boldsymbol{y} = \boldsymbol{T}^T\boldsymbol{X}\boldsymbol{b}.$$

- Next we add explicit constraints. $\boldsymbol{R}\boldsymbol{b} = \boldsymbol{r}$ is the explicit constraints and singularity of $\boldsymbol{V}$ imposes implicit constraints $\boldsymbol{T}^T\boldsymbol{X}\boldsymbol{b} = \boldsymbol{T}^T\boldsymbol{y}$. Collectively we write this as

$$\boldsymbol{R}_0\boldsymbol{b} = \boldsymbol{r}_0,$$

where

$$\boldsymbol{R}_0 = \begin{pmatrix} \boldsymbol{T}^T\boldsymbol{X} \\ \boldsymbol{R} \end{pmatrix} \quad \text{and} \quad \boldsymbol{r}_0 = \begin{pmatrix} \boldsymbol{T}^T\boldsymbol{y} \\ \boldsymbol{r} \end{pmatrix}.$$

- In summary, Aitken model with linear constraints $\boldsymbol{R}\boldsymbol{b} = \boldsymbol{r}$ and a singular $\boldsymbol{V}$ is equivalent to the constrained model

$$\boldsymbol{S}^T\boldsymbol{y} = \boldsymbol{S}^T\boldsymbol{X}\boldsymbol{b} + \boldsymbol{u},$$

with $\mathrm{E}(\boldsymbol{u}) = \boldsymbol{0}_n$, $\mathrm{Cov}(\boldsymbol{u}) = \sigma^2\boldsymbol{D}_s$, and linear constraints

$$\boldsymbol{R}_0\boldsymbol{b} = \boldsymbol{r}_0.$$

- (Estimability and MVAUE) Consider the Aitken model ($\mathrm{E}(\boldsymbol{y}) = \boldsymbol{X}\boldsymbol{b}$, $\mathrm{Cov}(\boldsymbol{y}) = \sigma^2\boldsymbol{V}$) with linear constraints $\boldsymbol{R}\boldsymbol{b} = \boldsymbol{r}$. A linear function $\boldsymbol{\Lambda}\boldsymbol{b}$ is estimable if and only if $\mathcal{C}(\boldsymbol{\Lambda}^T) \subset \mathcal{C}((\boldsymbol{X}^T, \boldsymbol{R}^T))$. And the best (minimum) variance affine unbiased estimator (MVAUE) is

$$\widehat{\boldsymbol{\Lambda}\boldsymbol{b}} = \boldsymbol{\Lambda}\boldsymbol{G}^-\boldsymbol{X}^T\boldsymbol{V}^+\boldsymbol{y} + \boldsymbol{\Lambda}\boldsymbol{G}^-\boldsymbol{R}_0^T(\boldsymbol{R}_0\boldsymbol{G}^-\boldsymbol{R}_0^T)^-(\boldsymbol{r}_0 - \boldsymbol{R}_0\boldsymbol{G}^-\boldsymbol{X}^T\boldsymbol{V}^+\boldsymbol{y}),$$

where

$$\boldsymbol{R}_0 = \begin{pmatrix} \boldsymbol{T}^T\boldsymbol{X} \\ \boldsymbol{R} \end{pmatrix}, \qquad \boldsymbol{r}_0 = \begin{pmatrix} \boldsymbol{T}^T\boldsymbol{y} \\ \boldsymbol{r} \end{pmatrix}, \qquad \boldsymbol{G} = \boldsymbol{X}^T\boldsymbol{V}^+\boldsymbol{X} + \boldsymbol{R}_0^T\boldsymbol{R}_0,$$

and $\boldsymbol{T}$ is a matrix of maximum rank such that $\boldsymbol{V}\boldsymbol{T} = \boldsymbol{0}$. The variance matrix of $\widehat{\boldsymbol{\Lambda}\boldsymbol{b}}$ is

$$\mathrm{Cov}(\widehat{\boldsymbol{\Lambda}\boldsymbol{b}}) = \sigma^2\boldsymbol{\Lambda}\boldsymbol{G}^-\boldsymbol{\Lambda}^T - \sigma^2\boldsymbol{\Lambda}\boldsymbol{G}^-\boldsymbol{R}_0^T(\boldsymbol{R}_0\boldsymbol{G}^-\boldsymbol{R}_0^T)^-\boldsymbol{R}_0\boldsymbol{G}^-\boldsymbol{\Lambda}^T.$$

Proof. Use previous result for Aitken model with linear constraints and a non-singular $\boldsymbol{V}$. □

Normal distribution (JM 5.2)

- So far we studied the following linear models
  - Method of least squares: find $\boldsymbol{b}$ that minimizes $\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\|_2$ (approximation)

- Linear mean model: $\mathrm{E}(\boldsymbol{y}) = \boldsymbol{X}\boldsymbol{b}$ (unbiasedness, estimability)
- Gauss-Markov model: $\mathrm{E}(\boldsymbol{y}) = \boldsymbol{X}\boldsymbol{b}$, $\mathrm{Cov}(\boldsymbol{y}) = \sigma^2\boldsymbol{I}$ (MVAUE, BLUE)
- Aitken model: $\mathrm{E}(\boldsymbol{y}) = \boldsymbol{X}\boldsymbol{b}$, $\mathrm{Cov}(\boldsymbol{y}) = \sigma^2\boldsymbol{V}$ (MVAUE, BLUE)

Next we add distribution assumption $\boldsymbol{y} \sim N(\boldsymbol{X}\boldsymbol{b}, \sigma^2\boldsymbol{V})$ so we can do more: confidence interval, testing, best estimator of σ^2,...

- Moment generating function (mgf) of random variables.
  - Assume $X \sim F_X(t) = P(X \le t)$. If $\mathrm{E}(e^{tX}) < \infty$ for all t in a neighborhood of 0, i.e., $|t| < \delta$ for some $\delta > 0$, then the function
  $$m_X(t) = \mathrm{E}(e^{tX})$$
  is defined for all t such that $|t| < \delta$. $m_X(t)$ is the *moment generating function* (mgf) of X.
  - If $m_X(t)$ exists, then $\mathrm{E}(|X|^j) < \infty$ for $j = 0, 1, \ldots$.
  - Differentiating mgf and setting $t = 0$ yields moments: $m_X(0) = 1$, $m'_X(0) = \mathrm{E}(X)$, $m''_X(0) = \mathrm{E}(X^2)$, ...
  In general, $m_X^{(j)}(0) = \mathrm{E}(X^j)$, $j = 0, 1, 2, \ldots$
  - (mgf determines distribution) If X_1 and X_2 are random variables with mgf $m_{X_1}(t)$ and $m_{X_2}(t)$, then $m_{X_1}(t) = m_{X_2}(t)$ for all t in a neighborhood of 0 if and only if $F_{X_1}(t) = F_{X_2}(t)$ for all t.
  - If $X_1, X_2, \ldots, X_n$ are independent with mgfs $m_{X_1}(t), \ldots, m_{X_n}(t)$ and
  $$Y = a_0 + \sum_{i=1}^{n} a_i X_i,$$
  then
  $$m_Y(t) = e^{ta_0} \prod_{i=1}^{n} m_{X_i}(a_i t).$$
- Moment generating function (mgf) of random vectors.
  - Let $\boldsymbol{X} = (X_1, \ldots, X_p)^T \in \mathbb{R}^p$ be a random vector. Then
  $$m_{\boldsymbol{X}}(\boldsymbol{t}) = \mathrm{E}(e^{\boldsymbol{t}^T\boldsymbol{X}}) = \mathrm{E}(e^{t_1X_1 + \cdots + t_pX_p}), \quad \boldsymbol{t} \in \mathbb{R}^p$$
  is the mgf of $\boldsymbol{X}$ provided that $\mathrm{E}(e^{\boldsymbol{t}^T\boldsymbol{X}}) < \infty$ for all $\boldsymbol{t}$ such that $\|\boldsymbol{t}\| < \delta$ for some $\delta > 0$.

– If $m_{\boldsymbol{X}}(\boldsymbol{t})$ exists, then $\mathrm{E}(|X_1|^{a_1} \cdots |X_p|^{a_p}) < \infty$ for $a_j = 0, 1, 2, \ldots$ Especially $\mathrm{E}(|X_j|^{a_j}) < \infty$, $a_j = 0, 1, 2, \ldots$ Also

$$m_{X_j}(t) = m_{\boldsymbol{X}}(t\boldsymbol{e}_j) = m_{\boldsymbol{X}}((0, \ldots, t, \ldots, 0)^T).$$

– Differentiating mgf and setting $\boldsymbol{t} = \boldsymbol{0}$ yields moments. Let $\boldsymbol{n} = (n_1, \ldots, n_p)$ and $\boldsymbol{t}^{\boldsymbol{n}} = \prod_{i=1}^p t_i^{n_i}$. Then

$$\frac{\partial^{\sum_{i=1}^p n_i}}{\partial \boldsymbol{t}^{\boldsymbol{n}}} m_{\boldsymbol{X}}(\boldsymbol{t})\bigg|_{\boldsymbol{t}=\boldsymbol{0}} = \mathrm{E}\left(\prod_{i=1}^p X_i^{n_i}\right).$$

The first two moments are

$$\begin{aligned} \nabla m_{\boldsymbol{X}}(\boldsymbol{t})|_{\boldsymbol{t}=\boldsymbol{0}} &= \mathrm{E}(\boldsymbol{X}) \\ d^2 m_{\boldsymbol{X}}(\boldsymbol{t})|_{\boldsymbol{t}=\boldsymbol{0}} &= \mathrm{E}(\boldsymbol{X}\boldsymbol{X}^T). \end{aligned}$$

– (mgf determines distribution) $m_{\boldsymbol{X}}(\boldsymbol{t}) = m_{\boldsymbol{Y}}(\boldsymbol{t})$ for all $\|\boldsymbol{t}\| < \delta$ ($\delta > 0$) if and only if $F_{\boldsymbol{X}}(\boldsymbol{t}) = F_{\boldsymbol{Y}}(\boldsymbol{t})$ for all $\boldsymbol{t}$.

– Suppose that $\boldsymbol{X}$ is partitioned as $\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1 \\ \vdots \\ \boldsymbol{X}_m \end{pmatrix}$ and has mgf $m_{\boldsymbol{X}}(\boldsymbol{t}) = \mathrm{E}(e^{\boldsymbol{t}^T\boldsymbol{X}})$. Then $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_m$ are independent if and only if

$$m_{\boldsymbol{X}}(\boldsymbol{t}) = \prod_{i=1}^m m_{\boldsymbol{X}_i}(\boldsymbol{t}_i)$$

for all $\boldsymbol{t} = \begin{pmatrix} \boldsymbol{t}_1 \\ \vdots \\ \boldsymbol{t}_m \end{pmatrix}$ in a neighborhood of $\boldsymbol{0}$.

– If $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are independent random vectors and

$$\boldsymbol{Y} = \boldsymbol{A}_0 + \sum_{i=1}^n \boldsymbol{A}_i \boldsymbol{X}_i,$$

then

$$m_{\boldsymbol{Y}}(\boldsymbol{t}) = e^{\boldsymbol{t}^T \boldsymbol{A}_0} \prod_{i=1}^n m_{\boldsymbol{X}_i}(\boldsymbol{A}_i^T \boldsymbol{t}).$$

- A random variable Z has a *standard normal distribution*, denoted $Z \sim N(0,1)$, if

$$F_Z(t) = P(Z \leq t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\, dz,$$

or equivalently Z has density

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \quad -\infty < z < \infty.$$

or equivalently

$$m_Z(t) = E(e^{tZ}) = e^{t^2/2}, \quad -\infty < t < \infty.$$

- Non-standard normal random variable.
  - Definition 1: A random variable X has *normal distribution* with mean μ and variance σ^2, denoted $X \sim N(\mu, \sigma^2)$, if

$$X =^D \mu + \sigma Z,$$

    where $Z \sim N(0,1)$.
  - Definition 2: $X \sim N(\mu, \sigma^2)$ if

$$m_X(t) = E(e^{tX}) = e^{t\mu + \sigma^2 t^2/2}, \quad -\infty < t < \infty.$$

  - In both definitions, $\sigma^2 = 0$ is allowed. If $\sigma^2 > 0$, it has a density

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}, \quad -\infty < x < \infty.$$

- The *standard multivariate normal* is a vector of independent standard normals, denoted $\boldsymbol{Z} \sim N_p(\mathbf{0}_p, \boldsymbol{I}_p)$. The joint density is

$$f_{\boldsymbol{Z}}(\boldsymbol{z}) = \frac{1}{(2\pi)^{p/2}} e^{-\sum_{i=1}^{p} z_i^2/2}.$$

The mgf is

$$m_{\boldsymbol{Z}}(\boldsymbol{t}) = \prod_{i=1}^{p} m_{Z_i}(t_i) = \prod_{i=1}^{p} e^{t_i^2/2} = e^{\sum_{i=1}^{p} t_i^2/2} = e^{\boldsymbol{t}^T \boldsymbol{t}/2}.$$

- Consider the affine transformation $\boldsymbol{X} = \boldsymbol{\mu} + \boldsymbol{A}\boldsymbol{Z}$, where $\boldsymbol{Z} \sim N_p(\boldsymbol{0}_p, \boldsymbol{I}_p)$. $\boldsymbol{X}$ has mean and variance

$$\mathrm{E}(\boldsymbol{X}) \;=\; \boldsymbol{\mu}, \quad \mathrm{Var}(\boldsymbol{X}) = \boldsymbol{A}\boldsymbol{A}^T$$

and the moment generating function is

$$m_{\boldsymbol{X}}(\boldsymbol{t}) = \mathrm{E}(e^{\boldsymbol{t}^T(\boldsymbol{\mu}+\boldsymbol{A}\boldsymbol{Z})}) = e^{\boldsymbol{t}^T\boldsymbol{\mu}}\mathrm{E}e^{\boldsymbol{t}^T\boldsymbol{A}\boldsymbol{Z}} = e^{\boldsymbol{t}^T\boldsymbol{\mu}+\boldsymbol{t}^T\boldsymbol{A}\boldsymbol{A}^T\boldsymbol{t}/2}.$$

- $\boldsymbol{X} \in \mathbb{R}^p$ has a *multivariate normal distribution* with mean $\boldsymbol{\mu} \in \mathbb{R}^p$ and covariance $\boldsymbol{V} \in \mathbb{R}^{p\times p}$ ($\boldsymbol{V}$ psd), denoted $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{V})$, if its mgf takes the form

$$m_{\boldsymbol{X}}(\boldsymbol{t}) = e^{\boldsymbol{t}^T\boldsymbol{\mu}+\boldsymbol{t}^T\boldsymbol{V}\boldsymbol{t}/2}, \quad \boldsymbol{t} \in \mathbb{R}^p.$$

Remarks

  - We already see any affine transform of a multivariate standard normal is normal. Conversely, any multivariate normal $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{V})$ is also an affine transform of a multivariate standard normal. Just take $\boldsymbol{A}$ to be the Cholesky factor or symmetric square root of $\boldsymbol{V}$.
  - $\boldsymbol{V}$ can be singular, in which case the density does not exist. Suppose $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{V})$ and $\boldsymbol{V}$ is psd with rank $s < p$. From eigen-decomposition

$$\begin{aligned}
\boldsymbol{V} &= \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^T \\
&= (\boldsymbol{S}_{p\times s}, \boldsymbol{T}_{p\times(p-s)}) \begin{pmatrix} d_1 & & & & & \\ & \ddots & & & & \\ & & d_s & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{pmatrix} \begin{pmatrix} \boldsymbol{S}^T_{p\times s} \\ \boldsymbol{T}^T_{p\times(p-s)} \end{pmatrix} \\
&= \boldsymbol{S}\mathrm{diag}(d_1, \ldots, d_s)\boldsymbol{S}^T \\
&= \boldsymbol{S}\boldsymbol{D}_s\boldsymbol{S}^T,
\end{aligned}$$

where $d_1, \ldots, d_s$ are positive eigenvalues, $\boldsymbol{S}^T\boldsymbol{S} = \boldsymbol{I}_s$, $\boldsymbol{T}^T\boldsymbol{T} = \boldsymbol{I}_{p-s}$, and $\boldsymbol{S}^T\boldsymbol{T} = \boldsymbol{0}_{s\times(p-s)}$. Then

$$\begin{aligned}
\boldsymbol{S}^T\boldsymbol{X} &\sim N_s(\boldsymbol{S}^T\boldsymbol{\mu}, \boldsymbol{D}_s) \\
\boldsymbol{T}^T\boldsymbol{X} &= \boldsymbol{T}^T\boldsymbol{\mu}.
\end{aligned}$$

The second set of equations indicates that a singular $\boldsymbol{V}$ imposes constraints on the values $\boldsymbol{X}$ can take. The probability mass lies in a subspace of dimension s.

- If $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{V})$ and $\boldsymbol{V}$ is non-singular, then
  - $\boldsymbol{V} = \boldsymbol{A}\boldsymbol{A}^T$ for some non-singular $\boldsymbol{A}$
  - $\boldsymbol{A}^{-1}(\boldsymbol{X} - \boldsymbol{\mu}) \sim N_p(\boldsymbol{0}_p, \boldsymbol{I}_p)$
  - The density of $\boldsymbol{X}$ is

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{V}|^{1/2}} e^{-(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{V}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})/2}.$$

  Proof. The first fact follows by taking $\boldsymbol{A}$ to be the Cholesky factor or square root of $\boldsymbol{V}$. The second fact is trivial. For the third fact, we use the change of variable formula. Let $\boldsymbol{Z} \sim N_p(\boldsymbol{0}_p, \boldsymbol{I}_p)$. By the change of variable formula $\boldsymbol{X} = \boldsymbol{\mu} + \boldsymbol{A}\boldsymbol{Z}$ has density

$$\begin{aligned} f_{\boldsymbol{X}}(\boldsymbol{x}) &= \frac{1}{(2\pi)^{p/2}} e^{-(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{A}^{-T}\boldsymbol{A}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})/2} |\det(\boldsymbol{A}^{-1})| \\ &= \frac{1}{(2\pi)^{p/2}|\boldsymbol{V}|^{1/2}} e^{-(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{V}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})/2}. \end{aligned}$$

  □

- (Any affine transform of normal is normal) If $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{V})$ and $\boldsymbol{Y} = \boldsymbol{a} + \boldsymbol{B}\boldsymbol{X}$, where $\boldsymbol{a} \in \mathbb{R}^q$ and $\boldsymbol{B} \in \mathbb{R}^{q \times p}$, then $\boldsymbol{Y} \sim N_q(\boldsymbol{a} + \boldsymbol{B}\boldsymbol{\mu}, \boldsymbol{B}\boldsymbol{V}\boldsymbol{B}^T)$.

  Proof. Check the mgf. □

- (Marginal of normal is normal) If $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{V})$, then any subvector of $\boldsymbol{X}$ is normal too.

  Proof. Immediate corollary of the preceding result. □

- A convenient fact about normal random variables/vectors is that zero correlation/covariance implies independence.

If $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{V})$ and is partitioned as

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1 \\ \vdots \\ \boldsymbol{X}_m \end{pmatrix}, \quad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \vdots \\ \boldsymbol{\mu}_m \end{pmatrix}, \quad \boldsymbol{V} = \begin{pmatrix} \boldsymbol{V}_{11} & \cdots & \boldsymbol{V}_{1m} \\ \vdots & & \\ \boldsymbol{V}_{m1} & \cdots & \boldsymbol{V}_{mm} \end{pmatrix},$$

then $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_m$ are jointly independent if and only if $\boldsymbol{V}_{ij} = \boldsymbol{0}$ for all $i \neq j$.

Proof. If $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_m$ are jointly independent, then $\boldsymbol{V}_{ij} = \mathrm{Cov}(\boldsymbol{X}_i, \boldsymbol{X}_j) = \mathrm{E}(\boldsymbol{X}_i - \boldsymbol{\mu}_i)(\boldsymbol{X}_j - \boldsymbol{\mu}_j)^T = \mathrm{E}(\boldsymbol{X}_i - \boldsymbol{\mu}_i)\mathrm{E}(\boldsymbol{X}_j - \boldsymbol{\mu}_j)^T = \boldsymbol{0}_{p_i}\boldsymbol{0}_{p_j}^T = \boldsymbol{0}_{p_i \times p_j}$. Conversely, if $\boldsymbol{V}_{ij} = \boldsymbol{0}$ for all $i \neq j$, then the mgf of $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_m)^T$ is

$$\begin{aligned} m_{\boldsymbol{X}}(\boldsymbol{t}) &= e^{\boldsymbol{t}^T\boldsymbol{\mu} + \boldsymbol{t}^T\boldsymbol{V}\boldsymbol{t}/2} \\ &= e^{\sum_{i=1}^m \boldsymbol{t}_i^T\boldsymbol{\mu}_i + \sum_{i=1}^m \boldsymbol{t}_i^T\boldsymbol{V}_{ii}\boldsymbol{t}_i/2} \\ &= m_{\boldsymbol{X}_1}(\boldsymbol{t}_1) \cdots m_{\boldsymbol{X}_m}(\boldsymbol{t}_m). \end{aligned}$$

Thus $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_m$ are jointly independent. □

- Let $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{V})$, $\boldsymbol{Y}_1 = \boldsymbol{a}_1 + \boldsymbol{B}_1\boldsymbol{X}$, and $\boldsymbol{Y}_2 = \boldsymbol{a}_2 + \boldsymbol{B}_2\boldsymbol{X}$. Then $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ are independent if and only if $\boldsymbol{B}_1\boldsymbol{V}\boldsymbol{B}_2^T = \boldsymbol{0}$.

  Proof. Note $\mathrm{Cov}(\boldsymbol{Y}_1, \boldsymbol{Y}_2) = \boldsymbol{B}_1\mathrm{Cov}(\boldsymbol{X})\boldsymbol{B}_2^T = \boldsymbol{B}_1\boldsymbol{V}\boldsymbol{B}_2^T$. □

Chi-square and related distributions (JM 5.3)

- Let $\boldsymbol{Z} \sim N_p(\boldsymbol{0}_p, \boldsymbol{I}_p)$. Then $U = \|\boldsymbol{Z}\|_2^2 = \sum_{i=1}^p Z_i^2$ has the *chi-square distribution* with p degrees of freedom, denoted by $U \sim \chi_p^2$.
  - The mgf of U is

$$\begin{aligned} m_U(t) &= \mathrm{E}(e^{tU}) = \mathrm{E}\left(e^{t\sum_{i=1}^p Z_i^2}\right) \\ &= \prod_{i=1}^p \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{tz_i^2 - z_i^2/2}\, dz_i \\ &= \prod_{i=1}^p \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{z_i^2}{2(1-2t)^{-1}}}\, dz_i \\ &= \prod_{i=1}^p (1-2t)^{-1/2} \\ &= (1-2t)^{-p/2}. \end{aligned}$$

– The first two moments of U are

$$\mathrm{E}(U) = p, \quad \mathrm{Var}(U) = 2p.$$

– The density of U is

$$f_U(u) = \frac{u^{(p-2)/2}e^{-u/2}}{\Gamma(p/2)2^{p/2}}, \quad u > 0.$$

- Let $J \sim \text{Poisson}(\phi)$ and $U|J = j \sim \chi^2_{p+2j}$. Then the unconditional distribution of U is the *noncentral chi-square distribution* with noncentrality parameter ϕ, denoted by $U \sim \chi^2_p(\phi)$.

  – The density of U is

$$f_U(u) = \sum_{j=0}^{\infty} e^{-\phi}\frac{\phi^j}{j!} \times \frac{u^{(p+2j-2)/2}e^{-u/2}}{\Gamma((p+2j)/2)2^{(p+2j)/2}}, \quad u > 0.$$

  – The mgf of U is

$$\begin{aligned} m_U(t) &= \mathrm{E}(e^{tU}) = \mathrm{E}[\mathrm{E}(e^{tU}|J)] \\ &= \mathrm{E}(1-2t)^{-(p+2J)/2} \\ &= (1-2t)^{-p/2}\mathrm{E}[(1-2t)^{-J}] \\ &= (1-2t)^{-p/2}e^{2\phi t/(1-2t)}. \end{aligned}$$

  – The first two moments of U are

$$\mathrm{E}(U) = p + 2\phi, \quad \mathrm{Var}(U) = 2p + 8\phi.$$

- If $U_i \sim \chi^2_{p_i}(\phi_i)$, $i = 1, \ldots, m$, are jointly independent, then $U = \sum_{i=1}^m U_i \sim \chi_p(\phi)$, where $p = \sum_{i=1}^m p_i$ and $\phi = \sum_{i=1}^m \phi_i$.

  Proof. The mgf of U is

$$m_U(t) = \prod_{i=1}^m m_{U_i}(t) = \prod_{i=1}^m (1-2t)^{-p_i/2}e^{2t\phi_i/(1-2t)} = (1-2t)^{-p/2}e^{2t\phi/(1-2t)}.$$

  □

- If $X \sim N(\mu, 1)$, then $U = X^2 \sim \chi^2_1(\mu^2/2)$.

Proof. The mgf of U is

$$\begin{aligned} m_U(t) &= \mathrm{E}(e^{tU}) = \mathrm{E}(e^{tX^2}) \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{tx^2-(x-\mu)^2/2} \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1-2t}{2}(x-\frac{\mu}{1-2t})^2-\mu^2+\frac{\mu^2}{1-2t}} \\ &= (1-2t)^{-1/2} \times e^{(\mu^2/2)2t/(1-2t)}, \end{aligned}$$

which matches that of $\chi^2_1(\mu^2/2)$. □

- If $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{I}_p)$, then $W = \|\boldsymbol{X}\|_2^2 = \sum_{i=1}^p X_i^2 \sim \chi^2_p(\frac{1}{2}\boldsymbol{\mu}^T\boldsymbol{\mu})$.

  Proof. Immediate corollary of the preceding two results. □

- If $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{V})$ and $\boldsymbol{V}$ is non-singular. Then $W = \boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X} \sim \chi^2_p(\frac{1}{2}\boldsymbol{\mu}^T\boldsymbol{V}^{-1}\boldsymbol{\mu})$.

  Proof. Let $\boldsymbol{V} = \boldsymbol{A}\boldsymbol{A}^T$. Then $\boldsymbol{Z} = \boldsymbol{A}^{-1}\boldsymbol{X} \sim N_p(\boldsymbol{A}^{-1}\boldsymbol{\mu}, \boldsymbol{I}_p)$. And

  $$W = \boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X} = \boldsymbol{Z}^T\boldsymbol{Z} \sim \chi^2_p(\frac{1}{2}\boldsymbol{\mu}^T\boldsymbol{V}^{-1}\boldsymbol{\mu}).$$

  □

- (Stochastic monotonicity of non-central chi-square) Let $U \sim \chi^2_p(\phi)$. Then $P(U > c)$ is increasing in ϕ for fixed p and $c > 0$.

  Proof. For an algebraic proof, see JM Result 5.11 (p106). For a probabilistic proof, we fix p and assume $\phi_2 > \phi_1 > 0$. Let $X_1, X_2, \ldots$ be a sequence of independent standard normals, $J_1 \sim \text{Poisson}(\phi_1)$, $\Delta J \sim \text{Poisson}(\phi_2 - \phi_1)$, and $X_1, X_2, \ldots, J_1, \Delta J)$ are jointly independent. Let $J_2 = J_1 + \Delta J \sim \text{Poisson}(\phi_2)$. Define

  $$\begin{aligned} U_1|J_1 &= \sum_{i=1}^{p+2J_1} X_i^2 \\ U_2|J_2 &= \sum_{i=1}^{p+2J_2} X_i^2. \end{aligned}$$

  Then $U_1 \sim \chi^2_p(\phi_1)$, $U_2 \sim \chi^2_p(\phi_2)$, and $U_2 \geq U_1$ almost surely. Hence $P(U_2 > c) \geq P(U_1 > c)$. □

- Let $U_1 \sim \chi^2_{p_1}$ and $U_2 \sim \chi^2_{p_2}$ be independent. Then $F = \frac{U_1/p_1}{U_2/p_2}$ has the *F-distribution* with p_1 and p_2 degrees of freedom, denoted by $F \sim F_{p_1,p_2}$.

  - Then density of F is

$$f_F(f) = \frac{\Gamma(\frac{p_1+p_2}{2})\left(\frac{p_2}{p_1}\right)^{p_1/2}}{\Gamma(\frac{p_1}{2})\Gamma(\frac{p_2}{2})} f^{p_1/2-1}\left(1+\frac{p_1}{p_2}f\right)^{-(p_1+p_2)/2}, \quad f>0.$$

  - Not all moments of F exist.

$$\begin{aligned} \mathrm{E}(F) &= \frac{p_2}{p_2-2} \text{ for } p_2 > 2 \\ \mathrm{Var}(F) &= \frac{2p_2^2(p_1+p_2-2)}{p_1(p_2-2)^2(p_2-4)} \text{ for } p_2 > 4. \end{aligned}$$

- Let $U_1 \sim \chi^2_{p_1}(\phi)$ and $U_2 \sim \chi^2_{p_2}$ be independent. Then $F = \frac{U_1/p_1}{U_2/p_2}$ has the *noncentral F-distribution* with p_1 and p_2 degrees of freedom and *noncentrality* parameter ϕ, denoted by $F \sim F_{p_1,p_2}(\phi)$.

- (Stochastic monotonicity of noncentral F) Let $W \sim F_{p_1,p_2}(\phi)$. Then $P(W > c)$ is strictly increasing in ϕ for fixed p_1 and p_2.

  Proof. It follows from stochastic monotonicity of non-central chi-square. □

- Let $U \sim N(\mu, 1)$ and $V \sim \chi^2_k$ be independent. Then $T = U/\sqrt{V/k}$ has the *noncentral Student's t-distribution* with k degrees of freedom and noncentrality parameter μ, denoted $T \sim t_k(\mu)$.

  If $T \sim t_k(\mu)$, then $T^2 = F_{1,k}(\mu^2/2)$.

  If $\mu = 0$, the distribution is the *Student's t distribution*, denoted by $T \sim t_k$, and has density

$$f_T(t) = \frac{\Gamma((k+1)/2)}{\Gamma(k/2)\sqrt{\pi k}} \times (1+t^2/k)^{-(k+1)/2}.$$

13 Lecture 13: Oct 21

Announcement

- HW5 posted and due this Fri Oct 25.

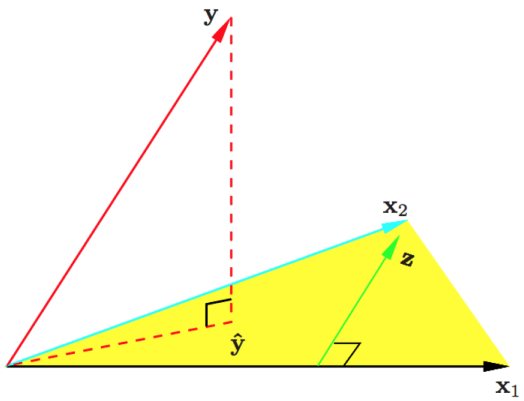
Last time

- Multivariate normal $N_p(\boldsymbol{\mu}, \boldsymbol{V})$, mgf $m(\boldsymbol{t}) = e^{\boldsymbol{t}^T\boldsymbol{\mu}+\boldsymbol{t}^T\boldsymbol{V}\boldsymbol{t}/2}$, correlation and independence, existence of density when $\boldsymbol{V}$ is non-singular.
- (Central) Chi-square distribution: If $\boldsymbol{Z} \sim N_p(\boldsymbol{0}_p, \boldsymbol{I}_p)$, then $U = \|\boldsymbol{Z}\|_2^2 \sim \chi_p^2$.
- Noncentral chi-square distribution:
  - If $\boldsymbol{Z} \sim N_p(\boldsymbol{\mu}, \boldsymbol{I}_p)$, then $U = \|\boldsymbol{Z}\|_2^2 \sim \chi_p^2(\boldsymbol{\mu}^T\boldsymbol{\mu}/2)$.
  - If $\boldsymbol{Z} \sim N_p(\boldsymbol{\mu}, \boldsymbol{V})$, then $U = \boldsymbol{Z}^T\boldsymbol{V}^{-1}\boldsymbol{Z} \sim \chi_p^2(\boldsymbol{\mu}^T\boldsymbol{V}^{-1}\boldsymbol{\mu}/2)$.
  - $U \sim \chi_p^2(\phi)$ is stochastic monotone in p and ϕ.
  - If $U_i \sim \chi_{p_i}^2(\phi_i)$ are independent, then $\sum_i U_i \sim \chi_{\sum_i p_i}^2(\sum_i \phi_i)$.
- (Central) F distribution: If $U_1 \sim \chi_{p_1}^2$ and $U_2 \sim \chi_{p_2}^2$, then $(U_1/p_1)/(U_2/p_2) \sim F_{p_1,p_2}$.
- Noncentral F distribution: If $U_1 \sim \chi_{p_1}^2(\phi)$ and $U_2 \sim \chi_{p_2}^2$, then $F = (U_1/p_1)/(U_2/p_2) \sim F_{p_1,p_2}(\phi)$. $F_{p_1,p_2}(\phi)$ is stochastic monotone in ϕ.
- Noncentral t distribution: If $U \sim N(\mu, 1)$ and $V \sim \chi_p^2$, then $T = U/\sqrt{V/p} \sim t_p(\mu)$.

Today

- Distribution of quadratic forms (JM 5.4)
- Cochran's theorem (JM 5.5)

Distribution of quadratic forms (JM 5.4)



- Motivation: From the geometry of least squares problem, we know

$$\boldsymbol{y} \;=\; \hat{\boldsymbol{y}} + \hat{\boldsymbol{e}} = \boldsymbol{P_X y} + (\boldsymbol{I} - \boldsymbol{P_X})\boldsymbol{y}$$

  and

$$\|\boldsymbol{y}\|_2^2 \;=\; \|\hat{\boldsymbol{y}}\|_2^2 + \|\hat{\boldsymbol{e}}\|_2^2 = \boldsymbol{y}^T\boldsymbol{P_X y} + \boldsymbol{y}^T(\boldsymbol{I} - \boldsymbol{P_X})\boldsymbol{y}.$$

  Now assuming $\boldsymbol{y} \sim N_n(\boldsymbol{Xb}, \sigma^2\boldsymbol{I}_n)$, we would like to know the distribution of the sums of squares $\boldsymbol{y}^T\boldsymbol{P_X y}$ and $\boldsymbol{y}^T(\boldsymbol{I} - \boldsymbol{P_X})\boldsymbol{y}$.

- A symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is idempotent with rank s if and only if there exists a matrix $\boldsymbol{G} \in \mathbb{R}^{n\times s}$ with orthonormal columns, that is $\boldsymbol{G}^T\boldsymbol{G} = \boldsymbol{I}_s$, such that $\boldsymbol{A} = \boldsymbol{G}\boldsymbol{G}^T$.

  Proof. The "if part" is easy. For the "only if part", recall that the eigenvalues of an idempotent matrix are either 1 or 0 and $\text{rank}(\boldsymbol{A}) = s$ equals the number of nonzero eigenvalues. Thus by the eigen-decomposition,

$$\boldsymbol{A} = (\boldsymbol{Q}_1, \boldsymbol{Q}_2)\begin{pmatrix}\boldsymbol{I}_s & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}\end{pmatrix}\begin{pmatrix}\boldsymbol{Q}_1^T\\ \boldsymbol{Q}_2^T\end{pmatrix} = \boldsymbol{Q}_1\boldsymbol{Q}_1^T,$$

  where $\boldsymbol{Q}_1 \in \mathbb{R}^{n\times s}$ and $\boldsymbol{Q}_2 \in \mathbb{R}^{n\times(n-s)}$. □

- Let $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{I}_p)$ and $\boldsymbol{A}$ be symmetric and idempotent with rank s. Then $\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X} \sim \chi_s^2(\frac{1}{2}\boldsymbol{\mu}^T\boldsymbol{A}\boldsymbol{\mu})$.

Proof. By the preceding result, $\boldsymbol{A} = \boldsymbol{G}\boldsymbol{G}^T$, where $\boldsymbol{G} \in \mathbb{R}^{p\times s}$ and $\boldsymbol{G}\boldsymbol{G}^T = \boldsymbol{I}_s$. Then $\boldsymbol{G}^T\boldsymbol{X} \sim N_s(\boldsymbol{G}^T\boldsymbol{\mu}, \boldsymbol{I}_s)$ and we have

$$\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X} = (\boldsymbol{G}^T\boldsymbol{X})^T\boldsymbol{G}^T\boldsymbol{X} = \|\boldsymbol{G}^T\boldsymbol{X}\|_2^2 \sim \chi_s^2\left(\frac{1}{2}\boldsymbol{\mu}^T\boldsymbol{A}\boldsymbol{\mu}\right).$$

□

- (General case) Let $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{V})$ with $\boldsymbol{V}$ non-singular, and $\boldsymbol{A} \in \mathbb{R}^{p\times p}$ be symmetric. If $\boldsymbol{A}\boldsymbol{V}$ is idempotent with rank s, then $\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X} \sim \chi_s^2(\frac{1}{2}\boldsymbol{\mu}^T\boldsymbol{A}\boldsymbol{\mu})$.

  Proof. Let $\boldsymbol{V} = \boldsymbol{\Gamma}\boldsymbol{\Gamma}^T$, where $\boldsymbol{\Gamma}$ can be the Cholesky factor or symmetric square root of $\boldsymbol{V}$, and $\boldsymbol{Y} = \boldsymbol{\Gamma}^{-1}\boldsymbol{X} \sim N_p(\boldsymbol{\Gamma}^{-1}\boldsymbol{\mu}, \boldsymbol{I}_p)$. Note

  $$\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X} = \boldsymbol{Y}^T\boldsymbol{\Gamma}^T\boldsymbol{A}\boldsymbol{\Gamma}\boldsymbol{Y}.$$

  The preceding results applies if $\boldsymbol{\Gamma}^T\boldsymbol{A}\boldsymbol{\Gamma}$ is symmetric (trivial) and idempotent. Idempotency holds since

  $$(\boldsymbol{\Gamma}^T\boldsymbol{A}\boldsymbol{\Gamma})(\boldsymbol{\Gamma}^T\boldsymbol{A}\boldsymbol{\Gamma}) = \boldsymbol{\Gamma}^T\boldsymbol{A}\boldsymbol{V}\boldsymbol{A}\boldsymbol{\Gamma} = \boldsymbol{\Gamma}^T\boldsymbol{A}\boldsymbol{\Gamma}.$$

  The last equation is true since $\boldsymbol{A}\boldsymbol{V}$ is a projection onto $\mathcal{C}(\boldsymbol{A}\boldsymbol{V}) = \mathcal{C}(\boldsymbol{A})$. Also note that $\text{rank}(\boldsymbol{\Gamma}^T\boldsymbol{A}\boldsymbol{\Gamma}) = \text{rank}(\boldsymbol{A}) = \text{rank}(\boldsymbol{A}\boldsymbol{V}) = s$. Therefore by the preceding result, $\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X} \sim \chi_s^2(\phi)$ with $\phi = \frac{1}{2}\boldsymbol{\mu}^T\boldsymbol{\Gamma}^{-T}\boldsymbol{\Gamma}^T\boldsymbol{A}\boldsymbol{\Gamma}\boldsymbol{\Gamma}^{-1}\boldsymbol{\mu} = \frac{1}{2}\boldsymbol{\mu}^T\boldsymbol{A}\boldsymbol{\mu}$. □

- Consider the normal linear model $\boldsymbol{y} \sim N_n(\boldsymbol{X}\boldsymbol{b}, \sigma^2\boldsymbol{I}_n)$.
  - Using $\boldsymbol{A} = (1/\sigma^2)(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}})$, we have

    $$\text{SSE}/\sigma^2 = \|\hat{\boldsymbol{e}}\|_2^2/\sigma^2 = \boldsymbol{y}^T\boldsymbol{A}\boldsymbol{y} \sim \chi_{n-r}^2,$$

    where $r = \text{rank}(\boldsymbol{X})$. Note the noncentrality parameter is

    $$\phi = \frac{1}{2}(\boldsymbol{X}\boldsymbol{b})^T(1/\sigma^2)(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}})(\boldsymbol{X}\boldsymbol{b}) = 0 \text{ for all } \boldsymbol{b}.$$

  - Using $\boldsymbol{A} = (1/\sigma^2)\boldsymbol{P}_{\boldsymbol{X}}$, we have

    $$\text{SSR}/\sigma^2 = \|\hat{\boldsymbol{y}}\|_2^2/\sigma^2 = \boldsymbol{y}^T\boldsymbol{A}\boldsymbol{y} \sim \chi_r^2(\phi)$$

    with the noncentrality parameter

    $$\phi = \frac{1}{2}(\boldsymbol{X}\boldsymbol{b})^T(1/\sigma^2)\boldsymbol{P}_{\boldsymbol{X}}(\boldsymbol{X}\boldsymbol{b}) = \frac{1}{2\sigma^2}\|\boldsymbol{X}\boldsymbol{b}\|_2^2.$$

– The joint distribution of $\hat{\boldsymbol{y}}$ and $\hat{\boldsymbol{e}}$ is

$$\begin{pmatrix}\hat{\boldsymbol{y}}\\ \hat{\boldsymbol{e}}\end{pmatrix} = \begin{pmatrix}\boldsymbol{P_X}\\ \boldsymbol{I}_n - \boldsymbol{P_X}\end{pmatrix}\boldsymbol{y} \sim N_{2n}\left(\begin{pmatrix}\boldsymbol{Xb}\\ \boldsymbol{0}_n\end{pmatrix}, \begin{pmatrix}\sigma^2\boldsymbol{P_X} & \boldsymbol{0}\\ \boldsymbol{0} & \sigma^2(\boldsymbol{I}-\boldsymbol{P_X})\end{pmatrix}\right).$$

So $\hat{\boldsymbol{y}}$ is independent of $\hat{\boldsymbol{e}}$. Thus $\|\hat{\boldsymbol{y}}\|_2^2$ is independent of $\|\hat{\boldsymbol{e}}\|_2^2$ and

$$F = \frac{\|\hat{\boldsymbol{y}}\|_2^2/r}{\|\hat{\boldsymbol{e}}\|_2^2/(n-r)} \sim F_{r,n-r}\left(\frac{1}{2\sigma^2}\|\boldsymbol{Xb}\|_2^2\right).$$

- (Independence between two linear forms of a multivariate normal) In last lecture we showed the following result.

  Let $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{V})$. Then $\boldsymbol{a}_1 + \boldsymbol{B}_1\boldsymbol{X}$ and $\boldsymbol{a}_2 + \boldsymbol{B}_2\boldsymbol{X}$ are independent if and only if $\boldsymbol{B}_1\boldsymbol{V}\boldsymbol{B}_2^T = \boldsymbol{0}$.

- (Independence between linear and quadratic forms of a multivariate normal) Let $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{V})$ and $\boldsymbol{A}$ is symmetric with rank s. If $\boldsymbol{BVA} = \boldsymbol{0}$, then $\boldsymbol{BX}$ and $\boldsymbol{X}^T\boldsymbol{AX}$ are independent.

  Proof. By eigen-decomposition, $\boldsymbol{A} = \boldsymbol{Q}_1\boldsymbol{\Lambda}_1\boldsymbol{Q}_1^T$, where $\boldsymbol{Q}_1^T\boldsymbol{Q}_1 = \boldsymbol{I}_s$ and $\boldsymbol{\Lambda}_1 \in \mathbb{R}^{s\times s}$ is non-singular. Consider the joint distribution

  $$\begin{pmatrix}\boldsymbol{BX}\\ \boldsymbol{Q}_1^T\boldsymbol{X}\end{pmatrix} \sim N\left(\begin{pmatrix}\boldsymbol{B\mu}\\ \boldsymbol{Q}_1^T\boldsymbol{\mu}\end{pmatrix}, \begin{pmatrix}\boldsymbol{BVB}^T & \boldsymbol{BVQ}_1\\ \boldsymbol{Q}_1^T\boldsymbol{VB}^T & \boldsymbol{Q}_1^T\boldsymbol{VQ}_1\end{pmatrix}\right).$$

  By hypothesis

  $$\boldsymbol{BVA} = \boldsymbol{BVQ}_1\boldsymbol{\Lambda}_1\boldsymbol{Q}_1^T = \boldsymbol{0}.$$

  Post-multiplying both sides by $\boldsymbol{Q}_1\boldsymbol{\Lambda}_1^{-1}$ gives $\boldsymbol{BVQ}_1 = \boldsymbol{0}$, which implies that $\boldsymbol{BX}$ is independent of both $\boldsymbol{Q}_1^T\boldsymbol{X}$ and $\boldsymbol{X}^T\boldsymbol{Q}_1\boldsymbol{\Lambda}_1\boldsymbol{Q}_1\boldsymbol{X} = \boldsymbol{X}^T\boldsymbol{AX}$. □

- (Independence between two quadratic forms of a multivariate normal) Let $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{V})$, $\boldsymbol{A}$ be symmetric with rank r, and $\boldsymbol{B}$ be symmetric with rank s. If $\boldsymbol{BVA} = \boldsymbol{0}$, then $\boldsymbol{X}^T\boldsymbol{AX}$ and $\boldsymbol{X}^T\boldsymbol{BX}$ are independent.

  Proof. Again by eigen-decomposition,

  $$\begin{aligned}\boldsymbol{A} &= \boldsymbol{Q}_1\boldsymbol{\Lambda}_1\boldsymbol{Q}_1^T, \quad \text{where } \boldsymbol{Q}_1 \in \mathbb{R}^{p\times r}, \boldsymbol{\Lambda}_1 \in \mathbb{R}^{r\times r} \text{ nonsingular}\\ \boldsymbol{B} &= \boldsymbol{Q}_2\boldsymbol{\Lambda}_2\boldsymbol{Q}_2^T, \quad \text{where } \boldsymbol{Q}_2 \in \mathbb{R}^{p\times s}, \boldsymbol{\Lambda}_2 \in \mathbb{R}^{s\times s} \text{ nonsingular.}\end{aligned}$$

Now consider the joint distribution

$$\begin{pmatrix} \boldsymbol{Q}_1^T\boldsymbol{X} \\ \boldsymbol{Q}_2^T\boldsymbol{X} \end{pmatrix} \sim N\left(\begin{pmatrix} \boldsymbol{Q}_1^T\boldsymbol{\mu} \\ \boldsymbol{Q}_2^T\boldsymbol{\mu} \end{pmatrix}, \begin{pmatrix} \boldsymbol{Q}_1^T\boldsymbol{V}\boldsymbol{Q}_2 & \boldsymbol{Q}_1^T\boldsymbol{V}\boldsymbol{Q}_2 \\ \boldsymbol{Q}_2^T\boldsymbol{V}\boldsymbol{Q}_1 & \boldsymbol{Q}_2^T\boldsymbol{V}\boldsymbol{Q}_2 \end{pmatrix} \right).$$

By hypothesis

$$\boldsymbol{B}\boldsymbol{V}\boldsymbol{A} = \boldsymbol{Q}_2\boldsymbol{\Lambda}_2\boldsymbol{Q}_2^T\boldsymbol{V}\boldsymbol{Q}_1\boldsymbol{\Lambda}_1\boldsymbol{Q}_1^T = \boldsymbol{0}.$$

Pre-multiplying both sides by $\boldsymbol{\Lambda}_2^{-1}\boldsymbol{Q}_2^T$ and then post-multiplying both sides by $\boldsymbol{Q}_1\boldsymbol{\Lambda}_1^{-1}$ gives

$$\boldsymbol{Q}_2^T\boldsymbol{V}\boldsymbol{Q}_1 = \boldsymbol{0}.$$

Therefore $\boldsymbol{Q}_1^T\boldsymbol{X}$ is independent of $\boldsymbol{Q}_2^T\boldsymbol{X}$, which implies $\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X} = \boldsymbol{X}^T\boldsymbol{Q}_1\boldsymbol{\Lambda}_1\boldsymbol{Q}_1^T$ is independent of $\boldsymbol{X}^T\boldsymbol{B}\boldsymbol{X} = \boldsymbol{X}^T\boldsymbol{Q}_2\boldsymbol{\Lambda}_2\boldsymbol{Q}_2^T$. □

14 Lecture 14: Oct 23

Announcement

- HW5 due this Fri Oct 25.
- Homework session this afternoon.

Last time

- Distribution of quadratic forms (JM 5.4)
  - $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{V})$ with $\boldsymbol{V}$ non-singular, $\boldsymbol{A} \in \mathbb{R}^{p \times p}$ be symmetric, and $\boldsymbol{AV}$ is idempotent with rank s, then $\boldsymbol{X}^T\boldsymbol{AX} \sim \chi^2_s(\frac{1}{2}\boldsymbol{\mu}^T\boldsymbol{A\mu})$.
  - See HW5 for a more general version.
- Independence between linear and quadratic forms of a normal $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \boldsymbol{V})$
  - $(\boldsymbol{a} + \boldsymbol{AX}) \perp (\boldsymbol{b} + \boldsymbol{BX})$ if $\boldsymbol{AVB}^T = \boldsymbol{0}$.
  - $\boldsymbol{BX} \perp \boldsymbol{X}^T\boldsymbol{AX}$ if $\boldsymbol{A}$ is symmetric and $\boldsymbol{BVA} = \boldsymbol{0}$.
  - $\boldsymbol{X}^T\boldsymbol{AX} \perp \boldsymbol{X}^T\boldsymbol{BX}$ if $\boldsymbol{A}$ and $\boldsymbol{B}$ are symmetric and $\boldsymbol{BVA} = \boldsymbol{0}$.

Today

- Cochran's theorem (JM 5.5)
- Statistical inference for normal Gauss-Markov model $\boldsymbol{y} \sim N_n(\boldsymbol{Xb}, \sigma^2\boldsymbol{I})$

Cochran theorem (JM 5.5)

- We already saw that, under the normal linear model $\boldsymbol{y} \sim N_n(\boldsymbol{Xb}, \sigma^2\boldsymbol{I})$, the sum of squares $\|\hat{\boldsymbol{y}}\|_2^2 = \|\boldsymbol{y}^T\boldsymbol{P_X y}\|_2^2$ and $\|\hat{\boldsymbol{e}}\|_2^2 = \|\boldsymbol{y}^T(\boldsymbol{I} - \boldsymbol{P_X})\boldsymbol{y}\|_2^2$ are independent chi-square distributions.

  The Cochran theorem deals with the distributions of more general sum of squares from normal linear model.

- (Cochran theorem) Let $\boldsymbol{y} \sim N_n(\boldsymbol{\mu}, \sigma^2 \boldsymbol{I}_n)$ and $\boldsymbol{A}_i$, $i = 1, \ldots, k$, be symmetric idempotent matrix with rank s_i. If $\sum_{i=1}^k \boldsymbol{A}_i = \boldsymbol{I}_n$, then $(1/\sigma^2)\boldsymbol{y}^T \boldsymbol{A}_i \boldsymbol{y}$ are independent $\chi^2_{s_i}(\phi_i)$, with $\phi_i = \frac{1}{2\sigma^2}\boldsymbol{\mu}^T \boldsymbol{A}_i \boldsymbol{\mu}$ and $\sum_{i=1}^k s_i = n$.

  Proof. Since $\boldsymbol{A}_i$ is symmetric and idempotent with rank s_i, $\boldsymbol{A}_i = \boldsymbol{Q}_i \boldsymbol{Q}_i^T$ with $\boldsymbol{Q}_i \in \mathbb{R}^{n \times s_i}$ and $\boldsymbol{Q}_i^T \boldsymbol{Q}_i = \boldsymbol{I}_{s_i}$. Define $\boldsymbol{Q} = (\boldsymbol{Q}_1, \ldots, \boldsymbol{Q}_k) \in \mathbb{R}^{n \times \sum_{i=1}^k s_i}$. Note

$$\begin{aligned} \boldsymbol{Q}^T \boldsymbol{Q} &= \boldsymbol{I}_{\sum_{i=1}^k s_i} \\ \boldsymbol{Q}\boldsymbol{Q}^T &= \sum_{i=1}^k \boldsymbol{Q}_i \boldsymbol{Q}_i^T = \sum_{i=1}^k \boldsymbol{A}_i = \boldsymbol{I}_n. \end{aligned}$$

  From the second equation, we have

$$\sum_{i=1}^k s_i = \sum_{i=1}^k \operatorname{rank}(\boldsymbol{A}_i) = \sum_{i=1}^k \operatorname{tr}(\boldsymbol{A}_i) = \operatorname{tr}(\boldsymbol{I}_n) = n.$$

  Now

$$\boldsymbol{Q}^T \boldsymbol{y} = \begin{pmatrix} \boldsymbol{Q}_1^T \boldsymbol{y} \\ \vdots \\ \boldsymbol{Q}_k^T \boldsymbol{y} \end{pmatrix} \sim N_n \left(\begin{pmatrix} \boldsymbol{Q}_1^T \boldsymbol{\mu} \\ \vdots \\ \boldsymbol{Q}_k^T \boldsymbol{\mu} \end{pmatrix}, \sigma^2 \boldsymbol{I}_n \right),$$

  implying that $\boldsymbol{Q}_i^T \boldsymbol{y} \sim N_{s_i}(\boldsymbol{Q}_i^T \boldsymbol{\mu}, \sigma^2 \boldsymbol{I}_{s_i})$ are jointly independent. Therefore $(1/\sigma^2)\boldsymbol{y}^T \boldsymbol{A}_i \boldsymbol{y} = (1/\sigma^2)\|\boldsymbol{Q}_i^T \boldsymbol{y}\|_2^2 \sim \chi^2_{s_i}(\frac{1}{2\sigma^2}\boldsymbol{\mu}\boldsymbol{A}_i\boldsymbol{\mu})$ are jointly independent. □

- Application to the one-way ANOVA: $y_{ij} = \mu + \alpha_i + e_{ij}$, or $\boldsymbol{y} = N_n(\boldsymbol{X}\boldsymbol{b}, \sigma^2 \boldsymbol{I}_n)$ where

$$\boldsymbol{X}\boldsymbol{b} = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{1}_{n_1} & & & \\ \mathbf{1}_{n_2} & & \mathbf{1}_{n_2} & & \\ \vdots & & & \ddots & \\ \mathbf{1}_{n_a} & & & & \mathbf{1}_{n_a} \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_a \end{pmatrix}.$$

Note

$$
\begin{aligned}
\boldsymbol{P_1} &= n^{-1}\mathbf{1}\mathbf{1}^T \\
\boldsymbol{P_X} &= \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T \\
&= \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{1}_{n_1} & & & \\ \mathbf{1}_{n_2} & & \mathbf{1}_{n_2} & & \\ \vdots & & & \ddots & \\ \mathbf{1}_{n_a} & & & & \mathbf{1}_{n_a} \end{pmatrix} \begin{pmatrix} 0 & & & & \\ & n_1^{-1} & & & \\ & & n_2^{-1} & & \\ & & & \ddots & \\ & & & & n_a^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{1}_{n_1} & & & \\ \mathbf{1}_{n_2} & & \mathbf{1}_{n_2} & & \\ \vdots & & & \ddots & \\ \mathbf{1}_{n_a} & & & & \mathbf{1}_{n_a} \end{pmatrix}^T \\
&= \begin{pmatrix} n_1^{-1}\mathbf{1}_{n_1}\mathbf{1}_{n_1}^T & & & \\ & n_2^{-1}\mathbf{1}_{n_2}\mathbf{1}_{n_2}^T & & \\ & & \ddots & \\ & & & n_a^{-1}\mathbf{1}_{n_a}\mathbf{1}_{n_a}^T \end{pmatrix}.
\end{aligned}
$$

Define

$$
\begin{aligned}
\boldsymbol{A}_1 &= \boldsymbol{P_1} \\
\boldsymbol{A}_2 &= \boldsymbol{P_X} - \boldsymbol{P_1} \\
\boldsymbol{A}_3 &= \boldsymbol{I}_n - \boldsymbol{P_X}.
\end{aligned}
$$

with corresponding quadratic forms

$$
\begin{aligned}
\text{SSM} &= \boldsymbol{y}^T\boldsymbol{A}_1\boldsymbol{y} = n\bar{y}^2, \quad (1/\sigma^2)\boldsymbol{y}^T\boldsymbol{A}_1\boldsymbol{y} \sim \chi^2_1(\phi_1), \\
\phi_1 &= \frac{1}{2\sigma^2}(\boldsymbol{X}\boldsymbol{b})^T\boldsymbol{A}_1(\boldsymbol{X}\boldsymbol{b}) = \frac{n(\mu+\bar{\alpha})^2}{2\sigma^2} \\
\text{SSA}_{\text{cfm}} &= \boldsymbol{y}^T\boldsymbol{A}_2\boldsymbol{y} = \sum_{i=1}^{a} n_i\bar{y}_{i\cdot}^2 - n\bar{y}^2, \quad (1/\sigma^2)\boldsymbol{y}^T\boldsymbol{A}_2\boldsymbol{y} \sim \chi^2_{a-1}(\phi_2), \\
\phi_2 &= \frac{1}{2\sigma^2}(\boldsymbol{X}\boldsymbol{b})^T\boldsymbol{A}_2(\boldsymbol{X}\boldsymbol{b}) = \frac{\sum_{i=1}^{a} n_i(\alpha_i-\bar{\alpha})^2}{2\sigma^2} \\
\text{SSE} &= \boldsymbol{y}^T\boldsymbol{A}_3\boldsymbol{y} = \boldsymbol{y}^T\boldsymbol{y} - n\bar{y}^2 = \sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij}-\bar{y})^2, \quad (1/\sigma^2)\boldsymbol{y}^T\boldsymbol{A}_3\boldsymbol{y} \sim \chi^2_{n-a}(\phi_3), \\
\phi_3 &= \frac{1}{2\sigma^2}(\boldsymbol{X}\boldsymbol{b})^T\boldsymbol{A}_3(\boldsymbol{X}\boldsymbol{b}) = 0.
\end{aligned}
$$

We have the classical ANOVA table

Source	df	Projection	SS	Noncentrality
Mean	1	$\boldsymbol{P_1}$	SSM=$n\bar{y}^2$	$\frac{1}{2\sigma^2}n(\mu+\bar{\alpha})^2$
Group	$a-1$	$\boldsymbol{P_X}-\boldsymbol{P_1}$	$\text{SSA}_{\text{cfm}}=\sum_{i=1}^{a} n_i\bar{y}_i^2-n\bar{y}^2$	$\frac{1}{2\sigma^2}\sum_{i=1}^{a} n_i(\alpha_i-\bar{\alpha})^2$
Error	$n-a$	$\boldsymbol{I}-\boldsymbol{P_X}$	SSE=$\sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2$	0
Total	n	$\boldsymbol{I}$	SST=$\sum_i\sum_j y_{ij}^2$	$\frac{1}{2\sigma^2}\sum_{i=1}^{a} n_i(\mu+\alpha_i)^2$

Estimation under the normal Gauss-Markov model (JM 6.2)

Assume normal Gauss-Markov model $\boldsymbol{y}\sim N_n(\boldsymbol{Xb},\sigma^2\boldsymbol{I})$.

- The density of $\boldsymbol{y}\sim N_n(\boldsymbol{Xb},\sigma^2\boldsymbol{I})$ is

$$\begin{aligned} f(\boldsymbol{y}\mid\boldsymbol{b},\sigma^2) &= \frac{1}{(2\pi\sigma^2)^{n/2}}e^{-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{Xb})^T(\boldsymbol{y}-\boldsymbol{Xb})} \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}}e^{-\frac{1}{2\sigma^2}\boldsymbol{y}^T\boldsymbol{y}+\frac{1}{\sigma^2}\boldsymbol{y}^T\boldsymbol{Xb}-\frac{1}{2\sigma^2}\boldsymbol{b}^T\boldsymbol{X}^T\boldsymbol{Xb}}. \end{aligned}$$

- Recall Theorem 6.2.25 of Casella and Berger (2001).

  Let $\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n$ be iid observations from an exponential family with pdf or pmf of the form

$$f(\boldsymbol{x}|\boldsymbol{\theta})=h(\boldsymbol{x})c(\boldsymbol{\theta})\exp\left(\sum_{j=1}^{k}w(\theta_j)t_j(\boldsymbol{x})\right),$$

  where $\boldsymbol{\theta}=(\theta_1,\ldots,\theta_k)$. Then the statistics

$$T(\boldsymbol{X})=\left(\sum_{i=1}^{n}t_1(\boldsymbol{X}_i),\ldots,\sum_{i=1}^{n}t_k(\boldsymbol{X}_i)\right)$$

  is complete as long as the parameter space $\boldsymbol{\Theta}$ contains an open set in $\mathbb{R}^k$.

- Also by Theorem 6.2.28 of Casella and Berger (2001):

  If a minimal sufficient statistic exists, then any complete statistic is also a minimal sufficient statistic.

- From the density of the normal Gauss-Markov model, we see

$$T(\boldsymbol{y})=(\boldsymbol{y}^T\boldsymbol{y},\boldsymbol{X}^T\boldsymbol{y})$$

  is a complete and minimal sufficient statistic for $(\sigma^2,\boldsymbol{b})$.

- Recall the Rao-Blackwell theorem (Casella and Berger, 2001, Theorem 7.3.17). Any function of a sufficient statistic has the smallest variance among all unbiased estimators of its expectation.

- Under the normal Gauss-Markov model $\boldsymbol{y} \sim N_n(\boldsymbol{X}\boldsymbol{b}, \sigma^2\boldsymbol{I})$, the least squares estimator $\boldsymbol{\Lambda}\hat{\boldsymbol{b}}$ of an estimable function $\boldsymbol{\Lambda}\boldsymbol{b}$ has the smallest variance among *all* unbiased estimators.

  Remark: The least squares estimator $\Lambda\hat{\boldsymbol{b}}$ is also called the MVUE (minimum variance unbiased estimator) under normal assumption.

  Proof. The least squares estimator $\boldsymbol{\Lambda}\hat{\boldsymbol{b}} = \boldsymbol{\Lambda}(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{y}$ is a function of the complete sufficient statistic $T(\boldsymbol{y}) = (\boldsymbol{y}^T\boldsymbol{y}, \boldsymbol{X}^T\boldsymbol{y})$ and has expectation $\mathrm{E}(\boldsymbol{\Lambda}\hat{\boldsymbol{b}}) = \boldsymbol{\Lambda}\boldsymbol{b}$. □

- (MLE) Under the normal Gauss-Markov model $\boldsymbol{y} \sim N_n(\boldsymbol{X}\boldsymbol{b}, \sigma^2\boldsymbol{I})$, the maximum likelihood estimator (MLE) of $(\boldsymbol{b}, \sigma^2)$ is

$$\left(\hat{\boldsymbol{b}}, \frac{\mathrm{SSE}}{n}\right),$$

  where $\hat{\boldsymbol{b}}$ is any least squares solution (solution to the normal equation) and $\mathrm{SSE} = \boldsymbol{y}^T(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{y}$.

  Remark: The MLE for σ^2 is biased.

  Proof. To maximize the density

$$f(\boldsymbol{y} \mid \boldsymbol{b}, \sigma^2) \;=\; \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{b})^T(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{b})}$$

  is equivalent to maximizing the log-likelihood

$$L(\boldsymbol{b}, \sigma^2|\boldsymbol{y}) = \ln f(\boldsymbol{y} \mid \boldsymbol{b}, \sigma^2) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}).$$

  For any σ^2, the quadratic form is minimized by any least squares solution (as we showed before) and optimal value SSE. To find the maximizing σ^2, we set the derivative to 0

$$-\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\mathrm{SSE} = 0,$$

  which yields $\hat{\sigma}^2 = \mathrm{SSE}/n$. □

- Under the normal Gauss-Markov model $\boldsymbol{y} \sim N_n(\boldsymbol{X}\boldsymbol{b}, \sigma^2\boldsymbol{I})$, the MLE of an estimable function $\boldsymbol{\Lambda}\boldsymbol{b}$ is $\boldsymbol{\Lambda}\hat{\boldsymbol{b}}$, where $\hat{\boldsymbol{b}}$ is any least squares solution.

  Proof. If follows from the invariance property of the MLE (Casella and Berger, 2001, Theorem 7.2.10). □

15 Lecture 15: Oct 28

Announcement

- HW5 due today.
- HW6 posted today and due next Monday.
- HW7 will be posted this Wed and due next Wed.
- HW5-7 are covered in Midterm 2.

Last time

- Cochran's theorem: Let $\boldsymbol{y} \sim N_n(\boldsymbol{\mu}, \sigma^2 \boldsymbol{I}_n)$ and $\boldsymbol{A}_i$, $i = 1, \ldots, k$, be symmetric idempotent matrix with rank s_i. If $\sum_{i=1}^k \boldsymbol{A}_i = \boldsymbol{I}_n$, then $(1/\sigma^2)\boldsymbol{y}^T \boldsymbol{A}_i \boldsymbol{y}$ are independent $\chi^2_{s_i}(\phi_i)$, with $\phi_i = \frac{1}{2\sigma^2}\boldsymbol{\mu}^T \boldsymbol{A}_i \boldsymbol{\mu}$ and $\sum_{i=1}^k s_i = n$.
- Application of Cochran's theorem to the one-way ANOVA model.
- Statistical inference for normal Gauss-Markov model $\boldsymbol{y} \sim N_n(\boldsymbol{X}\boldsymbol{b}, \sigma^2 \boldsymbol{I})$.
  - $(\boldsymbol{y}^T\boldsymbol{y}, \boldsymbol{X}^T\boldsymbol{y})$ is a complete and minimal sufficient statistic for $(\sigma^2, \boldsymbol{b})$.
  - The least squares estimator $\boldsymbol{\Lambda}\hat{\boldsymbol{b}}$ of an estimable function $\boldsymbol{\Lambda}\boldsymbol{b}$ is not only MVAUE, but also MVUE under normal Gauss-Markov assumption.
  - MLE of $(\boldsymbol{b}, \sigma^2)$ is $(\hat{\boldsymbol{b}}, \text{SSE}/n)$, where $\hat{\boldsymbol{b}}$ is any least squares solution and $\text{SSE} = \boldsymbol{y}^T(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{y}$.
  - MLE of any estimable function $\boldsymbol{\Lambda}\boldsymbol{b}$ is $\boldsymbol{\Lambda}\hat{\boldsymbol{b}}$.

Today

- Testing general linear hypothesis: first principles test
- Testing general linear hypothesis: LRT

General linear hypothesis (JM 6.3)

We assume normal Gauss-Markov model $\boldsymbol{y} \sim N_n(\boldsymbol{X}\boldsymbol{b}, \sigma^2\boldsymbol{I})$.

- In many applications, we are interested in testing a genera linear hypothesis

$$H_0 : \boldsymbol{R}\boldsymbol{b} = \boldsymbol{r} \quad \text{vs} \quad H_A : \boldsymbol{R}\boldsymbol{b} \neq \boldsymbol{r},$$

  where $\boldsymbol{R} \in \mathbb{R}^{s\times p}$ and $\boldsymbol{r} \in \mathbb{R}^s$.

- Examples
  - $H_A : b_j = 0$
  - $H_A : b_1 = b_2 = b_3 = 0$
  - $H_A : b_2 = b_3 = b_4$
  - $H_A : b_1 + b_3 = 1,\ b_2 = 3$
  - $H_A : \boldsymbol{b} \in \mathcal{C}(\boldsymbol{B})$

- We say a general linear hypothesis $H_0 : \boldsymbol{R}\boldsymbol{b} = \boldsymbol{r}$ is *testable* if $\boldsymbol{R}$ has full row rank s and $\boldsymbol{R}\boldsymbol{b}$ is estimable.

- Example (testing the interaction in two-way ANOVA): Consider the two-way ANOVA model with interaction

$$y_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}.$$

  In Lecture 6 (Sep 18), we showed that the interaction effects

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

  are estimable. There are $(a-1)(b-1)$ linearly independent interaction effects. Thus $s = (a-1)(b-1)$ in the linear hypothesis $H_0 : \boldsymbol{R}\boldsymbol{b} = \boldsymbol{0}_s$ for testing interaction effects.

  Recall that the design matrix for a two-way ANOVA with interaction has rank ab and that for a two-way ANOVA without interaction has rank $a+b-1$. The difference is $ab - (a+b-1) = (a-1)(b-1)$.

First principles test for a general linear hypothesis (JM 6.3)

- Consider testing a testable general linear hypothesis $\boldsymbol{Rb} = \boldsymbol{r}$ under the normal Gauss-Markov model $\boldsymbol{y} \sim N_n(\boldsymbol{Xb}, \sigma^2\boldsymbol{I})$.

- Since $\boldsymbol{Rb}$ is estimable, the MVAUE (also the MVUE and MLE)

$$\boldsymbol{R}\hat{\boldsymbol{b}} = \boldsymbol{R}(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{y}$$

  is a multivariate normal with mean

$$\boldsymbol{R}(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{b} = \boldsymbol{Rb}$$

  and covariance

$$\sigma^2\boldsymbol{R}(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{R}^T = \sigma^2\boldsymbol{R}(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{R}^T =: \sigma^2\boldsymbol{H}.$$

  Hence

$$\boldsymbol{R}\hat{\boldsymbol{b}} - \boldsymbol{r} \sim N_s(\boldsymbol{Rb} - \boldsymbol{r}, \sigma^2\boldsymbol{H}).$$

- Note

$$\begin{aligned} & \operatorname{rank}(\boldsymbol{H}) = \operatorname{rank}(\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{R}^T) \\ \geq\ & \operatorname{rank}(\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{R}^T) = \operatorname{rank}(\boldsymbol{R}^T) = s, \end{aligned}$$

  thus $\boldsymbol{H}$ is non-singular.

- By previous result (Lecture 12),

$$(\boldsymbol{R}\hat{\boldsymbol{b}} - \boldsymbol{r})^T(\sigma^2\boldsymbol{H})^{-1}(\boldsymbol{R}\hat{\boldsymbol{b}} - \boldsymbol{r}) \sim \chi^2_s(\phi)$$

  with noncentrality parameter

$$\phi = \frac{1}{2}(\boldsymbol{Rb} - \boldsymbol{r})^T(\sigma^2\boldsymbol{H})^{-1}(\boldsymbol{Rb} - \boldsymbol{r}).$$

- Since

$$[\boldsymbol{R}(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T](\sigma^2\boldsymbol{I})(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}}) = \boldsymbol{0},$$

  the quadratic form $(\boldsymbol{R}\hat{\boldsymbol{b}} - \boldsymbol{r})^T(\sigma^2\boldsymbol{H})^{-1}(\boldsymbol{R}\hat{\boldsymbol{b}} - \boldsymbol{r})$ is independent of the quadratic form $\text{SSE} = \boldsymbol{y}^T(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{y}$.

- Therefore the ratio is a F random variable

$$F = \frac{(\boldsymbol{R}\hat{\boldsymbol{b}} - \boldsymbol{r})^T \boldsymbol{H}^{-1} (\boldsymbol{R}\hat{\boldsymbol{b}} - \boldsymbol{r})/s}{\text{SSE}/(n-r)} \sim F_{s,n-r}(\phi).$$

  Under $H_0 : \boldsymbol{R}\boldsymbol{b} = \boldsymbol{r}$, $\phi = 0$ and the ratio is a central F distribution.
  Under $H_A : \boldsymbol{R}\boldsymbol{b} \neq \boldsymbol{r}$, $\phi > 0$ and the ratio is a noncentral F distribution.

- Consider one-way ANOVA model

$$y_{ij} = \mu + \alpha_i + e_{ij}.$$

  We want test the hypothesis that all effects α_i are equal. The two hypotheses

$$H_0 : \alpha_1 - \alpha_2 = 0, \alpha_1 - \alpha_3 = 0, \ldots, \alpha_a - \alpha_a = 0$$

  and

$$H_0 : \alpha_1 - \alpha_2 = 0, \alpha_2 - \alpha_3 = 0, \ldots, \alpha_{a-1} - \alpha_a = 0$$

  are two logically equivalent ways of expressing the test. Will they lead to the same test procedure?

- Two linear hypotheses $\boldsymbol{R}_1\boldsymbol{b} = \boldsymbol{r}_1$ and $\boldsymbol{R}_2\boldsymbol{b} = \boldsymbol{r}_2$ are *equivalent* if

$$\{\boldsymbol{b} : \boldsymbol{R}_1\boldsymbol{b} = \boldsymbol{r}_1\} = \{\boldsymbol{b} : \boldsymbol{R}_2\boldsymbol{b} = \boldsymbol{r}_2\}.$$

- (Invariance of the first principles test under equivalent linear hypotheses) Suppose two linear hypotheses $\boldsymbol{R}_1\boldsymbol{b} = \boldsymbol{r}_1$ and $\boldsymbol{R}_2\boldsymbol{b} = \boldsymbol{r}_2$ are *equivalent* and both $\boldsymbol{R}_1$ and $\boldsymbol{R}_2$ have full row rank s. The points in the set $\{\boldsymbol{b} : \boldsymbol{R}_1\boldsymbol{b} = \boldsymbol{r}_1\}$ are characterized by

$$\boldsymbol{R}_1^- \boldsymbol{r}_1 + (\boldsymbol{I} - \boldsymbol{R}_1^- \boldsymbol{R}_1)\boldsymbol{q}.$$

  Now

$$\boldsymbol{R}_2[\boldsymbol{R}_1^- \boldsymbol{r}_1 + (\boldsymbol{I} - \boldsymbol{R}_1^- \boldsymbol{R}_1)\boldsymbol{q}] = \boldsymbol{r}_2$$

  for all $\boldsymbol{q}$ implies that $\boldsymbol{R}_2\boldsymbol{R}_1^- \boldsymbol{r}_1 = \boldsymbol{r}_2$ (taking $\boldsymbol{q} = \boldsymbol{0}$) and

$$\boldsymbol{R}_2(\boldsymbol{I} - \boldsymbol{R}_1^- \boldsymbol{R}_1) = \boldsymbol{0}_{s \times p}.$$

Thus $\mathcal{C}(\boldsymbol{R}_2^T) \subset \mathcal{C}(\boldsymbol{R}_1^T)$. Reversing the roles of $(\boldsymbol{R}_1, \boldsymbol{b}_1)$ and $(\boldsymbol{R}_2, \boldsymbol{b}_2)$ shows that $\mathcal{C}(\boldsymbol{R}_1^T) \subset \mathcal{C}(\boldsymbol{R}_2^T)$. Hence $\mathcal{C}(\boldsymbol{R}_1^T) = \mathcal{C}(\boldsymbol{R}_2^T)$.

$\boldsymbol{R}_2 = \boldsymbol{T}\boldsymbol{R}_1$ for some transformation matrix $\boldsymbol{T} \in \mathbb{R}^{s\times s}$. $\boldsymbol{T}$ must be non-singular, otherwise $\text{rank}(\boldsymbol{R}_2) \leq \text{rank}(\boldsymbol{T}) < s$, a contradiction with $\boldsymbol{R}_2$ has full row rank. Also we have

$$\boldsymbol{R}_2\boldsymbol{R}_1^-\boldsymbol{r}_1 = \boldsymbol{T}\boldsymbol{R}_1\boldsymbol{R}_1^-\boldsymbol{r}_1 = \boldsymbol{T}\boldsymbol{r}_1 = \boldsymbol{r}_2.$$

Thus the numerator of the F statistic for testing $H_0 : \boldsymbol{R}_2\boldsymbol{b} = \boldsymbol{r}_2$

$$\begin{aligned} & (\boldsymbol{R}_2\boldsymbol{b} - \boldsymbol{r}_2)^T[\boldsymbol{R}_2(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{R}_2^T]^{-1}(\boldsymbol{R}_2\boldsymbol{b} - \boldsymbol{r}_2) \\ = \; & (\boldsymbol{R}_1\boldsymbol{b} - \boldsymbol{r}_1)^T\boldsymbol{T}^T[\boldsymbol{T}\boldsymbol{R}_1(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{R}_1^T\boldsymbol{T}^T]^{-1}\boldsymbol{T}(\boldsymbol{R}_1\boldsymbol{b} - \boldsymbol{r}_1) \\ = \; & (\boldsymbol{R}_1\boldsymbol{b} - \boldsymbol{r}_1)^T[\boldsymbol{R}_1(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{R}_1^T]^{-1}(\boldsymbol{R}_1\boldsymbol{b} - \boldsymbol{r}_1). \end{aligned}$$

is same as that for testing $H_0 : \boldsymbol{R}_1\boldsymbol{b} = \boldsymbol{r}_1$.

Likelihood ratio test (LRT) for a general linear hypothesis (JM 6.4-6.5)

- We derive the LRT for testing the general linear hypothesis $H_0 : \boldsymbol{R}\boldsymbol{b} = \boldsymbol{r}$.
- The log-likelihood function is

$$\begin{aligned} L(\boldsymbol{b}, \sigma^2|\boldsymbol{y}) &= -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}) \\ &= -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}Q(\boldsymbol{b}), \end{aligned}$$

where $Q(\boldsymbol{b}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b})$. As derived in the last lecture, for any $\boldsymbol{b}$, the maximizing σ^2 is given by

$$\hat{\sigma}^2 = \frac{Q(\boldsymbol{b})}{n}$$

and the resultant log-likelihood is

$$-\frac{n}{2}\ln[2\pi Q(\boldsymbol{b})/n] - \frac{n}{2}.$$

We only need to derive the constrained and unconstrained MLE for $\boldsymbol{b}$. That is the constrained and unconstrained maximizers of $Q(\boldsymbol{b})$.

- The unconstrained MLE is the least squares solution $\hat{\boldsymbol{b}}$, i.e., any solution to the normal equation.

- To derive the constrained MLE, we need to solve the minimization problem

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{b})^T(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{b}) \\ \text{subject to} & \boldsymbol{R}\boldsymbol{b}=\boldsymbol{r}. \end{array}$$

  Setting the gradient of the Lagrangian function

$$\frac{1}{2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{b})^T(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{b})+\boldsymbol{\lambda}^T(\boldsymbol{R}\boldsymbol{b}-\boldsymbol{r})$$

  to zero leads to the linear equation

$$\begin{pmatrix} \boldsymbol{X}^T\boldsymbol{X} & \boldsymbol{R}^T \\ \boldsymbol{R} & \boldsymbol{0} \end{pmatrix}\begin{pmatrix} \boldsymbol{b} \\ \boldsymbol{\lambda} \end{pmatrix}=\begin{pmatrix} \boldsymbol{X}^T\boldsymbol{y} \\ \boldsymbol{r} \end{pmatrix}.$$

  This equation is always consistent since $\boldsymbol{X}^T\boldsymbol{y}\in\mathcal{C}(\boldsymbol{X}^T)=\mathcal{C}(\boldsymbol{X}^T\boldsymbol{X})$ and $\boldsymbol{r}=\boldsymbol{R}\boldsymbol{b}\in\mathcal{C}(\boldsymbol{R})$. By HW4 Q5(e), a generalized inverse of bordered Gramian matrix is

$$\begin{aligned} & \begin{pmatrix} \boldsymbol{X}^T\boldsymbol{X} & \boldsymbol{R}^T \\ \boldsymbol{R} & \boldsymbol{0} \end{pmatrix}^- \\ = & \begin{pmatrix} (\boldsymbol{X}^T\boldsymbol{X})^- - (\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{R}^T\boldsymbol{H}^{-1}\boldsymbol{R}(\boldsymbol{X}^T\boldsymbol{X})^- & (\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{R}^T\boldsymbol{H}^{-1} \\ \boldsymbol{H}^{-1}\boldsymbol{R}(\boldsymbol{X}^T\boldsymbol{X})^- & -\boldsymbol{H}^{-1} \end{pmatrix}, \end{aligned}$$

  where $\boldsymbol{H}=\boldsymbol{R}(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{R}^T$ is non-singular under the assumption $\boldsymbol{R}$ has full row rank. Therefore a solution is

$$\begin{aligned} \hat{\boldsymbol{b}}_0 &= [(\boldsymbol{X}^T\boldsymbol{X})^- - (\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{R}^T\boldsymbol{H}^{-1}\boldsymbol{R}(\boldsymbol{X}^T\boldsymbol{X})^-]\boldsymbol{X}^T\boldsymbol{y}+(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{R}^T\boldsymbol{H}^{-1}\boldsymbol{r} \\ &= \hat{\boldsymbol{b}}-(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{R}^T\boldsymbol{H}^{-1}\boldsymbol{R}\hat{\boldsymbol{b}}+(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{R}^T\boldsymbol{H}^{-1}\boldsymbol{r} \\ &= \hat{\boldsymbol{b}}-(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{R}^T\boldsymbol{H}^{-1}(\boldsymbol{R}\hat{\boldsymbol{b}}-\boldsymbol{r}). \end{aligned}$$

- The change in the SSEs of constrained and unconstrained models is

$$\begin{aligned} & Q(\hat{\boldsymbol{b}}_0)-Q(\hat{\boldsymbol{b}}) \\ = & (\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{b}}_0)^T(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{b}}_0)-(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{b}})^T(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{b}}) \\ = & 2\boldsymbol{y}^T\boldsymbol{X}(\hat{\boldsymbol{b}}-\hat{\boldsymbol{b}}_0)-2\hat{\boldsymbol{b}}\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{b}}+2\hat{\boldsymbol{b}}^T\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{b}}_0+(\hat{\boldsymbol{b}}_0-\hat{\boldsymbol{b}})^T\boldsymbol{X}^T\boldsymbol{X}(\hat{\boldsymbol{b}}_0-\hat{\boldsymbol{b}}) \\ = & 2(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{b}})^T\boldsymbol{X}(\hat{\boldsymbol{b}}-\hat{\boldsymbol{b}}_0)+(\hat{\boldsymbol{b}}_0-\hat{\boldsymbol{b}})^T\boldsymbol{X}^T\boldsymbol{X}(\hat{\boldsymbol{b}}_0-\hat{\boldsymbol{b}}) \\ = & 0+(\boldsymbol{R}\hat{\boldsymbol{b}}-\boldsymbol{r})^T\boldsymbol{H}^{-1}\boldsymbol{R}(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{R}^T\boldsymbol{H}^{-1}(\boldsymbol{R}\hat{\boldsymbol{b}}-\boldsymbol{r}) \\ = & (\boldsymbol{R}\hat{\boldsymbol{b}}-\boldsymbol{r})^T\boldsymbol{H}^{-1}(\boldsymbol{R}\hat{\boldsymbol{b}}-\boldsymbol{r}). \end{aligned}$$

- The LRT rejects when

$$\frac{\max_{\Omega_0} L(\boldsymbol{b}, \sigma^2)}{\max_{\Omega} L(\boldsymbol{b}, \sigma^2)} = \left(\frac{Q(\hat{\boldsymbol{b}}_0)}{Q(\hat{\boldsymbol{b}})} \right)^{-n/2}$$

  is small, or equivalently when

$$\frac{Q(\hat{\boldsymbol{b}}_0) - Q(\hat{\boldsymbol{b}})}{Q(\hat{\boldsymbol{b}})}$$

  is large, or equivalently when

$$\frac{(\boldsymbol{R}\hat{\boldsymbol{b}} - \boldsymbol{r})^T \boldsymbol{H}^{-1} (\boldsymbol{R}\hat{\boldsymbol{b}} - \boldsymbol{r})/s}{\text{SSE}/(n-r)}$$

  is large.

  Therefore the LRT is same as the first principles test!

16 Lecture 16: Oct 30

Announcement

- HW4 returned.
- HW4 Q5(e) mystery resolved: `http://hua-zhou.github.io/teaching/st552-2013fall/2013/10/29/hw4-Q5e-mystery-resolved.html`
- HW6 due next Monday (?)
- No HW session this afternoon.
- HW7 posted today and due next Wed.
- Midterm 2 covers Chapter 5-7 and HW 5-7.

Last time

- Testing general linear hypothesis $H_0 : \boldsymbol{R}\boldsymbol{b} = \boldsymbol{r}$: first principles test

$$F = \frac{(\boldsymbol{R}\hat{\boldsymbol{b}} - \boldsymbol{r})^T \boldsymbol{H}^{-1} (\boldsymbol{R}\hat{\boldsymbol{b}} - \boldsymbol{r})/s}{\text{SSE}/(n-r)} \sim F_{s,n-r}(\phi).$$

- Testing general linear hypothesis: LRT = first principles test
- The first principles test (and LRT) is same under equivalent linear hypotheses.

Today

- t test and confidence interval
- simultaneous confidence intervals and multiple comparison

t test and confidence interval

Assume the normal Gauss-Markov model $\boldsymbol{y} \sim N_n(\boldsymbol{X}\boldsymbol{b}, \sigma^2 \boldsymbol{I})$.

- If $\boldsymbol{r}^T\boldsymbol{b}$ is estimable, then the least squares estimator (MVAUE, MVUE, MLE) is

$$\boldsymbol{r}^T\hat{\boldsymbol{b}} \sim N(\boldsymbol{r}^T\boldsymbol{b}, \sigma^2\boldsymbol{r}^T(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{r}),$$

or equivalently

$$\frac{\boldsymbol{r}^T\hat{\boldsymbol{b}} - \boldsymbol{r}^T\boldsymbol{b}}{\sigma\sqrt{\boldsymbol{r}^T(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{r}}} \sim N(0,1).$$

- We estimate σ^2 by the usual $\hat{\sigma}^2 = \text{SSE}/(n-r)$. Then

$$\frac{\boldsymbol{r}^T\hat{\boldsymbol{b}} - \boldsymbol{r}^T\boldsymbol{b}}{\hat{\sigma}\sqrt{\boldsymbol{r}^T(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{r}}} = \frac{(\boldsymbol{r}^T\hat{\boldsymbol{b}} - \boldsymbol{r}^T\boldsymbol{b})/\sqrt{\sigma^2\boldsymbol{r}^T(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{r}}}{\text{SSE}/[\sigma^2(n-r)]} \sim t_{n-r}.$$

- A $(1-\alpha)$ confidence interval for $\boldsymbol{r}^T\boldsymbol{b}$ is given by

$$\boldsymbol{r}^T\hat{\boldsymbol{b}} \pm t_{n-r,\alpha/2}\hat{\sigma}\sqrt{\boldsymbol{r}^T(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{r}},$$

where $t_{n-r,\alpha/2}$ is the critical value of a central t distribution with $n-r$ degrees of freedom. We are assured that

$$\mathbf{P}(\boldsymbol{r}^T\hat{\boldsymbol{b}} - t_{n-r,\alpha/2}\hat{\sigma}\sqrt{\boldsymbol{r}^T(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{r}} \le \boldsymbol{r}^T\boldsymbol{b} \le \boldsymbol{r}^T\hat{\boldsymbol{b}} + t_{n-r,\alpha/2}\hat{\sigma}\sqrt{\boldsymbol{r}^T(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{r}}) = 1-\alpha.$$

- For two-sided testing

$$H_0 : \boldsymbol{r}^T\boldsymbol{b} = r \quad \text{vs} \quad H_A : \boldsymbol{r}^T\boldsymbol{b} \neq r.$$

The t test rejects H_0 when

$$\left|\frac{\boldsymbol{r}^T\hat{\boldsymbol{b}} - r}{\hat{\sigma}\sqrt{\boldsymbol{r}^T(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{r}}}\right| > t_{n-r,\alpha/2}.$$

One-sided tests are carried out similarly using critical value $t_{n-r,\alpha}$.

Simultaneous confidence intervals/multiple comparison (JM 6.6)

Assume the normal Gauss-Markov model $\boldsymbol{y} \sim N_n(\boldsymbol{X}\boldsymbol{b}, \sigma^2\boldsymbol{I})$.

- Now we consider estimating an estimable function $\boldsymbol{Rb}$ where $\boldsymbol{R}$ has full row rank s. The least squares estimator (also MVAUE, MVUE and MLE) is

$$\boldsymbol{R}\hat{\boldsymbol{b}} \sim N_s(\boldsymbol{Rb}, \sigma^2 \boldsymbol{H}),$$

where

$$\boldsymbol{H} = \boldsymbol{R}(\boldsymbol{X}^T\boldsymbol{X})^- \boldsymbol{R}^T.$$

- The $1-\alpha$ confidence interval for the i-th linear function $\boldsymbol{r}_i^T\boldsymbol{b}$ is $[l_i, u_i]$, where

$$\begin{aligned} l_i &= \boldsymbol{r}_i^T\hat{\boldsymbol{b}} - t_{n-r,\alpha/2}\hat{\sigma}\sqrt{\boldsymbol{r}_i^T(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{r}_i} = \boldsymbol{r}_i^T\hat{\boldsymbol{b}} - t_{n-r,\alpha/2}\hat{\sigma}\sqrt{h_{ii}} \\ u_i &= \boldsymbol{r}_i^T\hat{\boldsymbol{b}} + t_{n-r,\alpha/2}\hat{\sigma}\sqrt{\boldsymbol{r}_i^T(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{r}_i} = \boldsymbol{r}_i^T\hat{\boldsymbol{b}} + t_{n-r,\alpha/2}\hat{\sigma}\sqrt{h_{ii}}. \end{aligned}$$

That means

$$\mathbf{P}(l_i \le \boldsymbol{r}_i^T\boldsymbol{b} \le u_i) = 1-\alpha \quad \text{for } i = 1, \ldots, s,$$

and thus

$$\mathbf{P}(l_i \le \boldsymbol{r}_i^T\boldsymbol{b} \le u_i \text{ for all } i = 1, \ldots, s) \le 1-\alpha,$$

where the inequality can be strict. In other words, our joint confidence intervals don't have the right coverage ☹

Bonferroni correction

- (Bonferroni inequalities) Let E_i be a collection of events. Then

$$\begin{aligned} \mathbf{P}(\cup_j E_j) &= \mathbf{P}(\text{at least one } E_j) \le \sum_j \mathbf{P}(E_j) \\ \mathbf{P}(\cap_j E_j) &= \mathbf{P}(\text{all } E_j) \ge 1 - \sum_j \mathbf{P}(E_j^c). \end{aligned}$$

- Bonferroni method for simultaneous confidence intervals. Set

$$\begin{aligned} l_i &= \boldsymbol{r}_i^T\hat{\boldsymbol{b}} - t_{n-r,\alpha/(2s)}\hat{\sigma}\sqrt{\boldsymbol{r}_i^T(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{r}_i} = \boldsymbol{r}_i^T\hat{\boldsymbol{b}} - t_{n-r,\alpha/2}\hat{\sigma}\sqrt{h_{ii}} \\ u_i &= \boldsymbol{r}_i^T\hat{\boldsymbol{b}} + t_{n-r,\alpha/(2s)}\hat{\sigma}\sqrt{\boldsymbol{r}_i^T(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{r}_i} = \boldsymbol{r}_i^T\hat{\boldsymbol{b}} + t_{n-r,\alpha/2}\hat{\sigma}\sqrt{h_{ii}}. \end{aligned}$$

Then

$$\begin{aligned} & \mathbf{P}(l_i \le \boldsymbol{r}_i^T \boldsymbol{b} \le u_i \text{ for all } i = 1, \ldots, s) \\ \ge \; & 1 - \sum_{i=1}^{s} \mathbf{P}(\boldsymbol{r}_i^T \boldsymbol{b} \notin [l_i, u_i]) \\ = \; & 1 - s(\alpha/s) \\ = \; & 1 - \alpha. \end{aligned}$$

Now our joint confidence intervals have the right coverage ☺

- Obviously we can utilize these simultaneous intervals for testing $H_0 : \boldsymbol{Rb} = \boldsymbol{r}$. Simply reject H_0 when $|\boldsymbol{r}_i^T \hat{\boldsymbol{b}} - r_i| > t_{n-r,\alpha/2s} \hat{\sigma} \sqrt{h_{ii}}$ for any i.

Scheffé's method

- Scheffé's method constructs simultaneous confidence intervals

$$[\boldsymbol{u}^T \boldsymbol{R}\hat{\boldsymbol{b}} - c\hat{\sigma}\sqrt{\boldsymbol{u}^T \boldsymbol{H} \boldsymbol{u}}, \boldsymbol{u}^T \boldsymbol{R}\hat{\boldsymbol{b}} + c\hat{\sigma}\sqrt{\boldsymbol{u}^T \boldsymbol{H} \boldsymbol{u}}]$$

for *all* possible linear combinations $\boldsymbol{u}^T \boldsymbol{Rb}$ of the estimable function $\boldsymbol{Rb}$. The universal constant c needs to be chosen so that we have the right coverage.

- Observe that

$$\begin{aligned} & \mathbf{P}(\boldsymbol{u}^T \boldsymbol{Rb} \in [\boldsymbol{u}^T \boldsymbol{R}\hat{\boldsymbol{b}} - c\hat{\sigma}\sqrt{\boldsymbol{u}^T \boldsymbol{H} \boldsymbol{u}}, \boldsymbol{u}^T \boldsymbol{R}\hat{\boldsymbol{b}} + c\hat{\sigma}\sqrt{\boldsymbol{u}^T \boldsymbol{H} \boldsymbol{u}}] \text{ for all } \boldsymbol{u}) \\ = \; & \mathbf{P}\left(\frac{|\boldsymbol{u}^T \boldsymbol{R}\hat{\boldsymbol{b}} - \boldsymbol{u}^T \boldsymbol{Rb}|}{\hat{\sigma}\sqrt{\boldsymbol{u}^T \boldsymbol{H} \boldsymbol{u}}} \le c \text{ for all } \boldsymbol{u} \right) \\ = \; & \mathbf{P}\left(\max_{\boldsymbol{u}} \frac{|\boldsymbol{u}^T \boldsymbol{R}\hat{\boldsymbol{b}} - \boldsymbol{u}^T \boldsymbol{Rb}|}{\hat{\sigma}\sqrt{\boldsymbol{u}^T \boldsymbol{H} \boldsymbol{u}}} \le c \right) \\ = \; & \mathbf{P}\left(\max_{\boldsymbol{u}} \frac{[\boldsymbol{u}^T (\boldsymbol{R}\hat{\boldsymbol{b}} - \boldsymbol{Rb})]^2}{\hat{\sigma}^2 \boldsymbol{u}^T \boldsymbol{H} \boldsymbol{u}} \le c^2 \right) \\ = \; & \mathbf{P}\left(\frac{(\boldsymbol{R}\hat{\boldsymbol{b}} - \boldsymbol{Rb})^T \boldsymbol{H}^{-1} (\boldsymbol{R}\hat{\boldsymbol{b}} - \boldsymbol{Rb})}{\hat{\sigma}^2} \le c^2 \right) \\ = \; & \mathbf{P}\left(\frac{(\boldsymbol{R}\hat{\boldsymbol{b}} - \boldsymbol{Rb})^T \boldsymbol{H}^{-1} (\boldsymbol{R}\hat{\boldsymbol{b}} - \boldsymbol{Rb})}{s\hat{\sigma}^2} \le \frac{c^2}{s} \right). \end{aligned}$$

The key step (the fourth equality) is due to a generalized Cauchy-Schwartz inequality (HW7). Since

$$\frac{(\boldsymbol{R}\hat{\boldsymbol{b}} - \boldsymbol{R}\boldsymbol{b})^T \boldsymbol{H}^{-1} (\boldsymbol{R}\hat{\boldsymbol{b}} - \boldsymbol{R}\boldsymbol{b})}{s\hat{\sigma}^2} \sim F_{s,n-r},$$

we choose c such that $c^2/s = F_{s,n-r,\alpha}$, i.e.,

$$c = \sqrt{sF_{s,n-r,\alpha}}.$$

- Scheffé's simultaneous confidence intervals are just inversion of the F test (first principles test, LRT) and are invariant under equivalent linear hypotheses.

Tukey's method

- Consider the *balanced* one-way ANOVA model: $y_{ij} = \mu + \alpha_i + e_{ij}$, where $n_1 = \cdots = n_a = n$.
- We are interested constructing simultaneous confidence intervals for *all* pairwise differences of treatment effects $\alpha_i - \alpha_j$.
- The key observation is

$$\max_{i,j}(x_i - x_j) = \max_i x_i - \min_i x_i.$$

for any real numbers x_i.

- Under balanced one-way ANOVA

$$\bar{y}_{i\cdot} \sim N(\mu + \alpha_i, \sigma^2/n)$$

for $i = 1, \ldots, a$ and are independent. So

$$Z_i = \frac{\bar{y}_{i\cdot} - (\mu + \alpha_i)}{\sigma/\sqrt{n}}$$

are independent standard normals. Then

$$\begin{aligned} & \max_{i,j} \frac{(\bar{y}_{i\cdot} - \bar{y}_{j\cdot}) - (\alpha_i - \alpha_j)}{\sigma/\sqrt{n}} \\ = & \max_i \frac{\bar{y}_{i\cdot} - (\mu + \alpha_i)}{\sigma/\sqrt{n}} - \min_i \frac{\bar{y}_{i\cdot} - (\mu + \alpha_i)}{\sigma/\sqrt{n}} \\ = & \max_i Z_i - \min_i Z_i. \end{aligned}$$

Also recall that

$$\frac{\text{SSE}}{\sigma^2} = \frac{a(n-1)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{a(n-1)}$$

and is independent of Z_i.

- We tabulate the distribution of the random variable

$$W = \frac{\max_j Z_j - \min_j Z_j}{\sqrt{U/\nu}},$$

where Z_j are k iid standard normals and independent of $U \sim \chi^2_\nu$. Let $w_{k,\nu,\alpha}$ be the critical values of this distribution.

Then

$$\begin{aligned}
& 1-\alpha \\
= \;& \mathbf{P}\left(\frac{\max_i(\bar{y}_{i\cdot} - (\mu+\alpha_i))/(\sigma/\sqrt{n}) - \min_i(\bar{y}_{i\cdot} - (\mu+\alpha_i))/(\sigma/\sqrt{n})}{\sqrt{\hat{\sigma}^2/\sigma^2}} \le w_{a,a(n-1),\alpha} \right) \\
= \;& \mathbf{P}\left(\max_i(\bar{y}_{i\cdot} - (\mu+\alpha_i)) - \min_i(\bar{y}_{i\cdot} - (\mu+\alpha_i)) \le \frac{\hat{\sigma}}{\sqrt{n}} w_{a,a(n-1),\alpha} \right) \\
= \;& \mathbf{P}\left(\max_{i,j}[(\bar{y}_{i\cdot} - \bar{y}_{j\cdot}) - (\alpha_i - \alpha_j)] \le \frac{\hat{\sigma}}{\sqrt{n}} w_{a,a(n-1),\alpha} \right) \\
= \;& \mathbf{P}\left(|(\bar{y}_{i\cdot} - \bar{y}_{j\cdot}) - (\alpha_i - \alpha_j)| \le \frac{\hat{\sigma}}{\sqrt{n}} w_{a,a(n-1),\alpha} \text{ for all } i,j \right) \\
= \;& \mathbf{P}\left((\bar{y}_{i\cdot} - \bar{y}_{j\cdot}) - \frac{\hat{\sigma}}{\sqrt{n}} w_{a,a(n-1),\alpha} \le \alpha_i - \alpha_j \le (\bar{y}_{i\cdot} - \bar{y}_{j\cdot}) + \frac{\hat{\sigma}}{\sqrt{n}} w_{a,a(n-1),\alpha} \text{ for all } i,j \right).
\end{aligned}$$

Therefore

$$\left[(\bar{y}_{i\cdot} - \bar{y}_{j\cdot}) - \frac{\hat{\sigma}}{\sqrt{n}} w_{a,a(n-1),\alpha},\ (\bar{y}_{i\cdot} - \bar{y}_{j\cdot}) + \frac{\hat{\sigma}}{\sqrt{n}} w_{a,a(n-1),\alpha} \right]$$

are level $1-\alpha$ simultaneous confidence intervals for all pairwise differences of treatment means $\alpha_i - \alpha_j$.

- (Extension to contrasts) Any linear combination $\boldsymbol{u}^T\boldsymbol{\alpha} = \sum_{i=1}^a u_i\alpha_i$, where $\sum_{i=1}^a u_i = 0$, is called a *contrast*. Tukey's method can be extended to simultaneous confidence intervals for all contrasts in balanced one-way ANOVA (HW7)

$$\left[\sum_i u_i\bar{y}_{i\cdot} - \frac{\hat{\sigma}}{\sqrt{n}} w_{a,a(n-1),\alpha} \times \frac{1}{2}\sum_i |u_i|,\ \sum_i u_i\bar{y}_{i\cdot} + \frac{\hat{\sigma}}{\sqrt{n}} w_{a,a(n-1),\alpha} \times \frac{1}{2}\sum_i |u_i| \right].$$

Sequential sum of squares (JM 7.3)

Assume the normal Gauss-Markov model $\boldsymbol{y} \sim N_n(\boldsymbol{X}\boldsymbol{b}, \sigma^2\boldsymbol{I})$.

- Partition the design matrix as

$$\boldsymbol{X} = (\boldsymbol{X}_0, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_k),$$

  where $\boldsymbol{X}_j \in \mathbb{R}^{n \times p_j}$ and $\sum_{j=0}^{k} p_j = p$.

- Define

$$\begin{aligned} \boldsymbol{A}_0 &= \boldsymbol{P}_{\boldsymbol{X}_0} \\ \boldsymbol{A}_j &= \boldsymbol{P}_{(\boldsymbol{X}_0,\ldots,\boldsymbol{X}_j)} - \boldsymbol{P}_{(\boldsymbol{X}_0,\ldots,\boldsymbol{X}_{j-1})}, \quad j = 1, \ldots, k, \\ \boldsymbol{A}_{k+1} &= \boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}}. \end{aligned}$$

  Let

$$\begin{aligned} r_j &= \text{rank}(\boldsymbol{A}_j) = \text{rank}(\boldsymbol{P}_{(\boldsymbol{X}_0,\ldots,\boldsymbol{X}_j)}) - \text{rank}(\boldsymbol{P}_{(\boldsymbol{X}_0,\ldots,\boldsymbol{X}_{j-1})}) \\ &= \text{rank}((\boldsymbol{X}_0, \ldots, \boldsymbol{X}_j)) - \text{rank}((\boldsymbol{X}_0, \ldots, \boldsymbol{X}_{j-1})) \end{aligned}$$

  be the extra rank contribution from predictors in $\boldsymbol{X}_j$.

- $\boldsymbol{A}_j$ are orthogonal projections and

$$\sum_{j=0}^{k+1} \boldsymbol{A}_j = \boldsymbol{I}_n.$$

  Therefore we can apply the Cochran's theorem,

$$\frac{\text{SS}_j}{\sigma^2} = \frac{1}{\sigma^2}\boldsymbol{y}^T\boldsymbol{A}_j\boldsymbol{y} \sim \chi^2_{r_j}(\phi_j), \ j = 0, \ldots, k+1,$$

  with noncentrality parameter

$$\phi_j = \frac{1}{2\sigma^2}(\boldsymbol{X}\boldsymbol{b})^T\boldsymbol{A}_j(\boldsymbol{X}\boldsymbol{b})$$

  and are independent.

- ANOVA table for sequential SS (type I SS in SAS). See Table 1.

- Example SAS output. Two-way ANOVA with interaction.

Source	DF	Projection	SS	Noncentrality
$\boldsymbol{b}_0$	$r(\boldsymbol{X}_0)$	$\boldsymbol{A}_0\boldsymbol{P}_{\boldsymbol{X}_0}$	$\boldsymbol{y}^T\boldsymbol{A}_0\boldsymbol{y}$	$(2\sigma^2)^{-1}(\boldsymbol{X}\boldsymbol{b})^T\boldsymbol{A}_0(\boldsymbol{X}\boldsymbol{b})$
$\boldsymbol{b}_1$ after $\boldsymbol{b}_0$	$r(\boldsymbol{X}_0,\boldsymbol{X}_1)-r(\boldsymbol{X}_0)$	$\boldsymbol{A}_1=\boldsymbol{P}_{(\boldsymbol{X}_0,\boldsymbol{X}_1)}-\boldsymbol{P}_{\boldsymbol{X}_0}$	$\boldsymbol{y}^T\boldsymbol{A}_1\boldsymbol{y}$	$(2\sigma^2)^{-1}(\boldsymbol{X}\boldsymbol{b})^T\boldsymbol{A}_1(\boldsymbol{X}\boldsymbol{b})$
...				
$\boldsymbol{b}_j$ after $\boldsymbol{b}_0,\ldots,\boldsymbol{b}_{j-1}$	$r(\boldsymbol{X}_0,\ldots,\boldsymbol{X}_j)-r(\boldsymbol{X}_0,\ldots,\boldsymbol{X}_{j-1})$	$\boldsymbol{A}_j=\boldsymbol{P}_{(\boldsymbol{X}_0,\ldots,\boldsymbol{X}_j)}-\boldsymbol{P}_{(\boldsymbol{X}_0,\ldots,\boldsymbol{X}_{j-1})}$	$\boldsymbol{y}^T\boldsymbol{A}_j\boldsymbol{y}$	$(2\sigma^2)^{-1}(\boldsymbol{X}\boldsymbol{b})^T\boldsymbol{A}_j(\boldsymbol{X}\boldsymbol{b})$
...				
$\boldsymbol{b}_k$ after $\boldsymbol{b}_0,\ldots,\boldsymbol{b}_{k-1}$	$r(\boldsymbol{X}_0,\ldots,\boldsymbol{X}_k)-r(\boldsymbol{X}_0,\ldots,\boldsymbol{X}_{k-1})$	$\boldsymbol{A}_k=\boldsymbol{P}_{(\boldsymbol{X}_0,\ldots,\boldsymbol{X}_k)}-\boldsymbol{P}_{(\boldsymbol{X}_0,\ldots,\boldsymbol{X}_{k-1})}$	$\boldsymbol{y}^T\boldsymbol{A}_k\boldsymbol{y}$	$(2\sigma^2)^{-1}(\boldsymbol{X}\boldsymbol{b})^T\boldsymbol{A}_k(\boldsymbol{X}\boldsymbol{b})$
Error	$n-r(\boldsymbol{X})$	$\boldsymbol{A}_{k+1}=\boldsymbol{I}_n-\boldsymbol{P}_{\boldsymbol{X}}$	$\boldsymbol{y}^T(\boldsymbol{I}_n-\boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{y}$	0
Total	n	$\boldsymbol{I}$	$\boldsymbol{y}^T\boldsymbol{y}$	$(2\sigma^2)^{-1}(\boldsymbol{X}\boldsymbol{b})^T(\boldsymbol{X}\boldsymbol{b})$

Table 1: ANOVA table for sequential SS (type I SS in SAS).

```
proc glm data=bread;
   class height width;
   model sales=height width height*width;
```

The GLM Procedure
Dependent Variable: sales

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	1580.000000	316.000000	30.58	0.0003
Error	6	62.000000	10.333333		
Corrected Total	11	1642.000000			

Source	DF	Type I SS	Mean Square	F Value	Pr > F
height	2	1544.000000	772.000000	74.71	<.0001
width	1	12.000000	12.000000	1.16	0.3226
height*width	2	24.000000	12.000000	1.16	0.3747

Source	DF	Type III SS	Mean Square	F Value	Pr > F
height	2	1544.000000	772.000000	74.71	<.0001
width	1	12.000000	12.000000	1.16	0.3226
height*width	2	24.000000	12.000000	1.16	0.3747

17 Lecture 17: Nov 4

Announcement

- HW6 due this Wed
- HW7 due this Wed.
- Midterm 2 change of time?

Last time

- t test and confidence interval

$$\boldsymbol{r}^T\hat{\boldsymbol{b}} \pm t_{n-r,\alpha/2}\hat{\sigma}\sqrt{\boldsymbol{r}^T(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{r}}.$$

- Simultaneous confidence intervals and multiple comparison
  - Bonferroni for s linear hypotheses $\boldsymbol{r}_i^T\boldsymbol{b}$

$$\boldsymbol{r}_i^T\hat{\boldsymbol{b}} \pm t_{n-r,\alpha/(2s)}\hat{\sigma}\sqrt{\boldsymbol{r}_i^T(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{r}_i}.$$

  - Scheffé for all linear combinations $\boldsymbol{u}^T\boldsymbol{R}\boldsymbol{b}$

$$\boldsymbol{u}^T\boldsymbol{R}\hat{\boldsymbol{b}} \pm c\hat{\sigma}\sqrt{\boldsymbol{u}^T\boldsymbol{R}(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{R}^T\boldsymbol{u}}.$$

    where $c = \sqrt{sF_{s,n-r,\alpha}}$.
  - Tukey for all pairwise differences between treatment effects $\alpha_i - \alpha_j$ in balanced one-way ANOVA

$$(\bar{y}_{i\cdot} - \bar{y}_{j\cdot}) \pm \frac{\hat{\sigma}}{\sqrt{n}} w_{a,a(n-1),\alpha},$$

    where $w_{a,a(n-1),\alpha}$ is critical value of $W = (\max_j Z_j - \min_j Z_j)/\sqrt{U/(an-a)}$, $U \sim \chi^2_{a(n-1)}$.

- Sequential SS (type I SS).

Today

- Testing under the Aitken model $\boldsymbol{y} \sim N(\boldsymbol{X}\boldsymbol{b}, \sigma^2\boldsymbol{V})$.

Testing under the Aitken model

Assume the Aitken model $\boldsymbol{y} \sim N(\boldsymbol{X}\boldsymbol{b}, \sigma^2\boldsymbol{V})$, where $\boldsymbol{V}$ can be singular.

- Consider testing a reduced model versus a full model

$$H_0 : \boldsymbol{y} \sim (\boldsymbol{X}_0\boldsymbol{c}, \sigma^2\boldsymbol{V}) \quad \text{vs} \quad H_A : \boldsymbol{y} \sim (\boldsymbol{X}\boldsymbol{b}, \sigma^2\boldsymbol{V}),$$

  where $\mathcal{C}(\boldsymbol{X}_0) \subset \mathcal{C}(\boldsymbol{X})$. Equivalently, we are testing the column space hypothesis

$$H_0 : \boldsymbol{b} \in \mathcal{C}(\boldsymbol{B}) \quad \text{vs} \quad H_A : \boldsymbol{b} \in \mathbb{R}^p$$

  for some $\boldsymbol{B}$. It's a special case of the more general linear hypothesis $\boldsymbol{R}\boldsymbol{b} = \boldsymbol{r}$.

- The MVAUE of the estimable function $\boldsymbol{X}\boldsymbol{b}$ is

$$\widehat{\boldsymbol{\Lambda}\boldsymbol{b}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{V}_1^-\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{V}_1^-\boldsymbol{y},$$

  where $\boldsymbol{V}_1 = \boldsymbol{V} + \boldsymbol{X}\boldsymbol{X}^T$. We also showed that

$$\widehat{\boldsymbol{X}\boldsymbol{b}} = \boldsymbol{X}\hat{\boldsymbol{b}},$$

  where $\hat{\boldsymbol{b}} = (\boldsymbol{X}^T\boldsymbol{V}_1^-\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{V}_1^-\boldsymbol{y}$ minimizes the generalized least squares criterion $(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b})^T\boldsymbol{V}_1^-(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b})$. Therefore the SSE under the full model is

$$\begin{aligned}
\text{SSE} &= (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{b}})^T\boldsymbol{V}_1^-(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{b}}) \\
&= \boldsymbol{y}^T[\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{V}_1^-\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{V}_1^-]^T\boldsymbol{V}_1^-[\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{V}_1^-\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{V}_1^-]\boldsymbol{y} \\
&= \boldsymbol{y}^T[\boldsymbol{V}_1^- - \boldsymbol{V}_1^-\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{V}_1^-\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{V}_1^-]\boldsymbol{y} \\
&=: \boldsymbol{y}^T\boldsymbol{A}_1\boldsymbol{y}.
\end{aligned}$$

  The third equality uses HW4 Q5 (c) $\mathcal{C}(\boldsymbol{X}^T\boldsymbol{V}_1^-\boldsymbol{X}) = \mathcal{C}(\boldsymbol{X}^T)$.

  We observe that SSE is invariant to the choice of the generalized inverses $\boldsymbol{V}_1^-$ and $(\boldsymbol{X}^T\boldsymbol{V}_1^-\boldsymbol{X})^-$ (why?). Without loss of generality, we may assume $\boldsymbol{A}_1$ is symmetric by using the Moore-Penrose inverses throughout.

- Similarly, the SSE under the reduced model is

$$\begin{aligned}
\text{SSE}_0 &= \boldsymbol{y}^T[\boldsymbol{I} - \boldsymbol{X}_0(\boldsymbol{X}_0^T\boldsymbol{V}_0^-\boldsymbol{X}_0)^-\boldsymbol{X}_0^T\boldsymbol{V}_0^-]^T\boldsymbol{V}_0^-[\boldsymbol{I} - \boldsymbol{X}_0(\boldsymbol{X}_0^T\boldsymbol{V}_0^-\boldsymbol{X}_0)^-\boldsymbol{X}_0^T\boldsymbol{V}_0^-]\boldsymbol{y} \\
&= \boldsymbol{y}^T[\boldsymbol{V}_0^- - \boldsymbol{V}_0^-\boldsymbol{X}_0(\boldsymbol{X}_0^T\boldsymbol{V}_0^-\boldsymbol{X}_0)^-\boldsymbol{X}_0^T\boldsymbol{V}_0^-]\boldsymbol{y} \\
&=: \boldsymbol{y}^T\boldsymbol{A}_0\boldsymbol{y},
\end{aligned}$$

  where $\boldsymbol{V}_0 = \boldsymbol{V} + \boldsymbol{X}_0\boldsymbol{X}_0^T$.

- A reasonable test for a general linear hypothesis may be based on the change in SSEs under the constrained and full models

$$\frac{\text{SSE}_0 - \text{SSE}}{\text{SSE}}.$$

We need some preparatory results first.

- Assume $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \boldsymbol{V})$, where $\boldsymbol{\mu} \in \mathcal{C}(\boldsymbol{V})$ and $\boldsymbol{V}$ is an orthogonal projection matrix. Then

$$\boldsymbol{X}^T\boldsymbol{X} \sim \chi^2_{\text{rank}(\boldsymbol{V})}(\boldsymbol{\mu}^T\boldsymbol{\mu}/2).$$

Proof. $\boldsymbol{V} = \boldsymbol{O}\boldsymbol{O}^T$, where $\boldsymbol{O} \in \mathbb{R}^{n\times\text{rank}(\boldsymbol{V})}$ and $\boldsymbol{O}^T\boldsymbol{O} = \boldsymbol{I}_{\text{rank}(\boldsymbol{V})}$. Note $\mathcal{C}(\boldsymbol{V}) = \mathcal{C}(\boldsymbol{O})$. Let $\boldsymbol{\mu} = \boldsymbol{O}\boldsymbol{b}$ for some $\boldsymbol{b}$. Then $\boldsymbol{X} = \boldsymbol{O}\boldsymbol{Z}$, where $\boldsymbol{Z} \sim N(\boldsymbol{b}, \boldsymbol{I}_{\text{rank}(\boldsymbol{V})})$. Thus

$$\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{Z}^T\boldsymbol{O}^T\boldsymbol{O}\boldsymbol{Z} = \boldsymbol{Z}^T\boldsymbol{Z} \sim \chi^2_{\text{rank}(\boldsymbol{V})}(\boldsymbol{\mu}^T\boldsymbol{\mu}/2).$$

□

- Assume $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \boldsymbol{V})$, $\boldsymbol{V}$ possibly singular and $\boldsymbol{A}$ is symmetric. Then

$$\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X} \sim \chi^2_{\text{tr}(\boldsymbol{A}\boldsymbol{V})}(\boldsymbol{\mu}^T\boldsymbol{A}\boldsymbol{\mu}/2)$$

if (1) $\boldsymbol{V}\boldsymbol{A}\boldsymbol{V}\boldsymbol{A}\boldsymbol{V} = \boldsymbol{V}\boldsymbol{A}\boldsymbol{V}$, (2) $\boldsymbol{\mu}^T\boldsymbol{A}\boldsymbol{V}\boldsymbol{A}\boldsymbol{\mu} = \boldsymbol{\mu}^T\boldsymbol{A}\boldsymbol{\mu}$, and (3) $\boldsymbol{V}\boldsymbol{A}\boldsymbol{V}\boldsymbol{A}\boldsymbol{\mu} = \boldsymbol{V}\boldsymbol{A}\boldsymbol{\mu}$.

Remark: Previously (in Lecture 13) we showed the special case where $\boldsymbol{V}$ is nonsingular, $\boldsymbol{A}$ is symmetric, and $\boldsymbol{A}\boldsymbol{V}$ is idempotent.

Proof. $\boldsymbol{X} = \boldsymbol{\mu} + \boldsymbol{e}$, where $\boldsymbol{e} \sim N(\boldsymbol{0}, \boldsymbol{V})$ and thus $\boldsymbol{e} \in \mathcal{C}(\boldsymbol{V})$. Then

$$\begin{aligned}
& \boldsymbol{X}^T\boldsymbol{A}\boldsymbol{V}\boldsymbol{A}\boldsymbol{X} \\
= \; & (\boldsymbol{\mu}+\boldsymbol{e})^T(\boldsymbol{A}\boldsymbol{V}\boldsymbol{A})(\boldsymbol{\mu}+\boldsymbol{e}) \\
= \; & \boldsymbol{\mu}^T\boldsymbol{A}\boldsymbol{V}\boldsymbol{A}\boldsymbol{\mu} + 2\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{V}\boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{e}^T\boldsymbol{A}\boldsymbol{V}\boldsymbol{A}\boldsymbol{e} \\
= \; & \boldsymbol{\mu}^T\boldsymbol{A}\boldsymbol{\mu} + 2\boldsymbol{e}^T\boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e} \\
= \; & \boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X}.
\end{aligned}$$

Let $\boldsymbol{V} = \boldsymbol{Q}\boldsymbol{Q}^T$ where $\boldsymbol{Q}$ can be the Cholesky factor or symmetric square root. Note $\mathcal{C}(\boldsymbol{V}) = \mathcal{C}(\boldsymbol{Q})$. Then

$$\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X} = \boldsymbol{X}^T\boldsymbol{A}\boldsymbol{V}\boldsymbol{A}\boldsymbol{X} = (\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{Q})(\boldsymbol{Q}^T\boldsymbol{A}\boldsymbol{X})$$

and

$$\boldsymbol{Q}^T\boldsymbol{A}\boldsymbol{X} \sim N(\boldsymbol{Q}^T\boldsymbol{A}\boldsymbol{\mu}, \boldsymbol{Q}^T\boldsymbol{A}\boldsymbol{V}\boldsymbol{A}\boldsymbol{Q}).$$

By assumption (1), the covariance matrix is an orthogonal projection with rank

$$\begin{aligned} & \operatorname{rank}(\boldsymbol{Q}^T\boldsymbol{A}\boldsymbol{V}\boldsymbol{A}\boldsymbol{Q}) = \operatorname{rank}(\boldsymbol{Q}^T\boldsymbol{A}\boldsymbol{Q}) \\ = \quad & \operatorname{tr}(\boldsymbol{Q}^T\boldsymbol{A}\boldsymbol{Q}) = \operatorname{tr}(\boldsymbol{A}\boldsymbol{Q}\boldsymbol{Q}^T) = \operatorname{tr}(\boldsymbol{A}\boldsymbol{V}). \end{aligned}$$

We also need to check $\boldsymbol{Q}^T\boldsymbol{A}\boldsymbol{\mu} \in \mathcal{C}(\boldsymbol{Q}^T\boldsymbol{A}\boldsymbol{V}\boldsymbol{A}\boldsymbol{Q})$. This is true because

$$\begin{aligned} & \boldsymbol{Q}^T\boldsymbol{A}\boldsymbol{V}\boldsymbol{A}\boldsymbol{Q}(\boldsymbol{Q}^T\boldsymbol{A}\boldsymbol{\mu}) \\ = \quad & \boldsymbol{Q}^T\boldsymbol{A}\boldsymbol{V}\boldsymbol{A}\boldsymbol{V}\boldsymbol{A}\boldsymbol{\mu} \\ = \quad & \boldsymbol{Q}^T\boldsymbol{A}\boldsymbol{V}\boldsymbol{A}\boldsymbol{\mu} \\ = \quad & \boldsymbol{Q}^T\boldsymbol{A}\boldsymbol{\mu}. \end{aligned}$$

Thus the result follows form the preceding result. □

- If $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \boldsymbol{V})$, $\boldsymbol{V}$ possibly singular, and (1) $\boldsymbol{V}\boldsymbol{A}\boldsymbol{V}\boldsymbol{B}\boldsymbol{V} = \boldsymbol{0}$, (2) $\boldsymbol{V}\boldsymbol{A}\boldsymbol{V}\boldsymbol{B}\boldsymbol{\mu} = \boldsymbol{0}$, (3) $\boldsymbol{V}\boldsymbol{B}\boldsymbol{V}\boldsymbol{A}\boldsymbol{\mu} = \boldsymbol{0}$, (4) $\boldsymbol{\mu}^T\boldsymbol{A}\boldsymbol{V}\boldsymbol{B}\boldsymbol{\mu} = \boldsymbol{0}$, then $\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X}$ and $\boldsymbol{X}^T\boldsymbol{B}\boldsymbol{X}$ are independent.

  Remark: Previously (in Lecture 13) we showed the special case where $\boldsymbol{V}$ is nonsingular and $\boldsymbol{A}$ and $\boldsymbol{B}$ are symmetric.

  Proof. Let $\boldsymbol{V} = \boldsymbol{Q}\boldsymbol{\Lambda}_s\boldsymbol{Q}^T$, where Λ_s is diagonal containing all nonzero eigenvalues of $\boldsymbol{V}$. Then $\boldsymbol{X} = \boldsymbol{\mu} + \boldsymbol{Q}\boldsymbol{\Lambda}_s^{1/2}\boldsymbol{Z}$, where $\boldsymbol{Z} \sim N(\boldsymbol{0}, \boldsymbol{I})$, and

$$\begin{aligned} \boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X} &= (\boldsymbol{\mu} + \boldsymbol{Q}\boldsymbol{\Lambda}_s^{1/2}\boldsymbol{Z})^T\boldsymbol{A}(\boldsymbol{\mu} + \boldsymbol{Q}\boldsymbol{\Lambda}_s^{1/2}\boldsymbol{Z}) \\ &= \boldsymbol{\mu}^T\boldsymbol{A}\boldsymbol{\mu} + 2\boldsymbol{\mu}^T\boldsymbol{A}\boldsymbol{Q}\boldsymbol{\Lambda}_s^{1/2}\boldsymbol{Z} + \boldsymbol{Z}^T\boldsymbol{\Lambda}_s^{1/2}\boldsymbol{Q}^T\boldsymbol{A}\boldsymbol{Q}\boldsymbol{\Lambda}_s^{1/2}\boldsymbol{Z} \\ \boldsymbol{X}^T\boldsymbol{B}\boldsymbol{X} &= (\boldsymbol{\mu} + \boldsymbol{Q}\boldsymbol{\Lambda}_s^{1/2}\boldsymbol{Z})^T\boldsymbol{B}(\boldsymbol{\mu} + \boldsymbol{Q}\boldsymbol{\Lambda}_s^{1/2}\boldsymbol{Z}) \\ &= \boldsymbol{\mu}^T\boldsymbol{B}\boldsymbol{\mu} + 2\boldsymbol{\mu}^T\boldsymbol{B}\boldsymbol{Q}\boldsymbol{Z}\boldsymbol{\Lambda}_s^{1/2} + \boldsymbol{Z}^T\boldsymbol{\Lambda}_s^{1/2}\boldsymbol{Q}^T\boldsymbol{B}\boldsymbol{Q}\boldsymbol{\Lambda}_s^{1/2}\boldsymbol{Z}. \end{aligned}$$

  Now we check independence between the summands of the two expressions.

– From assumption (4),

$$\begin{aligned} & \boldsymbol{\mu}^T \boldsymbol{A}\boldsymbol{Q}\boldsymbol{\Lambda}_s^{1/2}\boldsymbol{\Lambda}_s^{1/2}\boldsymbol{Q}^T\boldsymbol{B}\boldsymbol{\mu} = \boldsymbol{\mu}^T\boldsymbol{A}\boldsymbol{V}\boldsymbol{B}\boldsymbol{\mu} = 0 \\ \Rightarrow \quad & \boldsymbol{\mu}^T\boldsymbol{A}\boldsymbol{Q}\boldsymbol{Z} \perp \boldsymbol{\mu}^T\boldsymbol{B}\boldsymbol{Q}\boldsymbol{Z}. \end{aligned}$$

– From assumption (3),

$$\begin{aligned} & \boldsymbol{V}\boldsymbol{B}\boldsymbol{V}\boldsymbol{A}\boldsymbol{\mu} = \boldsymbol{0} \\ \Rightarrow \quad & \boldsymbol{Q}\boldsymbol{\Lambda}_s\boldsymbol{Q}^T\boldsymbol{B}\boldsymbol{V}\boldsymbol{A}\boldsymbol{\mu} = \boldsymbol{0} \\ \Rightarrow \quad & \boldsymbol{Q}^T\boldsymbol{B}\boldsymbol{V}\boldsymbol{A}\boldsymbol{\mu} = \boldsymbol{0} \\ \Rightarrow \quad & \boldsymbol{\mu}^T\boldsymbol{A}\boldsymbol{Q}\boldsymbol{\Lambda}_s^{1/2}\boldsymbol{\Lambda}_s^{1/2}\boldsymbol{Q}^T\boldsymbol{B}\boldsymbol{Q} = \boldsymbol{0}^T \\ \Rightarrow \quad & \boldsymbol{\mu}^T\boldsymbol{A}\boldsymbol{Q}\boldsymbol{\Lambda}_s^{1/2}\boldsymbol{Z} \perp \boldsymbol{Z}^T\boldsymbol{\Lambda}_s^{1/2}\boldsymbol{Q}^T\boldsymbol{B}\boldsymbol{Q}\boldsymbol{\Lambda}_s^{1/2}\boldsymbol{Z}. \end{aligned}$$

– Similarly, by assumption (2), $\boldsymbol{\mu}^T\boldsymbol{B}\boldsymbol{Q}\boldsymbol{\Lambda}_s^{1/2}\boldsymbol{Z} \perp \boldsymbol{Z}^T\boldsymbol{\Lambda}_s^{1/2}\boldsymbol{Q}^T\boldsymbol{A}\boldsymbol{Q}\boldsymbol{\Lambda}_s^{1/2}\boldsymbol{Z}$.

– From assumption (1),

$$\begin{aligned} & \boldsymbol{V}\boldsymbol{A}\boldsymbol{V}\boldsymbol{B} = \boldsymbol{0} \\ = \quad & \boldsymbol{Q}\boldsymbol{\Lambda}_s\boldsymbol{Q}^T\boldsymbol{A}\boldsymbol{V}\boldsymbol{B} = \boldsymbol{0} \\ \Rightarrow \quad & \boldsymbol{Q}^T\boldsymbol{A}\boldsymbol{Q}\boldsymbol{\Lambda}_s^{1/2}\boldsymbol{\Lambda}_s^{1/2}\boldsymbol{Q}^T\boldsymbol{B}\boldsymbol{Q} = \boldsymbol{0} \\ \Rightarrow \quad & \boldsymbol{Z}^T\boldsymbol{\Lambda}_s^{1/2}\boldsymbol{Q}^T\boldsymbol{A}\boldsymbol{Q}\boldsymbol{\Lambda}_s^{1/2}\boldsymbol{Z} \perp \boldsymbol{Z}^T\boldsymbol{\Lambda}_s^{1/2}\boldsymbol{Q}^T\boldsymbol{B}\boldsymbol{Q}\boldsymbol{\Lambda}_s^{1/2}\boldsymbol{Z}. \end{aligned}$$

Therefore, we have $\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X} \perp \boldsymbol{X}^T\boldsymbol{B}\boldsymbol{X}$. □

18 Lecture 18: Nov 6

Announcement

- HW6 due this Wed.
- HW7 due next Mon.
- Midterm 2 changed to next Wed Nov 13 @ 11:45AM-1PM.
- Q&A for basic exam questions? (if interested, submit your questions by 11/27).

Last time

Testing under Aitken model $\boldsymbol{y} \sim N(\boldsymbol{X}\boldsymbol{b}, \sigma^2\boldsymbol{V})$ – preparatory results. Note correction to the proof about independence of quadratic forms result (thanks to Xue Feng).

Today

Testing under Aitken model $\boldsymbol{y} \sim N(\boldsymbol{X}\boldsymbol{b}, \sigma^2\boldsymbol{V})$.

Testing under the Aitken model (cont'd)

Test $H_0 : \boldsymbol{y} \sim (\boldsymbol{X}_0\boldsymbol{c}, \sigma^2\boldsymbol{V})$ vs $H_A : \boldsymbol{y} \sim (\boldsymbol{X}\boldsymbol{b}, \sigma^2\boldsymbol{V})$, $\boldsymbol{V}$ possibly singular.

- Now we can state and prove the results related to testing under Aitken model. Recall the notation

$$\begin{aligned} \text{SSE} &= \boldsymbol{y}^T\boldsymbol{A}_1\boldsymbol{y} \\ \text{SSE}_0 &= \boldsymbol{y}^T\boldsymbol{A}_0\boldsymbol{y}, \end{aligned}$$

  where

$$\begin{aligned} \boldsymbol{A}_1 &= \boldsymbol{V}_1^- - \boldsymbol{V}_1^-\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{V}_1^-\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{V}_1^-, \quad \boldsymbol{V}_1 = \boldsymbol{V} + \boldsymbol{X}\boldsymbol{X}^T \\ \boldsymbol{A}_0 &= \boldsymbol{V}_0^- - \boldsymbol{V}_0^-\boldsymbol{X}_0(\boldsymbol{X}_0^T\boldsymbol{V}_0^-\boldsymbol{X}_0)^-\boldsymbol{X}_0^T\boldsymbol{V}_0^-, \quad \boldsymbol{V}_0 = \boldsymbol{V} + \boldsymbol{X}_0\boldsymbol{X}_0^T. \end{aligned}$$

- **Theorem**

  1. $\frac{1}{\sigma^2}\boldsymbol{y}^T\boldsymbol{A}_1\boldsymbol{y} \sim \chi^2_{\text{tr}(\boldsymbol{A}_1\boldsymbol{V})}$.

2. If $\boldsymbol{Xb} \in \mathcal{C}(\boldsymbol{X}_0, \boldsymbol{V})$, then
$$\frac{1}{\sigma^2}\boldsymbol{y}^T(\boldsymbol{A}_0 - \boldsymbol{A}_1)\boldsymbol{y} \sim \chi^2_{\mathrm{tr}((\boldsymbol{A}_0-\boldsymbol{A}_1)\boldsymbol{V})}(\boldsymbol{b}^T\boldsymbol{X}^T\boldsymbol{A}_0\boldsymbol{Xb}/2).$$
3. If $\boldsymbol{Xb} \in \mathcal{C}(\boldsymbol{X}_0, \boldsymbol{V})$, then $\boldsymbol{y}^T\boldsymbol{A}_1\boldsymbol{y}$ is independent of $\boldsymbol{y}^T(\boldsymbol{A}_0 - \boldsymbol{A}_1)\boldsymbol{y}$.

- The proof involves checking various conditions in the preparatory results.

  – $\boldsymbol{VA}_1 = \boldsymbol{V}_1\boldsymbol{A}_1$
  $\boldsymbol{VA}_0 = \boldsymbol{V}_0\boldsymbol{A}_0$.

  Proof.
  $$\begin{aligned}
  \boldsymbol{VA}_1 &= (\boldsymbol{V}_1 - \boldsymbol{XX}^T)\boldsymbol{A}_1 \\
  &= \boldsymbol{V}_1\boldsymbol{A}_1 - \boldsymbol{XX}^T[\boldsymbol{V}_1^- - \boldsymbol{V}_1^-\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{V}_1^-\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{V}_1^-] \\
  &= \boldsymbol{V}_1\boldsymbol{A}_1 - \boldsymbol{XX}^T\boldsymbol{V}_1^- + \boldsymbol{XX}^T\boldsymbol{V}_1^-\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{V}_1^-\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{V}_1^- \\
  &= \boldsymbol{V}_1\boldsymbol{A}_1 - \boldsymbol{XX}^T\boldsymbol{V}_1^- + \boldsymbol{XX}^T\boldsymbol{V}_1^- \\
  &= \boldsymbol{V}_1\boldsymbol{A}_1.
  \end{aligned}$$
  Similarly $\boldsymbol{VA}_0 = \boldsymbol{V}_0\boldsymbol{A}_0$. □

  – (1) $\boldsymbol{VA}_1\boldsymbol{VA}_1\boldsymbol{V} = \boldsymbol{VA}_1\boldsymbol{V}$
  (2) $\boldsymbol{b}^T\boldsymbol{X}^T\boldsymbol{A}_1\boldsymbol{VA}_1\boldsymbol{Xb} = \boldsymbol{b}^T\boldsymbol{X}^T\boldsymbol{A}_1\boldsymbol{Xb}$
  (3) $\boldsymbol{VA}_1\boldsymbol{VA}_1\boldsymbol{Xb} = \boldsymbol{VA}_1\boldsymbol{Xb}$.
  Remark: The first part of the theorem is checked by this result. Note that the centrality parameter is $\boldsymbol{b}^T\boldsymbol{X}^T\boldsymbol{A}_1\boldsymbol{Xb} = 0$ since
  $$\boldsymbol{A}_1\boldsymbol{X} = [\boldsymbol{V}_1^- - \boldsymbol{V}_1^-\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{V}_1^-\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{V}_1^-]\boldsymbol{X} = \boldsymbol{0}.$$
  Proof. For (1), by previous fact,
  $$\begin{aligned}
  &\boldsymbol{VA}_1\boldsymbol{VA}_1\boldsymbol{V} = \boldsymbol{VA}_1\boldsymbol{V}_1\boldsymbol{A}_1\boldsymbol{V} \\
  = \ &\boldsymbol{V}[\boldsymbol{I} - \boldsymbol{V}_1^-\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{V}_1^-\boldsymbol{X})^-\boldsymbol{X}^T]\boldsymbol{V}_1^-\boldsymbol{V}_1\boldsymbol{V}_1^-[\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{V}_1^-\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{V}_1^-]\boldsymbol{V} \\
  = \ &\boldsymbol{V}[\boldsymbol{I} - \boldsymbol{V}_1^-\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{V}_1^-\boldsymbol{X})^-\boldsymbol{X}^T]\boldsymbol{V}_1^-[\boldsymbol{V}_1\boldsymbol{V}_1^- - \boldsymbol{V}_1\boldsymbol{V}_1^-\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{V}_1^-\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{V}_1^-]\boldsymbol{V} \\
  = \ &\boldsymbol{V}[\boldsymbol{I} - \boldsymbol{V}_1^-\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{V}_1^-\boldsymbol{X})^-\boldsymbol{X}^T]\boldsymbol{V}_1^-[\boldsymbol{V}_1\boldsymbol{V}_1^- - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{V}_1^-\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{V}_1^-]\boldsymbol{V} \\
  = \ &\boldsymbol{V}[\boldsymbol{I} - \boldsymbol{V}_1^-\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{V}_1^-\boldsymbol{X})^-\boldsymbol{X}^T]\boldsymbol{V}_1^-[\boldsymbol{V}_1\boldsymbol{V}_1^-\boldsymbol{V} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{V}_1^-\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{V}_1^-\boldsymbol{V}] \\
  = \ &\boldsymbol{V}[\boldsymbol{I} - \boldsymbol{V}_1^-\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{V}_1^-\boldsymbol{X})^-\boldsymbol{X}^T]\boldsymbol{V}_1^-[\boldsymbol{V} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{V}_1^-\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{V}_1^-\boldsymbol{V}] \\
  = \ &\boldsymbol{V}[\boldsymbol{I} - \boldsymbol{V}_1^-\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{V}_1^-\boldsymbol{X})^-\boldsymbol{X}^T]\boldsymbol{V}_1^-[\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{V}_1^-\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{V}_1^-]\boldsymbol{V} \\
  = \ &\boldsymbol{VA}_1\boldsymbol{V}.
  \end{aligned}$$

For (2),

$$\begin{aligned} & \boldsymbol{b}^T\boldsymbol{X}^T\boldsymbol{A}_1\boldsymbol{V}\boldsymbol{A}_1\boldsymbol{X}\boldsymbol{b} \\ = \; & \boldsymbol{b}^T\boldsymbol{X}^T\boldsymbol{A}_1\boldsymbol{V}_1\boldsymbol{A}_1\boldsymbol{X}\boldsymbol{b} \\ = \; & \boldsymbol{b}^T\boldsymbol{X}^T\boldsymbol{A}_1\boldsymbol{X}\boldsymbol{b} \quad (\text{since } \mathcal{C}(\boldsymbol{X}) \subset \mathcal{C}(\boldsymbol{V}_1)). \end{aligned}$$

For (3),

$$\begin{aligned} & \boldsymbol{V}\boldsymbol{A}_1\boldsymbol{V}\boldsymbol{A}_1\boldsymbol{X}\boldsymbol{b} \\ = \; & \boldsymbol{V}_1\boldsymbol{A}_1\boldsymbol{V}_1\boldsymbol{A}_1\boldsymbol{X}\boldsymbol{b} \\ = \; & \boldsymbol{V}_1\boldsymbol{A}_1\boldsymbol{X}\boldsymbol{b} \quad (\text{since } \mathcal{C}(\boldsymbol{X}) \subset \mathcal{C}(\boldsymbol{V}_1)) \\ = \; & \boldsymbol{V}\boldsymbol{A}_1\boldsymbol{X}\boldsymbol{b}. \end{aligned}$$

□

– $\boldsymbol{V}\boldsymbol{A}_1\boldsymbol{V}\boldsymbol{A}_0\boldsymbol{V} = \boldsymbol{V}\boldsymbol{A}_1\boldsymbol{V}$.

Proof.

$$\begin{aligned} & \boldsymbol{V}\boldsymbol{A}_1\boldsymbol{V}\boldsymbol{A}_0\boldsymbol{V} \\ = \; & \boldsymbol{V}\boldsymbol{A}_1\boldsymbol{V}_0[\boldsymbol{V}_0^- - \boldsymbol{V}_0^-\boldsymbol{X}_0(\boldsymbol{X}_0^T\boldsymbol{V}_0^-\boldsymbol{X}_0)^-\boldsymbol{X}_0^T\boldsymbol{V}_0^-]\boldsymbol{V} \\ = \; & \boldsymbol{V}\boldsymbol{A}_1[\boldsymbol{V} - \boldsymbol{X}_0(\boldsymbol{X}_0^T\boldsymbol{V}_0^-\boldsymbol{X}_0)^-\boldsymbol{X}_0^T\boldsymbol{V}_0^-\boldsymbol{V}] \\ = \; & \boldsymbol{V}\boldsymbol{A}_1\boldsymbol{V} \\ = \; & \boldsymbol{V}\boldsymbol{A}\boldsymbol{V}. \end{aligned}$$

The third equality is because

$$\begin{aligned} & \boldsymbol{A}_1\boldsymbol{X}_0 \\ = \; & [\boldsymbol{V}_1^- - \boldsymbol{V}_1^-\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{V}_1^-\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{V}_1^-]\boldsymbol{X}_0 \\ = \; & \boldsymbol{V}_1^-\boldsymbol{X}_0 - \boldsymbol{V}_1^-\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{V}_1^-\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{V}_1^-\boldsymbol{X}_0 \\ = \; & \boldsymbol{V}_1^-\boldsymbol{X}_0 - \boldsymbol{V}_1^-\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{V}_1^-\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{V}_1^-\boldsymbol{X}\boldsymbol{T} \quad (\mathcal{C}(\boldsymbol{X}_0) \subset \mathcal{C}(\boldsymbol{X})) \\ = \; & \boldsymbol{V}_1^-\boldsymbol{X}_0 - \boldsymbol{V}_1^-\boldsymbol{X}_0 \\ = \; & \boldsymbol{0}. \end{aligned}$$

□

– (1) $\boldsymbol{V}(\boldsymbol{A}_0 - \boldsymbol{A}_1)\boldsymbol{V}(\boldsymbol{A}_0 - \boldsymbol{A}_1)\boldsymbol{V} = \boldsymbol{V}(\boldsymbol{A}_0 - \boldsymbol{A}_1)\boldsymbol{V}$.
(2) $\boldsymbol{b}^T\boldsymbol{X}^T(\boldsymbol{A}_0-\boldsymbol{A}_1)\boldsymbol{V}(\boldsymbol{A}_0-\boldsymbol{A}_1)\boldsymbol{X}\boldsymbol{b} = \boldsymbol{b}^T\boldsymbol{X}^T(\boldsymbol{A}_0-\boldsymbol{A}_1)\boldsymbol{X}\boldsymbol{b} = \boldsymbol{b}^T\boldsymbol{X}^T\boldsymbol{A}_0\boldsymbol{X}\boldsymbol{b}$.
(3) If $\boldsymbol{X}\boldsymbol{b} \in \mathcal{C}(\boldsymbol{X}_0, \boldsymbol{V})$, then $\boldsymbol{V}(\boldsymbol{A}_0-\boldsymbol{A}_1)\boldsymbol{V}(\boldsymbol{A}_0-\boldsymbol{A}_1)\boldsymbol{X}\boldsymbol{b} = \boldsymbol{V}(\boldsymbol{A}_0-\boldsymbol{A}_1)\boldsymbol{X}\boldsymbol{b}$.

Remark: The second part of the theorem is checked by this result. The non-centrality parameter is

$$\frac{1}{2}\boldsymbol{b}^T\boldsymbol{X}^T(\boldsymbol{A}_0 - \boldsymbol{A}_1)\boldsymbol{X}\boldsymbol{b} = \frac{1}{2}\boldsymbol{b}^T\boldsymbol{X}^T\boldsymbol{A}_0\boldsymbol{X}\boldsymbol{b}.$$

and the degrees of freedom is $\mathrm{tr}((\boldsymbol{A}_0 - \boldsymbol{A}_1)\boldsymbol{V})$.

Proof. For (1),

$$\begin{aligned} & \boldsymbol{V}(\boldsymbol{A}_0 - \boldsymbol{A}_1)\boldsymbol{V}(\boldsymbol{A}_0 - \boldsymbol{A}_1)\boldsymbol{V} \\ = \; & \boldsymbol{V}\boldsymbol{A}_0\boldsymbol{V}\boldsymbol{A}_0\boldsymbol{V} - 2\boldsymbol{V}\boldsymbol{A}_0\boldsymbol{V}\boldsymbol{A}_1\boldsymbol{V} + \boldsymbol{V}\boldsymbol{A}_1\boldsymbol{V}\boldsymbol{A}_1\boldsymbol{V} \\ = \; & \boldsymbol{V}\boldsymbol{A}_0\boldsymbol{V} - 2\boldsymbol{V}\boldsymbol{A}_1\boldsymbol{V} + \boldsymbol{V}\boldsymbol{A}_1\boldsymbol{V} \\ = \; & \boldsymbol{V}(\boldsymbol{A}_0 - \boldsymbol{A}_1)\boldsymbol{V}. \end{aligned}$$

For (2), recall that $\boldsymbol{A}_1\boldsymbol{X} = \boldsymbol{0}$. Then (2) follows from (same proof as the preceding result part (1))

$$\boldsymbol{V}_0\boldsymbol{A}_0\boldsymbol{V}_0\boldsymbol{A}_0\boldsymbol{V}_0 = \boldsymbol{V}_0\boldsymbol{A}_0\boldsymbol{V}_0.$$

For (3), since $\boldsymbol{A}_1\boldsymbol{X} = \boldsymbol{0}$, enough to show

$$\boldsymbol{V}\boldsymbol{A}_0\boldsymbol{V}\boldsymbol{A}_0\boldsymbol{X}\boldsymbol{b} - \boldsymbol{V}\boldsymbol{A}_1\boldsymbol{V}\boldsymbol{A}_0\boldsymbol{X}\boldsymbol{b} = \boldsymbol{V}\boldsymbol{A}_0\boldsymbol{X}\boldsymbol{b}.$$

With $\boldsymbol{V}\boldsymbol{A}_0\boldsymbol{V}\boldsymbol{A}_0\boldsymbol{X}\boldsymbol{b} = \boldsymbol{V}\boldsymbol{A}_0\boldsymbol{X}\boldsymbol{b}$, it's enough to show

$$\boldsymbol{V}\boldsymbol{A}_1\boldsymbol{V}\boldsymbol{A}_0\boldsymbol{X}\boldsymbol{b} = \boldsymbol{0}.$$

This is verified by

$$\begin{aligned} & \boldsymbol{V}\boldsymbol{A}_1\boldsymbol{V}\boldsymbol{A}_0\boldsymbol{X}\boldsymbol{b} \\ = \; & \boldsymbol{V}\boldsymbol{A}_1\boldsymbol{V}\boldsymbol{A}_0\boldsymbol{V}\boldsymbol{c} \quad (\text{because } \boldsymbol{X}\boldsymbol{b} \in \mathcal{C}(\boldsymbol{X}_0, \boldsymbol{V}) = \mathcal{C}(\boldsymbol{V}_0) \text{ and } \boldsymbol{A}_0\boldsymbol{X}_0 = \boldsymbol{0}) \\ = \; & \boldsymbol{V}\boldsymbol{A}_1\boldsymbol{V}\boldsymbol{c} \quad (\text{preceding result}) \\ = \; & \boldsymbol{V}\boldsymbol{A}_1\boldsymbol{X}\boldsymbol{b} \quad (\text{because } \boldsymbol{A}_1\boldsymbol{X}_0 = \boldsymbol{0}) \\ = \; & \boldsymbol{0}. \quad (\text{because } \boldsymbol{A}_1\boldsymbol{X} = \boldsymbol{0}) \end{aligned}$$

□

– (1) $\boldsymbol{V}\boldsymbol{A}_1\boldsymbol{V}(\boldsymbol{A}_0 - \boldsymbol{A}_1)\boldsymbol{V} = \boldsymbol{0}$
(2) If $\boldsymbol{X}\boldsymbol{b} \in \mathcal{C}(\boldsymbol{X}_0, \boldsymbol{V})$, then $\boldsymbol{V}\boldsymbol{A}_1\boldsymbol{V}(\boldsymbol{A}_0 - \boldsymbol{A}_1)\boldsymbol{X}\boldsymbol{b} = \boldsymbol{0}$.
(3) $\boldsymbol{V}(\boldsymbol{A}_0 - \boldsymbol{A}_1)\boldsymbol{V}\boldsymbol{A}_1\boldsymbol{X}\boldsymbol{b} = \boldsymbol{0}$
(4) $\boldsymbol{b}^T\boldsymbol{X}^T(\boldsymbol{A}_0 - \boldsymbol{A}_1)\boldsymbol{V}\boldsymbol{A}_1\boldsymbol{X}\boldsymbol{b} = 0$.

Remark: The third part of the theorem (independence) is checked by this result.

Proof. For (1),

$$\begin{aligned} & \boldsymbol{V}\boldsymbol{A}_1\boldsymbol{V}(\boldsymbol{A}_0 - \boldsymbol{A}_1)\boldsymbol{V} \\ = \; & \boldsymbol{V}\boldsymbol{A}_1\boldsymbol{V}\boldsymbol{A}_0\boldsymbol{V} - \boldsymbol{V}\boldsymbol{A}_1\boldsymbol{V}\boldsymbol{A}_1\boldsymbol{V} \\ = \; & \boldsymbol{V}\boldsymbol{A}_1\boldsymbol{V} - \boldsymbol{V}\boldsymbol{A}_1\boldsymbol{V} \\ = \; & \boldsymbol{0}. \end{aligned}$$

(2) follows from the proof in the preceding result.
(3) and (4) are trivial since $\boldsymbol{A}_1\boldsymbol{X} = \boldsymbol{0}$. □

- Before we state the F-test, we show that the constrained model is consistent if and only if $\boldsymbol{X}\boldsymbol{b} \in \mathcal{C}(\boldsymbol{X}_0)$.

1. (Null model) If $\boldsymbol{X}\boldsymbol{b} \in \mathcal{C}(\boldsymbol{X}_0)$, then $\mathbf{P}(\boldsymbol{Y} \in \mathcal{C}(\boldsymbol{X}_0, \boldsymbol{V})) = 1$ and

$$\boldsymbol{b}^T\boldsymbol{X}^T\boldsymbol{A}_0\boldsymbol{X}\boldsymbol{b} = 0.$$

2. (Alternative model) If $\boldsymbol{X}\boldsymbol{b} \notin \mathcal{C}(\boldsymbol{X}_0)$, then either
   – $\boldsymbol{X}\boldsymbol{b} \in \mathcal{C}(\boldsymbol{X}_0, \boldsymbol{V})$, $\mathbf{P}(\boldsymbol{Y} \in \mathcal{C}(\boldsymbol{X}_0, \boldsymbol{V})) = 1$, and $\boldsymbol{b}^T\boldsymbol{X}^T\boldsymbol{A}_0\boldsymbol{X}\boldsymbol{b} > 0$, or
   – $\boldsymbol{X}\boldsymbol{b} \notin \mathcal{C}(\boldsymbol{X}_0, \boldsymbol{V})$ and $\mathbf{P}(\boldsymbol{Y} \notin \mathcal{C}(\boldsymbol{X}_0, \boldsymbol{V})) = 1$.

Proof. For Part 1, note $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{e}$, where $\boldsymbol{e} \in \mathcal{C}(\boldsymbol{V})$ and $\boldsymbol{A}_0\boldsymbol{X}_0 = \boldsymbol{0}$. For part 2

– If $\boldsymbol{X}\boldsymbol{b} \in \mathcal{C}(\boldsymbol{X}_0, \boldsymbol{V})$, we need to show $\boldsymbol{b}^T\boldsymbol{X}^T\boldsymbol{A}_0\boldsymbol{X}\boldsymbol{b} > 0$. Here is a proof thanks to Po Ning. Suppose

$$\boldsymbol{b}^T\boldsymbol{X}^T\boldsymbol{A}_0\boldsymbol{X}\boldsymbol{b} = \boldsymbol{b}^T\boldsymbol{X}^T[\boldsymbol{V}_0^- - \boldsymbol{V}_0^-\boldsymbol{X}_0(\boldsymbol{X}_0^T\boldsymbol{V}_0^-\boldsymbol{X}_0)^-\boldsymbol{X}_0^T\boldsymbol{V}_0^-]\boldsymbol{X}\boldsymbol{b} = 0.$$

Note this quantity is invariant to the choice of generalized inverse $\boldsymbol{V}_0^-$ due to the fact $\boldsymbol{X}\boldsymbol{b} \in \mathcal{C}(\boldsymbol{X}_0, \boldsymbol{V}) = \mathcal{C}(\boldsymbol{V}_0)$ and $\mathcal{C}(\boldsymbol{X}_0) \subset \mathcal{C}(\boldsymbol{V}_0)$. Without loss of

generality, we can use $\boldsymbol{V}_0^+$, which is psd and thus can be decomposed as $\boldsymbol{V}^+ = \boldsymbol{L}\boldsymbol{L}^T$, $\boldsymbol{L} \in \mathbb{R}^{n\times r}$ full column rank. Therefore

$$\begin{aligned}
& \boldsymbol{b}^T\boldsymbol{X}^T[\boldsymbol{V}_0^- - \boldsymbol{V}_0^-\boldsymbol{X}_0(\boldsymbol{X}_0^T\boldsymbol{V}_0^-\boldsymbol{X}_0)^-\boldsymbol{X}_0^T\boldsymbol{V}_0^-]\boldsymbol{X}\boldsymbol{b} \\
= \; & \boldsymbol{b}^T\boldsymbol{X}^T[\boldsymbol{V}_0^+ - \boldsymbol{V}_0^+\boldsymbol{X}_0(\boldsymbol{X}_0^T\boldsymbol{V}_0^+\boldsymbol{X}_0)^-\boldsymbol{X}_0^T\boldsymbol{V}_0^+]\boldsymbol{X}\boldsymbol{b} \\
= \; & \boldsymbol{b}^T\boldsymbol{X}^T[\boldsymbol{L}\boldsymbol{L}^T - \boldsymbol{L}\boldsymbol{L}^T\boldsymbol{X}_0(\boldsymbol{X}_0^T\boldsymbol{V}_0^+\boldsymbol{X}_0)^-\boldsymbol{X}_0^T\boldsymbol{L}\boldsymbol{L}^T]\boldsymbol{X}\boldsymbol{b} \\
= \; & \boldsymbol{b}^T\boldsymbol{X}^T\boldsymbol{L}[\boldsymbol{I} - \boldsymbol{L}^T\boldsymbol{X}_0(\boldsymbol{X}_0^T\boldsymbol{L}\boldsymbol{L}^T\boldsymbol{X}_0)^-\boldsymbol{X}_0^T\boldsymbol{L}]\boldsymbol{L}^T\boldsymbol{X}\boldsymbol{b} \\
= \; & \boldsymbol{b}^T\boldsymbol{X}^T\boldsymbol{L}(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{Z}})\boldsymbol{L}^T\boldsymbol{X}\boldsymbol{b} \quad (\text{let } \boldsymbol{Z} = \boldsymbol{L}^T\boldsymbol{X}_0) \\
= \; & \boldsymbol{b}^T\boldsymbol{X}^T\boldsymbol{L}(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{Z}})(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{Z}})\boldsymbol{L}^T\boldsymbol{X}\boldsymbol{b} \\
= \; & \|(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{Z}})\boldsymbol{L}^T\boldsymbol{X}\boldsymbol{b}\|_2^2 \\
= \; & 0,
\end{aligned}$$

implying that $(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{Z}})\boldsymbol{L}^T\boldsymbol{X}\boldsymbol{b} = \boldsymbol{0}$ and thus

$$\boldsymbol{L}^T\boldsymbol{X}\boldsymbol{b} \in \mathcal{C}(\boldsymbol{Z}) = \mathcal{C}(\boldsymbol{L}^T\boldsymbol{X}_0).$$

Then what?

– If $\boldsymbol{X}\boldsymbol{b} \notin \mathcal{C}(\boldsymbol{X}_0, \boldsymbol{V})$, then $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{e}$ cannot be in $\mathcal{C}(\boldsymbol{X}_0, \boldsymbol{V})$ since $\boldsymbol{e} \in \mathcal{C}(\boldsymbol{V})$. If it does, then $\boldsymbol{X}\boldsymbol{b} = \boldsymbol{Y} - \boldsymbol{e} \in \mathcal{C}(\boldsymbol{X}_0, \boldsymbol{V})$, a contradiction. Therefore $\mathbf{P}(\boldsymbol{Y} \notin \mathcal{C}(\boldsymbol{X}_0, \boldsymbol{V})) = 1$.

□

- (F test under Aitken model) The test rejects $H_0 : \boldsymbol{y} \sim N(\boldsymbol{X}_0\boldsymbol{c}, \sigma^2\boldsymbol{V})$
  - if $\boldsymbol{y} \notin \mathcal{C}(\boldsymbol{X}_0, \boldsymbol{V})$, or
  - if $\boldsymbol{y} \in \mathcal{C}(\boldsymbol{X}_0, \boldsymbol{V})$ and

$$\frac{(\text{SSE}_0 - \text{SSE})/\text{tr}((\boldsymbol{A}_0 - \boldsymbol{A}_1)\boldsymbol{V})}{\text{SSE}/\text{tr}(\boldsymbol{A}_1\boldsymbol{V})} > F_{\text{tr}((\boldsymbol{A}_0-\boldsymbol{A}_1)\boldsymbol{V}),\text{tr}(\boldsymbol{A}_1\boldsymbol{V}),\alpha}.$$

It is a level-α test because $\mathbf{P}(\boldsymbol{Y} \notin \mathcal{C}(\boldsymbol{X}_0, \boldsymbol{V})) = 0$ under H_0. The power of the test is always greater than α since if $\boldsymbol{b}^T\boldsymbol{X}^T\boldsymbol{A}_0\boldsymbol{X}\boldsymbol{b} = 0$, then the test will reject H_0 with probability 1.

19 Lecture 19: Nov 11

Announcement

- HW6 returned.
- HW7 due today.
- Midterm 2 this Wed Nov 13 @ 11:45AM-1PM.

Last time

Testing under Aitken model $\boldsymbol{y} \sim N(\boldsymbol{X}\boldsymbol{b}, \sigma^2\boldsymbol{V})$ – main results and proofs.

Today

- Q&A
- Estimation of σ^2.

Q&A for midterm 2

- Notes p101 about independence of numerator and denominator of the F statistic.
- Notes p103 about cancellation of $\boldsymbol{T}$.

MVQUE of σ^2

- So far we have focused on estimation and testing of $\boldsymbol{b}$. What's a good estimator of σ^2? We only know that the *least squares estimator*

  $$\hat{\sigma}^2 = \text{SSE}/(n - \text{rank}(\boldsymbol{X}))$$

  is unbiased under the Gauss-Markov assumption.

  With additional normality assumption, does this have minimum variance within certain class of estimators? Does this have minimum mean square error (MSE) within certain class of estimators?

- Since σ^2 is a quadratic concept, we consider estimation of σ^2 by a quadratic function of $\boldsymbol{y}$.

  We call any estimator $\boldsymbol{y}^T\boldsymbol{A}\boldsymbol{y}$, where $\boldsymbol{A}$ is symmetric and non-stochastic, a *quadratic estimator*.

- Any estimator $\hat{\sigma}^2$ such that

$$\mathrm{E}(\hat{\sigma}^2) = \sigma^2 \quad \text{for all } \boldsymbol{b} \in \mathbb{R}^p \text{ and } \sigma^2 > 0$$

  is called an *unbiased estimator* of σ^2.

- The *minimum variance unbiased estimator* (MVQUE, also called BQUE) of σ^2 is a quadratic unbiased estimator of σ^2, say $\hat{\sigma}^2$, such that

$$\mathrm{Var}(\hat{\sigma}^2) \leq \mathrm{Var}(\hat{\tau}^2)$$

  for all quadratic unbiased estimators $\hat{\tau}^2$ of $\hat{\sigma}^2$.

- Since we are dealing with quadratic function of a multivariate normal, we recall a useful result (shown in HW5).

  Let $\boldsymbol{Y} \sim N(\boldsymbol{\mu}, \boldsymbol{V})$, $\boldsymbol{V}$ nonsingular, and let $U = \boldsymbol{Y}^T\boldsymbol{A}\boldsymbol{Y}$ for $\boldsymbol{A}$ symmetric.

  1. The mgf of U is

$$\begin{aligned} m_U(t) &= |\boldsymbol{I} - 2t\boldsymbol{A}\boldsymbol{V}|^{-1/2} e^{-\boldsymbol{\mu}^T\boldsymbol{V}^{-1}\boldsymbol{\mu}/2 + \boldsymbol{\mu}^T(\boldsymbol{V}-2t\boldsymbol{V}\boldsymbol{A}\boldsymbol{V})^{-1}\boldsymbol{\mu}/2} \\ &= |\boldsymbol{I} - 2t\boldsymbol{A}\boldsymbol{V}|^{-1/2} e^{-t\boldsymbol{\mu}^T(\boldsymbol{I}-2t\boldsymbol{A}\boldsymbol{V})^{-1}\boldsymbol{A}\boldsymbol{\mu}}. \end{aligned}$$

  2. The mean and variance of U is

$$\begin{aligned} \mathrm{E}\, U &= \mathrm{tr}(\boldsymbol{A}\boldsymbol{V}) + \boldsymbol{\mu}^T\boldsymbol{A}\boldsymbol{\mu} \\ \mathrm{Var}\, U &= 2\mathrm{tr}(\boldsymbol{A}\boldsymbol{V})^2 + 4\boldsymbol{\mu}^T\boldsymbol{A}\boldsymbol{V}\boldsymbol{A}\boldsymbol{\mu}. \end{aligned}$$

- Under Gauss-Markov normal model $\boldsymbol{y} \sim N(\boldsymbol{X}\boldsymbol{b}, \sigma^2\boldsymbol{I})$, the least squares estimator

$$\hat{\sigma}^2 = \frac{\mathrm{SSE}}{n - \mathrm{rank}(\boldsymbol{X})} = \frac{\boldsymbol{y}^T(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{y}}{n - \mathrm{rank}(\boldsymbol{X})}$$

  is MVQUE.

Proof. Let $\hat{\sigma}^2 = \boldsymbol{y}^T\boldsymbol{A}\boldsymbol{y}$ be a quadratic estimator of σ^2. Then

$$\begin{aligned}\hat{\sigma}^2 &= (\boldsymbol{X}\boldsymbol{b}+\boldsymbol{e})^T\boldsymbol{A}(\boldsymbol{X}\boldsymbol{b}+\boldsymbol{e}) \\ &= \boldsymbol{b}^T\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X}\boldsymbol{b} + 2\boldsymbol{b}^T\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{e} + \boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}\end{aligned}$$

and

$$\mathrm{E}(\hat{\sigma}^2) = \boldsymbol{b}^T\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X}\boldsymbol{b} + \sigma^2\mathrm{tr}\boldsymbol{A} = \sigma^2$$

for all $\boldsymbol{b}$ and $\sigma^2 > 0$ if and only if

$$\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X} = \boldsymbol{0}_{p\times p} \quad \text{and} \quad \mathrm{tr}\boldsymbol{A} = 1.$$

The variance of $\hat{\sigma}^2$ is (why?)

$$\begin{aligned}\mathrm{Var}(\hat{\sigma}^2) &= \mathrm{Var}(\boldsymbol{y}^T\boldsymbol{A}\boldsymbol{y}) \\ &= 2\sigma^4(\mathrm{tr}\boldsymbol{A}^2 + 2\boldsymbol{\gamma}^T\boldsymbol{X}^T\boldsymbol{A}^2\boldsymbol{X}\boldsymbol{\gamma}),\end{aligned}$$

where $\boldsymbol{\gamma} = \boldsymbol{b}/\sigma$. To seek the MVQUE, we solve the optimization problem

$$\begin{aligned}&\text{minimize} \quad \mathrm{tr}\boldsymbol{A}^2 + 2\boldsymbol{\gamma}^T\boldsymbol{X}^T\boldsymbol{A}^2\boldsymbol{X}\boldsymbol{\gamma} \\ &\text{subject to} \quad \boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X} = \boldsymbol{0} \text{ and } \mathrm{tr}\boldsymbol{A} = 1.\end{aligned}$$

(It turns out the optimal $\boldsymbol{A}$ is independent of the unknown parameter γ.) We form the Lagrangian

$$L(\boldsymbol{A}, \lambda, \boldsymbol{L}) = \frac{1}{2}\mathrm{tr}\boldsymbol{A}^2 + \boldsymbol{\gamma}^T\boldsymbol{X}^T\boldsymbol{A}^2\boldsymbol{X}\boldsymbol{\gamma} - \lambda(\mathrm{tr}\boldsymbol{A} - 1) - \mathrm{tr}(\boldsymbol{L}^T\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X}),$$

where λ and $\boldsymbol{L} \in \mathbb{R}^{p\times p}$ are Lagrange multipliers. Since $\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X}$ is symmetric, we assume $\boldsymbol{L}$ is symmetric too. Setting derivative of the Lagrangian to 0 gives

$$\begin{aligned}\boldsymbol{A} + \boldsymbol{X}\boldsymbol{\gamma}\boldsymbol{\gamma}^T\boldsymbol{X}^T\boldsymbol{A} + \boldsymbol{A}\boldsymbol{X}\boldsymbol{\gamma}\boldsymbol{\gamma}^T\boldsymbol{X}^T - \lambda\boldsymbol{I} &= \boldsymbol{X}\boldsymbol{L}\boldsymbol{X}^T \\ \boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X} &= \boldsymbol{0}_{p\times p} \\ \mathrm{tr}\boldsymbol{A} &= 1\end{aligned}$$

(Check Lecture 8 p56 for derivatives of trace functions.) Pre- and post-multiplying the first equation by $\boldsymbol{P}_{\boldsymbol{X}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T$ yields

$$-\lambda\boldsymbol{P}_{\boldsymbol{X}} = \boldsymbol{X}^T\boldsymbol{L}\boldsymbol{X}.$$

Substitution back to the first equation shows

$$\boldsymbol{A} = \lambda(\boldsymbol{I} - \boldsymbol{P_X}) - (\boldsymbol{X}\boldsymbol{\gamma}\boldsymbol{\gamma}^T\boldsymbol{X}^T\boldsymbol{A} + \boldsymbol{A}\boldsymbol{X}\boldsymbol{\gamma}\boldsymbol{\gamma}^T\boldsymbol{X}^T).$$

Taking trace on both sides gives $\mathrm{tr}(\boldsymbol{A}) = \lambda(n-r) = 1$, i.e.,

$$\lambda = \frac{1}{n-r}.$$

Taking square on both sides gives

$$\begin{aligned}\boldsymbol{A}^2 &= \lambda^2(\boldsymbol{I} - \boldsymbol{P_X}) - \lambda\boldsymbol{A}\boldsymbol{X}\boldsymbol{\gamma}\boldsymbol{\gamma}^T\boldsymbol{X}^T \\ &\quad -\lambda\boldsymbol{X}\boldsymbol{\gamma}\boldsymbol{\gamma}^T\boldsymbol{X}^T\boldsymbol{A} + \boldsymbol{X}\boldsymbol{\gamma}\boldsymbol{\gamma}^T\boldsymbol{X}^T\boldsymbol{A}^2\boldsymbol{X}\boldsymbol{\gamma}\boldsymbol{\gamma}^T\boldsymbol{X}^T \\ &\quad +\boldsymbol{A}\boldsymbol{X}\boldsymbol{\gamma}\boldsymbol{\gamma}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\gamma}\boldsymbol{\gamma}^T\boldsymbol{X}^T\boldsymbol{A}.\end{aligned}$$

Taking trace on both sides

$$\mathrm{tr}(\boldsymbol{A}^2) = \frac{1}{n-r} + 2(\boldsymbol{\gamma}^T\boldsymbol{X}^T\boldsymbol{A}^2\boldsymbol{X}\boldsymbol{\gamma})(\boldsymbol{\gamma}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\gamma}).$$

Thus the objective function can be expressed as

$$\begin{aligned}&\mathrm{tr}(\boldsymbol{A}^2) + 2\boldsymbol{\gamma}^T\boldsymbol{X}^T\boldsymbol{A}^2\boldsymbol{X}\boldsymbol{\gamma} \\ = \;& \frac{1}{n-r} + 2(\boldsymbol{\gamma}^T\boldsymbol{X}^T\boldsymbol{A}^2\boldsymbol{X}\boldsymbol{\gamma})(1 + \boldsymbol{\gamma}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\gamma}),\end{aligned}$$

which is minimized by taking $\boldsymbol{A}\boldsymbol{X}\boldsymbol{\gamma} = \boldsymbol{0}$. Therefore

$$\boldsymbol{A} = \lambda(\boldsymbol{I} - \boldsymbol{P_X}) = \frac{\boldsymbol{I} - \boldsymbol{P_X}}{n-r}.$$

□

20 Lecture 20: Nov 18

Last time

Under Gauss-Markov normal model, the least squares estimator $\hat{\sigma}^2 = \text{SSE}/(n-r)$ is MVQUE.

Today

- Best quadratic invariant estimator of σ^2.
- Variance components and mixed model (Chapter 8): introduction.

Best quadratic invariant estimation of σ^2

- In this section, we seek the quadratic estimator of σ^2 with smallest mean squared error (MSE), within the class of invariant quadratic estimators. Recall that the MSE of an estimator of σ^2 is

$$\begin{aligned} \text{E}(\hat{\sigma}^2 - \sigma^2)^2 &= \text{E}[\hat{\sigma}^2 - \text{E}(\hat{\sigma}^2) + \text{E}(\hat{\sigma}^2) - \sigma^2] \\ &= \text{Var}(\hat{\sigma}^2) + [\text{E}(\hat{\sigma}^2) - \sigma^2]^2 \\ &= \text{Variance} + \text{Bias}^2. \end{aligned}$$

- Assume $\boldsymbol{y} \sim N(\boldsymbol{X}\boldsymbol{b}, \sigma^2\boldsymbol{V})$, $\boldsymbol{V}$ nonsingular. Let $\boldsymbol{y}^T\boldsymbol{A}\boldsymbol{y}$ be a quadratic estimator. In HW5, we showed that its first two moments are

$$\begin{aligned} \text{E}(\boldsymbol{y}^T\boldsymbol{A}\boldsymbol{y}) &= \boldsymbol{b}^T\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X}\boldsymbol{b} + \sigma^2\text{tr}(\boldsymbol{A}\boldsymbol{V}) \\ \text{Var}(\boldsymbol{y}^T\boldsymbol{A}\boldsymbol{y}) &= 4\sigma^2\boldsymbol{b}^T\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{V}\boldsymbol{A}\boldsymbol{X}\boldsymbol{b} + 2\sigma^4\text{tr}(\boldsymbol{A}\boldsymbol{V})^2 \end{aligned}$$

  and its mgf is

$$\begin{aligned} m(t) &= |\boldsymbol{I} - 2t\boldsymbol{A}\boldsymbol{V}|^{-1/2} e^{-\boldsymbol{b}^T\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}\boldsymbol{b}/2 + \boldsymbol{b}^T\boldsymbol{X}^T(\boldsymbol{V}-2t\boldsymbol{V}\boldsymbol{A}\boldsymbol{V})^{-1}\boldsymbol{X}\boldsymbol{b}/2} \\ &= |\boldsymbol{I} - 2t\boldsymbol{A}\boldsymbol{V}|^{-1/2} e^{-t\boldsymbol{b}^T\boldsymbol{X}^T(\boldsymbol{I}-2t\boldsymbol{A}\boldsymbol{V})^{-1}\boldsymbol{A}\boldsymbol{X}\boldsymbol{b}}. \end{aligned}$$

  Thus distribution is independent of $\boldsymbol{b}$ if and only if $\boldsymbol{A}\boldsymbol{X} = \boldsymbol{0}_{n\times p}$.

  Proof. The "if" part is trivial from the mgf. For the "only if" part, $\boldsymbol{b}^T\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{V}\boldsymbol{A}\boldsymbol{X}\boldsymbol{b} = 0$ for all $\boldsymbol{b}$ implies $\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{V}\boldsymbol{A}\boldsymbol{X} = \boldsymbol{0}_{p\times p}$. Since $\boldsymbol{V}$ is nonsingular, $\boldsymbol{A}\boldsymbol{X} = \boldsymbol{0}$. □

- We say a quadratic estimator $\boldsymbol{y}^T\boldsymbol{A}\boldsymbol{y}$ is *invariant* under translation of $\boldsymbol{b}$ if $\boldsymbol{A}\boldsymbol{X} = \boldsymbol{0}$.

  A quadratic invariant estimator $\hat{\sigma}^2$ is the *best quadratic invariant estimator* of σ^2 if

  $$\mathrm{E}(\hat{\sigma}^2 - \sigma^2)^2 \leq \mathrm{E}(\hat{\tau}^2 - \sigma^2)^2$$

  for all quadratic invariant estimator $\hat{\tau}^2$ of σ^2.

- Under Gauss-Markov normal model $\boldsymbol{y} \sim N(\boldsymbol{X}\boldsymbol{b}, \sigma^2\boldsymbol{I})$, the best quadratic invariant estimator of σ^2 is

  $$\hat{\sigma}^2 = \frac{\boldsymbol{y}^T(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{y}}{n - r + 2} = \frac{\mathrm{SSE}}{n - r + 2}.$$

  Remark: By introducing bias (shrinkage) to the unbiased estimator $\mathrm{SSE}/(n-r)$, we achieve optimal MSE.

  Proof. Let $\hat{\sigma}^2 = \boldsymbol{y}^T\boldsymbol{A}\boldsymbol{y}$ be a quadratic estimator. Invariance imposes $\boldsymbol{A}\boldsymbol{X} = \boldsymbol{0}$. Thus

  $$\hat{\sigma}^2 = \boldsymbol{y}^T\boldsymbol{A}\boldsymbol{y} = (\boldsymbol{X}\boldsymbol{b} + \boldsymbol{e})^T\boldsymbol{A}(\boldsymbol{X}\boldsymbol{b} + \boldsymbol{e}) = \boldsymbol{e}^T\boldsymbol{A}\boldsymbol{e}$$

  and the MSE is (why?)

  $$\mathrm{E}(\hat{\sigma}^2 - \sigma^2)^2 = 2\sigma^4\mathrm{tr}\boldsymbol{A}^2 + \sigma^4(1 - \mathrm{tr}\boldsymbol{A})^2.$$

  We seek the best quadratic invariant estimator by solving

  $$\begin{aligned} \text{minimize} \quad & 2\mathrm{tr}\boldsymbol{A}^2 + (1 - \mathrm{tr}\boldsymbol{A})^2 \\ \text{subject to} \quad & \boldsymbol{A}\boldsymbol{X} = \boldsymbol{0}. \end{aligned}$$

  We form the Lagrangian

  $$L(\boldsymbol{A}, \boldsymbol{L}) = \mathrm{tr}\boldsymbol{A}^2 + \frac{1}{2}(1 - \mathrm{tr}\boldsymbol{A})^2 - \mathrm{tr}(\boldsymbol{L}^T\boldsymbol{A}\boldsymbol{X})$$

  and set its gradient to zero

  $$\begin{aligned} 2\boldsymbol{A} - (1 - \mathrm{tr}\boldsymbol{A})\boldsymbol{I}_n &= \boldsymbol{L}\boldsymbol{X}^T \\ \boldsymbol{A}\boldsymbol{X} &= \boldsymbol{0}. \end{aligned}$$

Pre-multiplying the first equation by $\boldsymbol{A}$ yields

$$2\boldsymbol{A}^2 - (1 - \text{tr}\boldsymbol{A})\boldsymbol{A} = \boldsymbol{A}\boldsymbol{L}\boldsymbol{X}^T.$$

Post-multiplying by $\boldsymbol{P}_{\boldsymbol{X}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T$ shows

$$\boldsymbol{A}\boldsymbol{L}\boldsymbol{X}^T = \boldsymbol{0}.$$

Substitution back gives

$$\boldsymbol{A}^2 = \frac{1 - \text{tr}\boldsymbol{A}}{2}\boldsymbol{A}.$$

Let $\text{rank}(\boldsymbol{A}) = \rho$. By a result shown in Lecture 3 (p16),

$$\text{tr}\boldsymbol{A} = \frac{1 - \text{tr}\boldsymbol{A}}{2}\text{rank}(\boldsymbol{A}) = \frac{1 - \text{tr}\boldsymbol{A}}{2}\rho.$$

Thus

$$\text{tr}(\boldsymbol{A}) = \frac{\rho}{\rho + 2} \quad \text{and} \quad \text{tr}(\boldsymbol{A}^2) = \frac{\rho}{(\rho + 2)^2}.$$

The objective function becomes

$$\frac{2}{\rho + 2}.$$

Since $\boldsymbol{A}\boldsymbol{X} = \boldsymbol{0}$, the maximing ρ is $n - \text{rank}(\boldsymbol{X}) = n - r$. Note $(n - r + 2)\boldsymbol{A}$ is symmetric, idempotent, and orthogonal to $\mathcal{C}(\boldsymbol{X})$, thus equals $\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}}$. Therefore

$$\boldsymbol{A} = (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}})/(n - r + 2).$$

□

- In summary, we have the following estimators for σ^2
  - MVQUE (aka least square estimator): SSE$/(n - r)$
  - MLE: SSE$/n$
  - Best quadratic invariant estimate: SSE$/(n - r + 2)$

Variance components and mixed models

- Traditionally, linear models have been divided into three categories:
  - *Fixed effects model*: $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{e}$, where $\boldsymbol{b}$ is fixed.
  - *Random effects model*: $\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e}$, where $\boldsymbol{u}$ is random.
  - *Mixed effects model*: $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e}$, where $\boldsymbol{b}$ is fixed and $\boldsymbol{u}$ is random.
- In a mixed effects model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e}$$

  - $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ is a design matrix for fixed effects $\boldsymbol{b} \in \mathbb{R}^p$.
  - $\boldsymbol{Z} \in \mathbb{R}^{n \times q}$ is a design matrix for random effects $\boldsymbol{u} \in \mathbb{R}^q$.
  - The most general assumption is $\boldsymbol{e} \in N(\boldsymbol{0}_n, \boldsymbol{R})$, $\boldsymbol{u} \in N(\boldsymbol{0}_q, \boldsymbol{G})$, and $\boldsymbol{e}$ is independent of $\boldsymbol{u}$.

  In many applications, $\boldsymbol{e} \sim N(\boldsymbol{0}_n, \sigma^2 \boldsymbol{I}_n)$ and

$$\boldsymbol{Z}\boldsymbol{u} = \begin{pmatrix} \boldsymbol{Z}_1 & \cdots & \boldsymbol{Z}_m \end{pmatrix} \begin{pmatrix} \boldsymbol{u}_1 \\ \vdots \\ \boldsymbol{u}_m \end{pmatrix} = \boldsymbol{Z}_1\boldsymbol{u}_1 + \cdots + \boldsymbol{Z}_m\boldsymbol{u}_m,$$

  where $\boldsymbol{u}_i \sim N(\boldsymbol{0}_{q_i}, \sigma_i^2 \boldsymbol{I}_{q_i})$, $\sum_{i=1}^m q_i = q$. $\boldsymbol{e}$ and $\boldsymbol{u}_i$, $i = 1, \ldots, m$, are jointly independent. Then the covariance of responses $\boldsymbol{y}$

$$\boldsymbol{V}(\sigma^2, \sigma_1^2, \ldots, \sigma_m^2) = \sigma^2 \boldsymbol{I} + \sum_{i=1}^m \sigma_i^2 \boldsymbol{Z}_i \boldsymbol{Z}_i^T$$

  is a function of the *variance components* $(\sigma^2, \sigma_1^2, \ldots, \sigma_m^2)$.
- Primary goal of the mixed model (aka variance components model) is to
  - estimation and testing of the fixed effects $\boldsymbol{b}$
  - estimation and testing of the variance components $(\sigma^2, \sigma_1^2, \ldots, \sigma_m^2)$
  - prediction

- Example: One-way ANOVA with random treatment effects (JM 8.2)

$$y_{ij} = \mu + \alpha_i + e_{ij}, \quad i = 1, \ldots, a, j = 1, \ldots, n_i, \sum_{i=1}^{a} n_i = n.$$

So far we considered the fixed effects model where the treatment effects α_i are assumed to be constant. In random effects model, the common assumption is

$$\begin{aligned} \alpha_i &\sim N(0, \sigma_a^2) \\ e_{ij} &\sim N(0, \sigma^2) \end{aligned}$$

are jointly independent. Equivalently the model is

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{X}\boldsymbol{b} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e} \\ &= \mathbf{1}_n \mu + \begin{pmatrix} \mathbf{1}_{n_1} & & & \\ & \mathbf{1}_{n_2} & & \\ & & \ddots & \\ & & & \mathbf{1}_{n_a} \end{pmatrix} \boldsymbol{u} + \boldsymbol{e}, \end{aligned}$$

where $\boldsymbol{u} \sim N(\mathbf{0}_a, \sigma_a^2 \boldsymbol{I}_a)$ and $\boldsymbol{e} \sim N(\mathbf{0}_n, \sigma^2 \boldsymbol{I}_n)$ are independent. The covariance matrix is

$$\boldsymbol{V} = \text{BlkDiag}(\boldsymbol{V}_1, \ldots, \boldsymbol{V}_a) = \begin{pmatrix} \boldsymbol{V}_1 & & & \\ & \boldsymbol{V}_2 & & \\ & & \ddots & \\ & & & \boldsymbol{V}_a \end{pmatrix},$$

where $\boldsymbol{V}_i = \sigma^2 \boldsymbol{I}_{n_i} + \sigma_a^2 \mathbf{1}_{n_i} \mathbf{1}_{n_i}^T$.

The model parameters are $(\mu, \sigma^2, \sigma_a^2)$.

- Example: Two-way ANOVA with random effects

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}, \quad i = 1, \ldots, a,\ j = 1, \ldots, b,\ k = 1, \ldots, n_{ij},$$

where

$$\begin{aligned} \alpha_i &\sim N(0, \sigma_a^2), \quad \beta_j \sim N(0, \sigma_b^2) \\ \gamma_{ij} &\sim N(0, \sigma_c^2), \quad e_{ijk} \sim N(0, \sigma^2) \end{aligned}$$

are jointly independent.

Equivalently the model is

$$\begin{aligned}
\boldsymbol{y} &= \mathbf{1}_n\mu + \begin{pmatrix} \mathbf{1}_{n_{11}} & & \\ \mathbf{1}_{n_{12}} & & \\ \vdots & & \\ \mathbf{1}_{n_{1b}} & & \\ \vdots & & \vdots \\ & & \mathbf{1}_{n_{a1}} \\ & & \mathbf{1}_{n_{a2}} \\ & & \vdots \\ & & \mathbf{1}_{n_{ab}} \end{pmatrix} \boldsymbol{u}_a + \begin{pmatrix} \mathbf{1}_{n_{11}} & & & & \\ & \mathbf{1}_{n_{12}} & & & \\ & & \ddots & & \\ & & & \mathbf{1}_{1b} & \\ \vdots & & & & \vdots \\ \mathbf{1}_{n_{a1}} & & & & \\ & \mathbf{1}_{n_{a2}} & & & \\ & & \ddots & & \\ & & & & \mathbf{1}_{ab} \end{pmatrix} \boldsymbol{u}_b \\
&+ \begin{pmatrix} \mathbf{1}_{n_{11}} & & & & & & & & & \\ & \mathbf{1}_{n_{12}} & & & & & & & & \\ & & \ddots & & & & & & & \\ & & & \mathbf{1}_{1b} & & & & & & \\ & & & & \ddots & & & & & \\ & & & & & \mathbf{1}_{n_{a1}} & & & & \\ & & & & & & \mathbf{1}_{n_{a2}} & & & \\ & & & & & & & \ddots & & \\ & & & & & & & & \mathbf{1}_{ab} \end{pmatrix} \boldsymbol{u}_c + \boldsymbol{e},
\end{aligned}$$

where

$$\begin{aligned}
\boldsymbol{u}_a &\sim N(\mathbf{0}_a, \sigma_a^2\boldsymbol{I}_a), \quad \boldsymbol{u}_b \sim N(\mathbf{0}_b, \sigma_b^2\boldsymbol{I}_b), \\
\boldsymbol{u}_c &\sim N(\mathbf{0}_{ab}, \sigma_c^2\boldsymbol{I}_{ab}), \quad \boldsymbol{e} \sim N(\mathbf{0}_n, \sigma^2\boldsymbol{I}_n)
\end{aligned}$$

are jointly independent.

The model parameters are $(\mu, \sigma^2, \sigma_a^2, \sigma_b^2, \sigma_c^2)$.

- We may also have two-way ANOVA mixed model, where one factor has fixed effects and the other factor has random effects (JM Example 8.1 p190).

21 Lecture 21: Nov 20

Announcement

- HW7 returned.
- Midterm 2 returned (85.63 ± 6.32).
- HW8 posted and due Mon Dec 2.
- No afternoon session today.
- Email your comments (compliments?) on TA to me.

Last time

- Best (minimal MSE) quadratic invariant estimator of σ^2.
- Variance components and mixed model (Chapter 8): introduction.

Today

- Variance component estimation: MLE.
- Variance component estimation: REML.

Variance component estimation: MLE (JM 8.4.1)

In this section, we pursue MLE for the variance components model

$$\boldsymbol{y} \sim N_n(\boldsymbol{X}\boldsymbol{b}, \boldsymbol{V}),$$

where

$$\boldsymbol{V} = \sum_{i=0}^{m} \sigma_i^2 \boldsymbol{V}_i,$$

with $\boldsymbol{V}_i$ psd and $\boldsymbol{V}$ nonsingular. Parameters are $\boldsymbol{b}$ and $(\sigma_0^2, \sigma_1^2, \ldots, \sigma_m^2)$.

- Assume $\boldsymbol{V}$ is nonsingular, then the log-likelihood function is

$$\begin{aligned} & L(\boldsymbol{b}, \sigma_0^2, \sigma_1^2, \ldots, \sigma_m^2) \\ = & -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln\det(\boldsymbol{V}) - \frac{1}{2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{b})^T\boldsymbol{V}^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{b}). \end{aligned}$$

- In general we need some iterative optimization algorithm (Fisher scoring, EM algorithm) to find the MLE.

- The following facts are useful when deriving the derivatives of log-likelihood.

  1. The partial derivative of a matrix $\boldsymbol{B} = (b_{ij})$ with respect to a scalar parameter θ is the matrix

$$\frac{\partial}{\partial\theta}\boldsymbol{B} = \left(\frac{\partial}{\partial\theta}b_{ij}\right).$$

  2. Because the trace function is linear,

$$\frac{\partial}{\partial\theta}\mathrm{tr}(\boldsymbol{B}) = \mathrm{tr}\left(\frac{\partial}{\partial\theta}\boldsymbol{B}\right).$$

  3. The product rule of differentiation implies

$$\frac{\partial}{\partial\theta}(\boldsymbol{B}\boldsymbol{C}) = \left(\frac{\partial}{\partial\theta}\boldsymbol{B}\right)\boldsymbol{C} + \boldsymbol{B}\left(\frac{\partial}{\partial\theta}\boldsymbol{C}\right).$$

  4. The derivative of a matrix inverse is

$$\frac{\partial}{\partial\theta}\boldsymbol{B}^{-1} = -\boldsymbol{B}^{-1}\left(\frac{\partial}{\partial\theta}\boldsymbol{B}\right)\boldsymbol{B}^{-1}$$

  Proof. Solving for $\frac{\partial}{\partial\theta}\boldsymbol{B}^{-1}$ in

$$\begin{aligned} \mathbf{0} &= \frac{\partial}{\partial\theta}\boldsymbol{I} \\ &= \frac{\partial}{\partial\theta}(\boldsymbol{B}^{-1}\boldsymbol{B}) \\ &= \left(\frac{\partial}{\partial\theta}\boldsymbol{B}^{-1}\right)\boldsymbol{B} + \boldsymbol{B}^{-1}\left(\frac{\partial}{\partial\theta}\boldsymbol{B}\right). \end{aligned}$$

□

5. If $\boldsymbol{B}$ is a square nonsingular matrix, then

$$\begin{aligned}\frac{\partial}{\partial\theta}\det\boldsymbol{B} &= \det\boldsymbol{B}\cdot\mathrm{tr}\left(\boldsymbol{B}^{-1}\frac{\partial}{\partial\theta}\boldsymbol{B}\right),\\ \frac{\partial}{\partial\theta}\ln\det\boldsymbol{B} &= \mathrm{tr}\left(\boldsymbol{B}^{-1}\frac{\partial}{\partial\theta}\boldsymbol{B}\right).\end{aligned}$$

- The score, observed information matrix, and expected (Fisher) information matrix of the variance components model are listed below (HW8).

1. Score (gradient) vector is

$$\begin{aligned}\frac{\partial}{\partial b_i}L &= \boldsymbol{e}_i^T\boldsymbol{X}^T\boldsymbol{V}^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{b})\\ \frac{\partial}{\partial\sigma_i^2}L &= -\frac{1}{2}\mathrm{tr}(\boldsymbol{V}^{-1}\boldsymbol{V}_i)+\frac{1}{2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{b})^T\boldsymbol{V}^{-1}\boldsymbol{V}_i\boldsymbol{V}^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{b}).\end{aligned}$$

2. The observed information matrix has entries

$$\begin{aligned}-\frac{\partial^2}{\partial b_i\partial b_j}L &= \boldsymbol{e}_i^T\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}\boldsymbol{e}_j\\ -\frac{\partial^2}{\partial\sigma_i^2\partial b_j}L &= \boldsymbol{e}_j^T\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{V}_i\boldsymbol{V}^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{b})\\ -\frac{\partial^2}{\partial\sigma_i^2\partial\sigma_j^2}L &= -\frac{1}{2}\mathrm{tr}(\boldsymbol{V}^{-1}\boldsymbol{V}_i\boldsymbol{V}^{-1}\boldsymbol{V}_j)+(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{b})^T\boldsymbol{V}^{-1}\boldsymbol{V}_i\boldsymbol{V}^{-1}\boldsymbol{V}_j\boldsymbol{V}^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{b}).\end{aligned}$$

3. The expected (Fisher) information matrix has entries

$$\begin{aligned}\mathrm{E}\left(-\frac{\partial^2}{\partial b_i\partial b_j}L\right) &= \boldsymbol{e}_i^T\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}\boldsymbol{e}_j\\ \mathrm{E}\left(-\frac{\partial^2}{\partial\sigma_i^2\partial b_j}L\right) &= 0\\ \mathrm{E}\left(-\frac{\partial^2}{\partial\sigma_i^2\partial\sigma_j^2}L\right) &= \frac{1}{2}\mathrm{tr}(\boldsymbol{V}^{-1}\boldsymbol{V}_i\boldsymbol{V}^{-1}\boldsymbol{V}_j)\end{aligned}$$

and thus has a block diagonal form

$$\begin{pmatrix}\mathrm{E}(-d_{\boldsymbol{b}}^2L) & \boldsymbol{0}_p\\ \boldsymbol{0}_p^T & \mathrm{E}(-d_{\boldsymbol{\sigma}^2}^2L)\end{pmatrix}.$$

- Setting the score (gradient) vector to zero gives the *likelihood equation* for $\boldsymbol{b}$ and σ_i^2

$$\begin{aligned} \boldsymbol{X}^T\boldsymbol{V}^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{b}) &= \boldsymbol{0} \\ (\boldsymbol{y}-\boldsymbol{X}\boldsymbol{b})^T\boldsymbol{V}^{-1}\boldsymbol{V}_i\boldsymbol{V}^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{b}) &= \mathrm{tr}(\boldsymbol{V}^{-1}\boldsymbol{V}_i), \quad i=0,\ldots,m. \end{aligned}$$

  Root of the likelihood equation is the MLE.

- Example: One-way ANOVA with random effects

$$y_{ij} = \mu + \alpha_i + e_{ij}, \quad i = 1,\ldots,a, j = 1,\ldots,n_i,$$

  where $\alpha_i \sim N(0,\sigma_1^2)$ and $e_{ij} \sim N(0,\sigma_0^2)$ are independent. Here

$$\boldsymbol{X} = \boldsymbol{1}_n, \quad \boldsymbol{V}_0 = \boldsymbol{I}_n, \quad \boldsymbol{V}_1 = \mathrm{BlkDiag}(\sigma_1^2 \boldsymbol{1}_{n_i}\boldsymbol{1}_{n_i}^T),$$

  and

$$\begin{aligned} \boldsymbol{V} &= \mathrm{BlkDiag}\left(\sigma_0^2\boldsymbol{I}_{n_i} + \sigma_1^2\boldsymbol{1}_{n_i}\boldsymbol{1}_{n_i}^T\right), \\ \boldsymbol{V}^{-1} &= \mathrm{BlkDiag}\left(\sigma_0^{-2}\boldsymbol{I}_{n_i} - \sigma_0^{-2}\frac{\sigma_1^2}{\sigma_0^2+n_i\sigma_1^2}\boldsymbol{1}_{n_i}\boldsymbol{1}_{n_i}^T\right) \\ \boldsymbol{V}^{-2} &= \mathrm{BlkDiag}\left(\sigma_0^{-4}\boldsymbol{I}_{n_i} + \sigma_0^{-4}\frac{\sigma_1^2(n_i-2)}{\sigma_0^2+n_i\sigma_1^2}\boldsymbol{1}_{n_i}\boldsymbol{1}_{n_i}^T\right). \end{aligned}$$

  Setting the score vector to 0 gives the likelihood equation (check it! it might be wrong)

$$\begin{aligned} \sigma_0^{-2}\sum_i\left(n_i - \frac{\sigma_1^2 n_i^2}{\sigma_0^2+n_i\sigma_1^2}\right)\mu &= \sigma_0^{-2}\sum_i\left(1-\frac{\sigma_1^2 n_i}{\sigma_0^2+n_i\sigma_1^2}\right)\sum_j y_{ij} \\ \sigma_0^{-2}\sum_i\left(n_i - \frac{\sigma_1^2 n_i}{\sigma_0^2+n_i\sigma_1^2}\right) &= \sigma_0^{-4}\sum_i\left(\sum_j(y_{ij}-\mu)^2 + \frac{\sigma_1^2(n_i-2)}{\sigma_0^2+n_i\sigma_1^2}(\sum_j y_{ij}-n_i\mu)^2\right) \\ \sigma_0^{-2}\sigma_1^2\sum_i\left(n_i - \frac{\sigma_1^2 n_i^2}{\sigma_0^2+n_i\sigma_1^2}\right) &= \sigma_0^{-4}\sigma_1^2\sum_i\frac{\sigma_0^2-n_i\sigma_1^2}{\sigma_0^2+n_i\sigma_1^2}\left(\sum_j y_{ij}-n_i\mu\right)^2. \end{aligned}$$

  In the un-balanced case, there is no analytical solution to the likelihood equation. In the balanced case, analytical solution is available (HW8).

MLE under singular $\boldsymbol{V}$

- Suppose $\boldsymbol{Y} \sim N_n(\boldsymbol{\mu}, \boldsymbol{V})$, where $\boldsymbol{V}$ is singular. MLE depends on a density, which does not exist for singular $\boldsymbol{V}$ ☹

- Assume $r = \text{rank}(\boldsymbol{V}) < n$, let $\boldsymbol{B} \in \mathbb{R}^{n \times r}$ such that $\mathcal{C}(\boldsymbol{B}) = \mathcal{C}(\boldsymbol{V})$. Then

$$\boldsymbol{B}^T\boldsymbol{Y} \sim N_r(\boldsymbol{B}^T\boldsymbol{\mu}, \boldsymbol{B}^T\boldsymbol{V}\boldsymbol{B}),$$

  where the covariance matrix is nonsingular since

$$\begin{aligned} & \text{rank}(\boldsymbol{B}^T\boldsymbol{V}\boldsymbol{B}) = \text{rank}(\boldsymbol{B}^T\boldsymbol{U}_r\boldsymbol{D}_r\boldsymbol{U}_r^T\boldsymbol{B}) \\ = \;& \text{rank}(\boldsymbol{U}_r^T\boldsymbol{B}) = \text{rank}(\boldsymbol{U}_r^T\boldsymbol{U}_r\boldsymbol{T}) = \text{rank}(\boldsymbol{U}_r^T\boldsymbol{U}_r) = r. \end{aligned}$$

- The log-likelihood of $\boldsymbol{B}^T\boldsymbol{Y}$ is

$$\begin{aligned} & -\frac{r}{2}\ln(2\pi) - \frac{1}{2}\ln\det(\boldsymbol{B}^T\boldsymbol{V}\boldsymbol{B}) - \frac{1}{2}(\boldsymbol{B}^T\boldsymbol{y} - \boldsymbol{B}^T\boldsymbol{\mu})^T(\boldsymbol{B}^T\boldsymbol{V}\boldsymbol{B})^{-1}(\boldsymbol{B}^T\boldsymbol{y} - \boldsymbol{B}^T\boldsymbol{\mu}) \\ = \;& -\frac{r}{2}\ln(2\pi) - \frac{1}{2}\ln\det(\boldsymbol{B}^T\boldsymbol{V}\boldsymbol{B}) - \frac{1}{2}(\boldsymbol{y} - \boldsymbol{\mu})^T\boldsymbol{B}(\boldsymbol{B}^T\boldsymbol{V}\boldsymbol{B})^{-1}\boldsymbol{B}^T(\boldsymbol{y} - \boldsymbol{\mu}). \end{aligned}$$

- We now show that maximization of the log-density does *not* depend on choice of $\boldsymbol{B}$.

  Let $\boldsymbol{B}_1, \boldsymbol{B}_2 \in \mathbb{R}^{n \times r}$ be such that $\mathcal{C}(\boldsymbol{B}_1) = \mathcal{C}(\boldsymbol{B}_2) = \mathcal{C}(\boldsymbol{V})$ with $r = \text{rank}(\boldsymbol{V})$. Then

  1. $\det(\boldsymbol{B}_1^T\boldsymbol{V}\boldsymbol{B}_1) = c\,\det(\boldsymbol{B}_2^T\boldsymbol{V}\boldsymbol{B}_2)$ for some constant c.
  2. $\boldsymbol{B}_1(\boldsymbol{B}_1^T\boldsymbol{V}\boldsymbol{B}_1)^{-1}\boldsymbol{B}_1^T = \boldsymbol{B}_2(\boldsymbol{B}_2^T\boldsymbol{V}\boldsymbol{B}_2)^{-1}\boldsymbol{B}_2^T$.

  Proof. Since $\mathcal{C}(\boldsymbol{B}_1) = \mathcal{C}(\boldsymbol{B}_2)$, $\boldsymbol{B}_1 = \boldsymbol{B}_2\boldsymbol{T}$ for some nonsingular transformation matrix $\boldsymbol{T}$. Then

$$\det(\boldsymbol{B}_1^T\boldsymbol{V}\boldsymbol{B}_1) \;=\; \det(\boldsymbol{T}^T\boldsymbol{B}_2^T\boldsymbol{V}\boldsymbol{B}_2\boldsymbol{T}) = [\det(\boldsymbol{T})]^2\det(\boldsymbol{B}_2^T\boldsymbol{V}\boldsymbol{B}_2)$$

  and

$$\boldsymbol{B}_1(\boldsymbol{B}_1^T\boldsymbol{V}\boldsymbol{B}_1)^{-1}\boldsymbol{B}_1^T \;=\; \boldsymbol{B}_2\boldsymbol{T}(\boldsymbol{T}^T\boldsymbol{B}_2^T\boldsymbol{V}\boldsymbol{B}_2\boldsymbol{T})^{-1}\boldsymbol{T}^T\boldsymbol{B}_1^T = \boldsymbol{B}_2(\boldsymbol{B}_2^T\boldsymbol{V}\boldsymbol{B}_2)^{-1}\boldsymbol{B}_2^T.$$

  □

22 Lecture 22: Nov 25

Announcement

- HW8 due Mon Dec 2.
- Email your comments (compliments?) on TA to me.

Last time

- Variance component estimation: MLE.

Today

- Variance component estimation: REML.
- Variance component estimation: method of moment.
- Variance component testing: exact F tests.

Variance component estimation: REML (JM 8.4.2)

- Again we consider estimating the variance components $(\sigma_0^2, \ldots, \sigma_m^2)$ of the model

$$\boldsymbol{y} \sim N_n(\boldsymbol{X}\boldsymbol{b}, \boldsymbol{V}),$$

where

$$\boldsymbol{V} = \sum_{i=0}^{m} \sigma_i^2 \boldsymbol{V}_i,$$

with $\boldsymbol{V}_i$ psd and $\boldsymbol{V}$ nonsingular. Parameters are $\boldsymbol{b}$ and variance components $(\sigma_0^2, \ldots, \sigma_m^2)$.

- *Restricted (or residual) maximum likelihood* (REML) estimation involves finding the MLE of variance components from the distribution of residuals. This allows for estimation of the variance components without complication of the fixed effects.

- It is not clear how to define residuals since $\boldsymbol{V}$ is unknown.

  Let $\boldsymbol{M}$ be any projection (not necessarily orthogonal) onto $\mathcal{C}(\boldsymbol{X})$. Then residuals can be defined as

  $$(\boldsymbol{I}-\boldsymbol{M})\boldsymbol{y}.$$

  The distribution of the residuals is a singular normal

  $$(\boldsymbol{I}-\boldsymbol{M})\boldsymbol{Y} \sim N(\boldsymbol{0}, (\boldsymbol{I}-\boldsymbol{M})\boldsymbol{V}(\boldsymbol{I}-\boldsymbol{M})^T)$$

  and, due to nonsingularity of $\boldsymbol{V}$,

  $$\mathcal{C}((\boldsymbol{I}-\boldsymbol{M})\boldsymbol{V}(\boldsymbol{I}-\boldsymbol{M})^T) = \mathcal{C}(\boldsymbol{I}-\boldsymbol{M})$$

  Suppose $s = \text{rank}(\boldsymbol{X}) = \text{rank}(\boldsymbol{M})$, then

  $$\text{rank}(\boldsymbol{I}-\boldsymbol{M}) = n - s.$$

- We use the result in last section to do MLE on this singular normal distribution. Let $\boldsymbol{B} \in \mathbb{R}^{n\times(n-s)}$ such that $\mathcal{C}(\boldsymbol{B}) = \mathcal{C}(\boldsymbol{I}-\boldsymbol{M})$. Then

  $$\boldsymbol{B}^T(\boldsymbol{I}-\boldsymbol{M})\boldsymbol{Y} \sim N_{n-s}(\boldsymbol{0}, \boldsymbol{B}^T(\boldsymbol{I}-\boldsymbol{M})\boldsymbol{V}(\boldsymbol{I}-\boldsymbol{M})^T\boldsymbol{B})$$

  and the MLE of variance components maximizes the log-likelihood

  $$-\frac{n-s}{2}\ln(2\pi) - \frac{1}{2}\ln\det(\boldsymbol{B}^T(\boldsymbol{I}-\boldsymbol{M})\boldsymbol{V}(\boldsymbol{I}-\boldsymbol{M})^T\boldsymbol{B})$$
  $$-\frac{1}{2}\boldsymbol{y}^T(\boldsymbol{I}-\boldsymbol{M})^T\boldsymbol{B}[\boldsymbol{B}^T(\boldsymbol{I}-\boldsymbol{M})\boldsymbol{V}(\boldsymbol{I}-\boldsymbol{M})^T\boldsymbol{B}]^{-1}\boldsymbol{B}^T(\boldsymbol{I}-\boldsymbol{M})\boldsymbol{y}.$$

  In last section we saw that maximization of this log-likelihood does not depend on choice of $\boldsymbol{B}$. Does this depend on choice of projection matrix $\boldsymbol{M}$?

- $\mathcal{C}((\boldsymbol{I}-\boldsymbol{M})^T\boldsymbol{B}) = \mathcal{N}(\boldsymbol{X}^T)$.

  Proof. Clearly $\boldsymbol{B}^T(\boldsymbol{I}-\boldsymbol{M})\boldsymbol{X} = \boldsymbol{B}^T(\boldsymbol{X}-\boldsymbol{X}) = \boldsymbol{0}_{(n-s)\times p}$. So $\mathcal{C}((\boldsymbol{I}-\boldsymbol{M})^T\boldsymbol{B}) \subset \mathcal{N}(\boldsymbol{X}^T)$. It is enough to show $\text{rank}((\boldsymbol{I}-\boldsymbol{M})^T\boldsymbol{B}) = n-s = \dim(\mathcal{N}(\boldsymbol{X}^T))$. Since $\mathcal{C}(\boldsymbol{B}) = \mathcal{C}(\boldsymbol{I}-\boldsymbol{M})$ and $\boldsymbol{I}-\boldsymbol{M}$ is idemponent, $(\boldsymbol{I}-\boldsymbol{M})\boldsymbol{B} = \boldsymbol{B}$. Thus

  $$\text{rank}((\boldsymbol{I}-\boldsymbol{M})^T\boldsymbol{B}) = \text{rank}(\boldsymbol{B}^T(\boldsymbol{I}-\boldsymbol{M})) = \text{rank}(\boldsymbol{B}^T) = n-s.$$

  □

Remark: This result shows that for any projection $\boldsymbol{M}$ and $\boldsymbol{B}$,

$$(\boldsymbol{I}-\boldsymbol{M})^T\boldsymbol{B}=(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{T}$$

for some transformation matrix $\boldsymbol{T}\in\mathbb{R}^{n\times(n-s)}$ with full column rank and $\mathcal{C}(\boldsymbol{T})=\mathcal{N}(\boldsymbol{X}^T)$ (otherwise contradicting with $\text{rank}((\boldsymbol{I}-\boldsymbol{M})^T\boldsymbol{B})=n-s$). Therefore the MLE estimation of variance components does not depend on choice of either $\boldsymbol{M}$ or $\boldsymbol{B}$.

- It turns out we even do *not* have to make this choice! The entire REML procedure can be carried out using the original data.

- An alternative definition of REML is to do MLE from

$$\boldsymbol{B}^T\boldsymbol{Y}\sim N_{n-s}(\boldsymbol{0},\boldsymbol{B}^T\boldsymbol{V}\boldsymbol{B}),$$

where $\boldsymbol{B}\in\mathbb{R}^{n\times(n-s)}$ is any basis of $\mathcal{N}(\boldsymbol{X}^T)$. That is $\boldsymbol{B}^T\boldsymbol{X}=\boldsymbol{0}_{(n-s)\times p}$ and $\text{rank}(\boldsymbol{B})=n-s$.

Essentially we find a basis $\boldsymbol{B}$ of $\mathcal{N}(\boldsymbol{X}^T)$ and estimate the variance components from $\boldsymbol{B}^T\boldsymbol{Y}$. This is equivalent to the above procedure of choosing a projection $\boldsymbol{M}$ and then doing MLE from residuals (why?). Hence the name *restricted maximum likelihood* (REML).

- Thus we are doing MLE of variance components from the log-likelihood

$$-\frac{n-s}{2}\ln(2\pi)-\frac{1}{2}\ln\det(\boldsymbol{B}^T\boldsymbol{V}\boldsymbol{B})-\frac{1}{2}\boldsymbol{y}^T\boldsymbol{B}[\boldsymbol{B}^T\boldsymbol{V}\boldsymbol{B}]^{-1}\boldsymbol{B}^T\boldsymbol{y}.$$

Setting gradient to zero shows REML of the variance components has to satisfy the likelihood equation

$$\begin{aligned}&\boldsymbol{y}^T\boldsymbol{B}(\boldsymbol{B}^T\boldsymbol{V}\boldsymbol{B})^{-1}\boldsymbol{B}^T\boldsymbol{V}_i\boldsymbol{B}(\boldsymbol{B}^T\boldsymbol{V}\boldsymbol{B})^{-1}\boldsymbol{B}^T\boldsymbol{y}\\ =\;&\text{tr}((\boldsymbol{B}^T\boldsymbol{V}\boldsymbol{B})^{-1}\boldsymbol{B}^T\boldsymbol{V}_i\boldsymbol{B})\\ =\;&\text{tr}(\boldsymbol{B}(\boldsymbol{B}^T\boldsymbol{V}\boldsymbol{B})^{-1}\boldsymbol{B}^T\boldsymbol{V}_i),\quad i=0,\ldots,m.\end{aligned}$$

It turns out the likelihood equation is independent of the choice of basis $\boldsymbol{B}$.

- $\boldsymbol{B}(\boldsymbol{B}^T\boldsymbol{V}\boldsymbol{B})^{-1}\boldsymbol{B}^T=\boldsymbol{V}^{-1}-\boldsymbol{V}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-}\boldsymbol{X}^T\boldsymbol{V}^{-1}$.

Proof. Since $\boldsymbol{V}$ is pd, the symmetric square root $\boldsymbol{V}^{1/2}$ is pd too. It suffices to show

$$\boldsymbol{V}^{1/2}\boldsymbol{B}(\boldsymbol{B}^T\boldsymbol{V}\boldsymbol{B})^{-1}\boldsymbol{B}^T\boldsymbol{V}^{1/2} = \boldsymbol{I} - \boldsymbol{V}^{-1/2}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-}\boldsymbol{X}^T\boldsymbol{V}^{-1/2}.$$

The left side is the orthogonal projection onto $\mathcal{C}(\boldsymbol{V}^{1/2}\boldsymbol{B}) = \mathcal{C}(\boldsymbol{B}) = \mathcal{N}(\boldsymbol{X}^T)$. The right side is the orthogonal projection onto $\mathcal{N}(\boldsymbol{X}^T\boldsymbol{V}^{-1/2}) = \mathcal{N}(\boldsymbol{X}^T)$. Since orthogonal projection to a vector space is unique, proof is done. □

- In summary, let

$$\boldsymbol{A} := \boldsymbol{V}^{-1} - \boldsymbol{V}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-}\boldsymbol{X}^T\boldsymbol{V}^{-1}.$$

Then any solution $(\hat{\sigma}_0^2, \ldots, \hat{\sigma}_m^2)$ to the (restricted) likelihood equation

$$\boldsymbol{y}^T\boldsymbol{A}\boldsymbol{V}_i\boldsymbol{A}\boldsymbol{y} = \mathrm{tr}(\boldsymbol{A}\boldsymbol{V}_i), \quad i = 0, 1, \ldots, m.$$

is the REML.

- Example: REML for balanced one-way ANOVA with random effects (HW8).

Variance component estimation: method of moment (JM 8.4.3)

Also called the ANOVA method (JM 8.4.3).

- Again we consider estimating the variance components $(\sigma_0^2, \ldots, \sigma_m^2)$ of the model

$$\boldsymbol{y} \sim N_n(\boldsymbol{X}\boldsymbol{b}, \boldsymbol{V}),$$

where

$$\boldsymbol{V} = \sum_{i=0}^{m} \sigma_i^2 \boldsymbol{V}_i,$$

with $\boldsymbol{V}_i$ psd and $\boldsymbol{V}$ nonsingular.

- Recall

$$\begin{aligned} \mathrm{E}(\boldsymbol{y}^T\boldsymbol{A}\boldsymbol{y}) &= \mathrm{tr}(\boldsymbol{A}\boldsymbol{V}) + \boldsymbol{b}^T\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X}\boldsymbol{b} \\ &= \sum_{i=0}^{m} \sigma_i^2 \mathrm{tr}(\boldsymbol{A}\boldsymbol{V}_i) + \boldsymbol{b}^T\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X}\boldsymbol{b}. \end{aligned}$$

- Choose $\boldsymbol{A}_i$, $i = 0, \ldots, m$, such that $\boldsymbol{X}^T \boldsymbol{A}_i \boldsymbol{X} = \boldsymbol{0}$. This leads to a system of equations by setting $\mathrm{E}(\boldsymbol{y}^T \boldsymbol{A}_i \boldsymbol{y})$ to the observed $\boldsymbol{y}^T \boldsymbol{A}_i \boldsymbol{y}$

$$\begin{pmatrix} \mathrm{tr}(\boldsymbol{A}_0 \boldsymbol{V}_0) & \mathrm{tr}(\boldsymbol{A}_0 \boldsymbol{V}_1) & \cdots & \mathrm{tr}(\boldsymbol{A}_0 \boldsymbol{V}_m) \\ \vdots & & & \vdots \\ \mathrm{tr}(\boldsymbol{A}_m \boldsymbol{V}_0) & \mathrm{tr}(\boldsymbol{A}_m \boldsymbol{V}_1) & \cdots & \mathrm{tr}(\boldsymbol{A}_m \boldsymbol{V}_m) \end{pmatrix} \begin{pmatrix} \sigma_0^2 \\ \vdots \\ \sigma_m^2 \end{pmatrix} = \begin{pmatrix} \boldsymbol{y}^T \boldsymbol{A}_0 \boldsymbol{y} \\ \vdots \\ \boldsymbol{y}^T \boldsymbol{A}_m \boldsymbol{y} \end{pmatrix},$$

  which can be solved for the variance components.

- Example: MoM for *unbalanced* one-way ANOVA with random effects (HW8). Choose $\boldsymbol{A}_0 = \boldsymbol{P}_{\boldsymbol{Z}} - \boldsymbol{P}_{\boldsymbol{1}}$ and $\boldsymbol{A}_1 = \boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{Z}}$.

Variance components testing: Wald's exact F test

Proposed by Seely and El Bassiouni (1983).

- Consider the mixed linear model:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{Z}_1 \boldsymbol{u}_1 + \boldsymbol{Z}_2 \boldsymbol{u}_2 + \boldsymbol{e},$$

  where $\boldsymbol{Z}_1 \in \mathbb{R}^{n \times q_1}$ and $\boldsymbol{Z}_2 \in \mathbb{R}^{n \times q_2}$. $\boldsymbol{u}_1$, $\boldsymbol{u}_2$, and $\boldsymbol{e}$ are independent random vectors with

$$\boldsymbol{u}_1 \sim N(\boldsymbol{0}_{q_1}, \boldsymbol{R}), \quad \boldsymbol{u}_2 \sim N(\boldsymbol{0}_{q_2}, \sigma_2^2 \boldsymbol{I}_{q_2}), \quad \boldsymbol{e} \sim N(\boldsymbol{0}_n, \sigma_0^2 \boldsymbol{I}_n).$$

  Equivalently, we have the variance component model

$$\boldsymbol{Y} \sim N_n(\boldsymbol{X}\boldsymbol{b}, \boldsymbol{V}),$$

  where

$$\boldsymbol{V} = \sigma_0^2 \boldsymbol{I}_n + \boldsymbol{Z}_1 \boldsymbol{R} \boldsymbol{Z}_1^T + \sigma_2^2 \boldsymbol{Z}_2 \boldsymbol{Z}_2^T.$$

- We are interested testing $H_0 : \sigma_2^2 = 0$ vs $H_A : \sigma_2^2 > 0$.

- Idea of the Wald test is to treat $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ as fixed effects and use the change in SSE as the test statistic.

- Let $\boldsymbol{P}_1 = \boldsymbol{P}_{\boldsymbol{X},\boldsymbol{Z}_1}$ be the orthogonal projection onto $\mathcal{C}(\boldsymbol{X}, \boldsymbol{Z}_1)$.
  Let $\boldsymbol{P}_2 = \boldsymbol{P}_{\boldsymbol{X},\boldsymbol{Z}_1,\boldsymbol{Z}_2}$ be the orthogonal projection onto $\mathcal{C}(\boldsymbol{X}, \boldsymbol{Z}_1, \boldsymbol{Z}_2)$.

  Then

$$\begin{aligned}\text{SSE}_1 &= \boldsymbol{y}^T(\boldsymbol{I} - \boldsymbol{P}_1)\boldsymbol{y} \\ \text{SSE}_2 &= \boldsymbol{y}^T(\boldsymbol{I} - \boldsymbol{P}_2)\boldsymbol{y}\end{aligned}$$

  and

$$\text{SSE}_1 - \text{SSE}_2 = \boldsymbol{y}^T(\boldsymbol{P}_2 - \boldsymbol{P}_1)\boldsymbol{y}.$$

  Ideally we would like to conduct a F test based on the ratio

$$\frac{\text{SSE}_1 - \text{SSE}_2}{\text{SSE}_2}.$$

  By previous result (p90 of notes), it is enough to show the following facts.

  1. $\sigma_0^{-2}(\boldsymbol{I} - \boldsymbol{P}_2)\boldsymbol{V}$ is idempotent with $\text{rank}((\boldsymbol{I} - \boldsymbol{P}_2)\boldsymbol{V}) = \text{rank}(\boldsymbol{I} - \boldsymbol{P}_2)$.

     Proof. Just note that

$$\begin{aligned}& (\boldsymbol{I} - \boldsymbol{P}_2)\boldsymbol{V}(\boldsymbol{I} - \boldsymbol{P}_2) \\ = \;& \boldsymbol{V} - \boldsymbol{V}\boldsymbol{P}_2 - \boldsymbol{P}_2\boldsymbol{V} + \boldsymbol{P}_2\boldsymbol{V}\boldsymbol{P}_2 \\ = \;& \boldsymbol{V} - (\sigma_0^2\boldsymbol{P}_2 + \boldsymbol{Z}_1\boldsymbol{R}\boldsymbol{Z}_1^T + \sigma_2^2\boldsymbol{Z}_2\boldsymbol{Z}_2^T) \\ = \;& \sigma_0^2(\boldsymbol{I} - \boldsymbol{P}_2).\end{aligned}$$

     □

  2. $\sigma_0^{-2}(\boldsymbol{P}_2 - \boldsymbol{P}_1)\boldsymbol{V}$ is idempotent under the null hypothesis $\sigma_2^2 = 0$ with rank $\text{rank}((\boldsymbol{P}_2 - \boldsymbol{P}_1)\boldsymbol{V}) = \text{rank}(\boldsymbol{P}_2 - \boldsymbol{P}_1)$.

     Proof. Just note that, when $\sigma_2^2 = 0$,

$$\begin{aligned}& (\boldsymbol{P}_2 - \boldsymbol{P}_1)\boldsymbol{V}(\boldsymbol{P}_2 - \boldsymbol{P}_1) \\ = \;& \boldsymbol{P}_2\boldsymbol{V}\boldsymbol{P}_2 - \boldsymbol{P}_2\boldsymbol{V}\boldsymbol{P}_1 - \boldsymbol{P}_1\boldsymbol{V}\boldsymbol{P}_2 + \boldsymbol{P}_1\boldsymbol{V}\boldsymbol{P}_1 \\ = \;& (\sigma_0^2\boldsymbol{P}_2 + \boldsymbol{Z}_1\boldsymbol{R}\boldsymbol{Z}_1^T) - (\sigma_0^2\boldsymbol{P}_1 + \boldsymbol{Z}_1\boldsymbol{R}\boldsymbol{Z}_1^T) \\ & -(\sigma_0^2\boldsymbol{P}_1 + \boldsymbol{Z}_1\boldsymbol{R}\boldsymbol{Z}_1^T) + (\sigma_0^2\boldsymbol{P}_1 + \boldsymbol{Z}_1\boldsymbol{R}\boldsymbol{Z}_1^T) \\ = \;& \sigma_0^2(\boldsymbol{P}_2 - \boldsymbol{P}_1).\end{aligned}$$

     □

3. $(\boldsymbol{P}_2 - \boldsymbol{P}_1)\boldsymbol{V}(\boldsymbol{I} - \boldsymbol{P}_1) = \boldsymbol{0}$.

Proof. Just note that

$$\begin{aligned} & (\boldsymbol{P}_2 - \boldsymbol{P}_1)\boldsymbol{V}(\boldsymbol{I} - \boldsymbol{P}_1) \\ = \; & \boldsymbol{P}_2\boldsymbol{V} - \boldsymbol{P}_2\boldsymbol{V}\boldsymbol{P}_2 - \boldsymbol{P}_1\boldsymbol{V} + \boldsymbol{P}_1\boldsymbol{V}\boldsymbol{P}_2 \\ = \; & (\sigma_0^2\boldsymbol{P}_2 + \boldsymbol{Z}_1\boldsymbol{R}\boldsymbol{Z}_1^T + \sigma_2^2\boldsymbol{Z}_2\boldsymbol{Z}_2^T) - (\sigma_0^2\boldsymbol{P}_2 + \boldsymbol{Z}_1\boldsymbol{R}\boldsymbol{Z}_1^T + \sigma_2^2\boldsymbol{Z}_2\boldsymbol{Z}_2^T) \\ & -(\sigma_0^2\boldsymbol{P}_1 + \boldsymbol{Z}_1\boldsymbol{R}\boldsymbol{Z}_1^T + \sigma_2^2\boldsymbol{P}_1\boldsymbol{Z}_2\boldsymbol{Z}_2^T) + (\sigma_0^2\boldsymbol{P}_1 + \boldsymbol{Z}_1\boldsymbol{R}\boldsymbol{Z}_1^T + \sigma_2^2\boldsymbol{P}_1\boldsymbol{Z}_2\boldsymbol{Z}_2^T) \\ = \; & \boldsymbol{0}. \end{aligned}$$

□

- (Wald's F test) The test reject $H_0 : \sigma_2^2 = 0$ when

$$\frac{(\text{SSE}_1 - \text{SSE}_2)/\text{rank}(\boldsymbol{P}_2 - \boldsymbol{P}_1)}{\text{SSE}_1/\text{rank}(\boldsymbol{I} - \boldsymbol{P}_2)} > F_{\text{rank}(\boldsymbol{P}_2-\boldsymbol{P}_1),\text{rank}(\boldsymbol{I}-\boldsymbol{P}_2),\alpha}.$$

  Remark: When $\mathcal{C}(\boldsymbol{X}, \boldsymbol{Z}_1, \boldsymbol{Z}_2) = \mathcal{C}(\boldsymbol{X}, \boldsymbol{Z}_1)$, the test cannot be performed since the numerator is $0/0$.

  Remark: Under the alternative hypothesis $\sigma_2^2 > 0$, the numerator is not a χ^2 random variable anymore but can be numerically evaluated by a method of Davies (1980).

- Extension of Wald's F test to the case

$$\boldsymbol{V} = \sigma_0^2\boldsymbol{\Sigma} + \boldsymbol{Z}_1\boldsymbol{R}\boldsymbol{Z}_1^T + \sigma_2^2\boldsymbol{Z}_2\boldsymbol{Z}_2^T, \tag{4}$$

  where $\boldsymbol{\Sigma} \in \mathbb{R}^{n\times n}$ is a known psd matrix. Let $r = \text{rank}(\boldsymbol{\Sigma})$.

  Given eigen-decomposition $\boldsymbol{\Sigma} = \boldsymbol{U}_r\boldsymbol{D}_r\boldsymbol{U}_r^T$, define $\boldsymbol{T} = \boldsymbol{D}_r^{-1/2}\boldsymbol{U}_r^T$. Thus

$$\boldsymbol{T}\boldsymbol{Y} \sim N(\boldsymbol{T}\boldsymbol{X}\boldsymbol{b}, \sigma_0^2\boldsymbol{I}_r + (\boldsymbol{T}\boldsymbol{Z}_1)\boldsymbol{R}(\boldsymbol{T}\boldsymbol{Z}_1)^T + \sigma_2^2(\boldsymbol{T}\boldsymbol{Z}_2)(\boldsymbol{T}\boldsymbol{Z}_2)^T).$$

  Then Wald's F test can be applied to $\boldsymbol{T}\boldsymbol{Y}$.

Variance components testing: Öfversten's exact F test

Proposed by Öfversten (1993) and Christensen (1996).

- Consider the mixed linear model:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{Z}_1\boldsymbol{u}_1 + \boldsymbol{Z}_2\boldsymbol{u}_2 + \boldsymbol{e},$$

  where $\boldsymbol{Z}_1 \in \mathbb{R}^{n\times q_1}$ and $\boldsymbol{Z}_2 \in \mathbb{R}^{n\times q_2}$. $\boldsymbol{u}_1$, $\boldsymbol{u}_2$, and $\boldsymbol{e}$ are independent random vectors with

$$\boldsymbol{u}_1 \sim N(\boldsymbol{0}_{q_1}, \sigma_1^2\boldsymbol{I}_{q_1}), \quad \boldsymbol{u}_2 \sim N(\boldsymbol{0}_{q_2}, \sigma_2^2\boldsymbol{I}_{q_2}), \quad \boldsymbol{e} \sim N(\boldsymbol{0}_n, \sigma_0^2\boldsymbol{I}_n).$$

  Equivalently, we have the variance component model

$$\boldsymbol{Y} \sim N_n(\boldsymbol{X}\boldsymbol{b}, \boldsymbol{V}),$$

  where

$$\boldsymbol{V} = \sigma_0^2\boldsymbol{I}_n + \sigma_1^2\boldsymbol{Z}_1\boldsymbol{Z}_1^T + \sigma_2^2\boldsymbol{Z}_2\boldsymbol{Z}_2^T.$$

- We are interested in testing $\boldsymbol{H}_0 : \sigma_1^2 = 0$ vs $H_A : \sigma_1^2 > 0$.

- The idea is to massage the problem to the case (4) considered above.
  First perform QR (Gram-Schmidt) on the matrix $\begin{pmatrix}\boldsymbol{X} & \boldsymbol{Z}_2 & \boldsymbol{Z}_1 & \boldsymbol{I}_n\end{pmatrix}$ to obtain an orthonormal basis $\begin{pmatrix}\boldsymbol{C}_1 & \boldsymbol{C}_2 & \boldsymbol{C}_3 & \boldsymbol{C}_4\end{pmatrix}$ of $\mathbb{R}^n$, where
  - $\boldsymbol{C}_1$ is an orthonormal basis of $\mathcal{C}(\boldsymbol{X})$
  - $\boldsymbol{C}_2$ is an orthonormal basis of $\mathcal{C}(\boldsymbol{X}, \boldsymbol{Z}_2) - \mathcal{C}(\boldsymbol{X})$
  - $\boldsymbol{C}_3$ is an orthonormal basis of $\mathcal{C}(\boldsymbol{X}, \boldsymbol{Z}_2, \boldsymbol{Z}_1) - \mathcal{C}(\boldsymbol{X}, \boldsymbol{Z}_2)$
  - $\boldsymbol{C}_4$ is an orthonormal basis of $\mathcal{C}(\boldsymbol{X}, \boldsymbol{Z}_2, \boldsymbol{Z}_1)^\perp$.

  Then we choose λ and matrix $\boldsymbol{K}$ such that

$$\boldsymbol{C}_2^T\boldsymbol{Y} + \boldsymbol{K}\boldsymbol{C}_4^T\boldsymbol{Y} \sim N(\boldsymbol{0}, (\sigma_2^2 + \sigma_0^2/\lambda)\boldsymbol{C}_2^T\boldsymbol{Z}_2\boldsymbol{Z}_2^T\boldsymbol{C}_2 + \sigma_1^2\boldsymbol{C}_2^T\boldsymbol{Z}_1\boldsymbol{Z}_1^T\boldsymbol{C}_2),$$

  which is of form (4).

- If $\boldsymbol{C}_2^T\boldsymbol{Z}_1 = \boldsymbol{0}$, e.g., when $\mathcal{C}(\boldsymbol{Z}_1) \subset \mathcal{C}(\boldsymbol{X})$, then this test cannot be performed.

- If $\boldsymbol{C}_2^T\boldsymbol{Z}_2\boldsymbol{Z}_2^T\boldsymbol{C}_2 = \lambda\boldsymbol{I}$, note

$$\begin{aligned}\boldsymbol{C}_2^T\boldsymbol{Y} &\sim N(\boldsymbol{0}, \sigma_0^2\boldsymbol{I} + \sigma_2^2\boldsymbol{C}_2^T\boldsymbol{Z}_2\boldsymbol{Z}_2^T\boldsymbol{C}_2 + \sigma_1^2\boldsymbol{C}_2^T\boldsymbol{Z}_1\boldsymbol{Z}_1^T\boldsymbol{C}_2)\\ &= N(\boldsymbol{0}, (\sigma_0^2 + \lambda\sigma_2^2)\boldsymbol{I} + \sigma_1^2\boldsymbol{C}_2^T\boldsymbol{Z}_1\boldsymbol{Z}_1^T\boldsymbol{C}_2).\end{aligned}$$

  Then the ordinary Wald's F test can be applied without using the $\boldsymbol{K}\boldsymbol{C}_4^T\boldsymbol{Y}$ piece as long as $\boldsymbol{C}_2^T\boldsymbol{Z}_1 \neq \boldsymbol{0}$.

- In general, $\boldsymbol{C}_2^T\boldsymbol{Z}_2\boldsymbol{Z}_2^T\boldsymbol{C}_2 \neq \lambda\boldsymbol{I}$, then the test requires $\boldsymbol{K}\boldsymbol{C}_4^T\boldsymbol{Y}$. Note that

$$\boldsymbol{C}_2^T\boldsymbol{V}\boldsymbol{C}_4 = \boldsymbol{C}_2^T(\sigma_0^2\boldsymbol{I}_n + \sigma_1^2\boldsymbol{Z}_1\boldsymbol{Z}_1^T + \sigma_2^2\boldsymbol{Z}_2\boldsymbol{Z}_2^T)\boldsymbol{C}_4 = \boldsymbol{0}.$$

  Therefore $\boldsymbol{C}_2^T\boldsymbol{Y}$ is independent of $\boldsymbol{K}\boldsymbol{C}_4^T\boldsymbol{Y}$. We simply pick $\boldsymbol{K}$ such that

$$\boldsymbol{K}\boldsymbol{C}_4^T\boldsymbol{Y} \sim N(\boldsymbol{0}, \sigma_0^2(\lambda^{-1}\boldsymbol{C}_2^T\boldsymbol{Z}_2\boldsymbol{Z}_2^T\boldsymbol{C}_2 - \boldsymbol{I})).$$

  That is

$$\boldsymbol{K}\boldsymbol{K}^T = \lambda^{-1}\boldsymbol{C}_2^T\boldsymbol{Z}_2\boldsymbol{Z}_2^T\boldsymbol{C}_2 - \boldsymbol{I}.$$

  Apparently we need to choose λ such that $\lambda^{-1}\boldsymbol{C}_2^T\boldsymbol{Z}_2\boldsymbol{Z}_2^T\boldsymbol{C}_2 - \boldsymbol{I}$ is psd. Let $\boldsymbol{C}_2^T\boldsymbol{Z}_2\boldsymbol{Z}_2^T\boldsymbol{C}_2 = \boldsymbol{W}\boldsymbol{\Lambda}\boldsymbol{W}^T = \boldsymbol{W}\text{diag}(\lambda_i)\boldsymbol{W}^T$ be the eigendecomposition. Setting λ to be the smallest eigenvalue yields

$$\boldsymbol{K} = \boldsymbol{W}\text{diag}(\sqrt{\lambda_i/\lambda - 1}).$$

23 Lecture 23: Dec 2

Announcement

- HW8 due this Wed (?)
- Course evaluation: `https://classeval.ncsu.edu/`

Last time

- Variance component estimation: REML.
- Variance component estimation: method of moment/ANOVA method.
- Variance component testing: exact F tests.

Today

- Variance component testing: eLRT, eRLRT, score.

Exact LRT and RLRT with one variance component

This is proposed by Crainiceanu and Ruppert (2004). Slight notation change below.

- Consider the variance component model

$$\boldsymbol{Y} \sim N_n(\boldsymbol{X}\boldsymbol{b}, \boldsymbol{V}),$$

where

$$\boldsymbol{V} = \sigma_0^2 \boldsymbol{I}_n + \sigma_1^2 \boldsymbol{V}_1.$$

Let $\lambda = \sigma_1^2/\sigma_0^2$ be the signal-to-noise ratio, and write the covariance as

$$\boldsymbol{V} = \sigma_0^2(\boldsymbol{I}_n + \lambda \boldsymbol{V}_1) = \sigma_0^2 \boldsymbol{V}_\lambda.$$

The model parameters are $(\boldsymbol{b}, \sigma_0^2, \lambda)$. Denote $s = \text{rank}(\boldsymbol{X})$.

- Testing $H_0 : \sigma_1^2 = 0$ vs $H_A : \sigma_1^2 > 0$ is equivalent to testing $H_0 : \lambda = 0$ vs $H_A : \lambda > 0$.

- The log-likelihood function is

$$L(\boldsymbol{b}, \sigma_0^2, \lambda) = -\frac{n}{2}\ln\sigma_0^2 - \frac{1}{2}\ln\det(\boldsymbol{V}_\lambda) - \frac{1}{2\sigma_0^2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{b})^T\boldsymbol{V}_\lambda^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{b}).$$

- The likelihood ratio test (LRT) statistic is

$$\text{LRT} = 2\sup_{H_A} L(\boldsymbol{b}, \sigma_0^2, \lambda) - 2\sup_{H_0} L(\boldsymbol{b}, \sigma_0^2, \lambda).$$

  - Under the null $\lambda = 0$, it is a regular linear model with MLE

$$\begin{aligned}\hat{\boldsymbol{b}} &= (\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{y} \\ \hat{\sigma}_0^2 &= \frac{\text{SSE}}{n} = \frac{\boldsymbol{y}^T(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{y}}{n} =: \frac{\boldsymbol{y}^T\boldsymbol{A}_0\boldsymbol{y}}{n}.\end{aligned}$$

    Thus

$$2\sup_{H_0} L(\boldsymbol{b}, \sigma_0^2, \lambda) = -n\ln\boldsymbol{y}^T\boldsymbol{A}_0\boldsymbol{y} + n\ln n - n.$$

  - Under the alternative, for fixed $\lambda > 0$, the profile likelihood maximizers are

$$\begin{aligned}\hat{\boldsymbol{b}}(\lambda) &= (\boldsymbol{X}^T\boldsymbol{V}_\lambda^{-1}\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{V}_\lambda^{-1}\boldsymbol{y} \\ \hat{\sigma}_0^2(\lambda) &= \frac{\boldsymbol{y}^T[\boldsymbol{V}_\lambda^{-1} - \boldsymbol{V}_\lambda^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{V}_\lambda^{-1}\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{V}_\lambda^{-1}]\boldsymbol{y}}{n} =: \frac{\boldsymbol{y}^T\boldsymbol{A}_\lambda\boldsymbol{y}}{n}.\end{aligned}$$

    Thus

$$2\sup_{H_A} L(\boldsymbol{b}, \sigma_0^2, \lambda) = \sup_{\lambda\geq 0} -n\ln\boldsymbol{y}^T\boldsymbol{A}_\lambda\boldsymbol{y} - \ln\det(\boldsymbol{V}_\lambda) + n\ln n - n.$$

  - Therefore the LRT statistic is

$$\text{LRT} = \sup_{\lambda\geq 0}\ n\ln\boldsymbol{y}^T\boldsymbol{A}_0\boldsymbol{y} - n\ln\boldsymbol{y}^T\boldsymbol{A}_\lambda\boldsymbol{y} - \ln\det(\boldsymbol{V}_\lambda),$$

    where

$$\begin{aligned}\boldsymbol{A}_0 &= \boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}} \\ \boldsymbol{A}_\lambda &= \boldsymbol{V}_\lambda^{-1} - \boldsymbol{V}_\lambda^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{V}_\lambda^{-1}\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{V}_\lambda^{-1}.\end{aligned}$$

- First obtain an orthonormal basis $\begin{pmatrix}\boldsymbol{Q}_0 & \boldsymbol{Q}_1 & \boldsymbol{Q}_2\end{pmatrix}$ of $\mathbb{R}^n$ such that

- $\boldsymbol{Q}_0 \in \mathbb{R}^{n\times s}$ is an orthonormal basis of $\mathcal{C}(\boldsymbol{X})$,
- $\boldsymbol{Q}_1 \in \mathbb{R}^{n\times k}$ is an orthonormal basis from the eigendecomposition

$$\boldsymbol{A}_0\boldsymbol{V}_1\boldsymbol{A}_0 = \boldsymbol{Q}_1 \mathrm{diag}(\mu_1, \ldots, \mu_k)\boldsymbol{Q}_1^T,$$

  where $k = \mathrm{rank}(\boldsymbol{A}_0\boldsymbol{V}_1\boldsymbol{A}_0)$, and
- $\boldsymbol{Q}_2 \in \mathbb{R}^{n\times(n-s-k)}$ is an orthonormal basis of $\mathcal{C}(\boldsymbol{Q}_0, \boldsymbol{Q}_1)^\perp = \mathcal{C}(\boldsymbol{X}, \boldsymbol{Q}_1)^\perp$.

Define $\boldsymbol{Q} = (\boldsymbol{Q}_1, \boldsymbol{Q}_2) \in \mathbb{R}^{n\times(n-s)}$, an orthonormal basis of $\mathcal{C}(\boldsymbol{X})^\perp = \mathcal{N}(\boldsymbol{X}^T)$. Then the following results hold.

1. $\boldsymbol{A}_0 = \boldsymbol{Q}\boldsymbol{Q}^T$.

   Proof. Both sides are the orthogonal projection onto $\mathcal{N}(\boldsymbol{X}^T)$. □

2. $\boldsymbol{Q}^T\boldsymbol{V}_\lambda\boldsymbol{Q} = \mathrm{diag}(1+\lambda\mu_1, \ldots, 1+\lambda\mu_k, 1, \ldots, 1)$.

   Proof. Note $\boldsymbol{A}_0\boldsymbol{Q} = (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{Q} = \boldsymbol{Q}$. Then

$$\begin{aligned}
& \boldsymbol{Q}^T\boldsymbol{V}_\lambda\boldsymbol{Q} \\
= \; & \boldsymbol{Q}^T\boldsymbol{A}_0\boldsymbol{V}_\lambda\boldsymbol{A}_0\boldsymbol{Q} \\
= \; & \boldsymbol{Q}^T\boldsymbol{A}_0(\boldsymbol{I}_n + \lambda\boldsymbol{V}_1)\boldsymbol{A}_0\boldsymbol{Q} \\
= \; & \boldsymbol{Q}^T\boldsymbol{A}_0\boldsymbol{Q} + \lambda\boldsymbol{Q}^T\boldsymbol{Q}_1\mathrm{diag}(\mu_1, \ldots, \mu_k)\boldsymbol{Q}_1^T\boldsymbol{Q} \\
= \; & \boldsymbol{Q}^T\boldsymbol{Q} + \lambda \begin{pmatrix} \boldsymbol{I}_k \\ \boldsymbol{0} \end{pmatrix} \mathrm{diag}(\mu_1, \ldots, \mu_k) \begin{pmatrix} \boldsymbol{I}_k & \boldsymbol{0} \end{pmatrix} \\
= \; & \boldsymbol{I}_{n-s} + \lambda\mathrm{diag}(\mu_1, \ldots, \mu_k, 0, \ldots, 0) \\
= \; & \mathrm{diag}(1+\mu_1, \ldots, 1+\mu_k, 1, \ldots, 1).
\end{aligned}$$

   □

3. $\boldsymbol{A}_\lambda = \boldsymbol{Q}\mathrm{diag}((1+\mu_1)^{-1}, \ldots, (1+\mu_k)^{-1}, 1, \ldots, 1)\boldsymbol{Q}^T$.

   Proof. Since $\boldsymbol{Q}$ form a basis of $\mathcal{N}(\boldsymbol{X}^T)$, by a previous result (p143),

$$\boldsymbol{A}_\lambda = \boldsymbol{V}_\lambda^{-1} - \boldsymbol{V}_\lambda^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{V}_\lambda^{-1}\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{V}_\lambda^{-1} = \boldsymbol{Q}(\boldsymbol{Q}^T\boldsymbol{V}_\lambda\boldsymbol{Q})^{-1}\boldsymbol{Q}^T.$$

   Then substitute the result 2 for $\boldsymbol{Q}^T\boldsymbol{V}_\lambda\boldsymbol{Q}$. □

4. Under the alternative $\lambda > 0$,

$$\sigma_0^{-1}\boldsymbol{Q}^T\boldsymbol{y} \sim N(\boldsymbol{0}_{n-s}, \text{diag}(1+\lambda\mu_1, \ldots, 1+\lambda\mu_k, 1, \ldots, 1)).$$

Under the null $\lambda = 0$,

$$\sigma_0^{-1}\boldsymbol{Q}^T\boldsymbol{y} \sim N(\boldsymbol{0}_{n-s}, \boldsymbol{I}_{n-s}).$$

Proof. Since $\mathcal{C}(\boldsymbol{Q}) = \mathcal{N}(\boldsymbol{X}^T)$, $\boldsymbol{Q}^T\boldsymbol{X}\boldsymbol{b} = \boldsymbol{0}_{n-s}$. Covariance follows from result 2. □

5. Let the eigenvalues of $\boldsymbol{V}_1$ be $(\xi_1, \ldots, \xi_\ell)$, where $\ell = \text{rank}(\boldsymbol{V}_1)$. Then

$$\ln\det(\boldsymbol{V}_\lambda) = \sum_{i=1}^{\ell} \ln(1+\lambda\xi_i).$$

Proof. This is trivial. □

- Putting things together, under the null,

$$\begin{aligned}
\text{LRT} &= \sup_{\lambda\geq 0}\ n\ln \boldsymbol{y}^T\boldsymbol{A}_0\boldsymbol{y} - n\ln \boldsymbol{y}^T\boldsymbol{A}_\lambda\boldsymbol{y} - \ln\det(\boldsymbol{V}_\lambda) \\
&= \sup_{\lambda\geq 0}\ n\ln \boldsymbol{y}^T\boldsymbol{Q}\boldsymbol{Q}^T\boldsymbol{y} \\
&\quad -n\ln \boldsymbol{y}^T\boldsymbol{Q}\text{diag}((1+\lambda\mu_1)^{-1}, \ldots, (1+\lambda\mu_k)^{-1}, 1, \ldots, 1)\boldsymbol{Q}^T\boldsymbol{y} \\
&\quad -\ln\det(\boldsymbol{V}_\lambda) \\
&=^{\mathcal{D}} \sup_{\lambda\geq 0}\ n\ln \frac{\sum_{i=1}^{n-s} w_i^2}{\sum_{i=1}^{k}\frac{w_i^2}{1+\lambda\mu_i} + \sum_{i=k+1}^{n-s} w_i^2} - \sum_{i=1}^{l}\ln(1+\lambda\xi_i),
\end{aligned}$$

where w_i are $(n-s)$ independent standard normals.

- The null distribution can be obtained from computer simulation:

Given $\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{V}_1$, and simulation replicates B
Eigen-decomposition: $\boldsymbol{V}_1 = \boldsymbol{U}_\ell\text{diag}(\xi_1, \ldots, \xi_l)\boldsymbol{U}_\ell^T$
Regress $\boldsymbol{U}_\ell\text{diag}(\sqrt{\xi_1}, \ldots, \sqrt{\xi_\ell})$ on $\boldsymbol{X}$ and obtain residuals $\boldsymbol{A}_0\boldsymbol{V}_1^{1/2}$
Obtain eigenvalues $(\mu_1, \ldots, \mu_k)$ of $\boldsymbol{A}_0\boldsymbol{V}_1\boldsymbol{A}_0$
for $b = 1$ **to** B **do**
  Simulate $(w_1, \ldots, w_{n-s})$ independent standard normals

Find and record the maximal value of

$$f(\lambda) = n \ln \frac{\sum_{i=1}^{n-s} w_i^2}{\sum_{i=1}^{k} \frac{w_i^2}{1+\lambda\mu_i} + \sum_{i=k+1}^{n-s} w_i^2} - \sum_{i=1}^{l} \ln(1+\lambda\xi_i)$$

over $\lambda \in [0, \infty)$

end for

Remark: The maximization task is subject to an MM algorithm.

- 0 is a local maximum of $f(\lambda)$ if $f'(\lambda) <= 0$. Therefore the probability of having a local maximum at 0 is

$$\text{Prob}\left(\frac{\sum_{i=1}^{k} \mu_i w_i^2}{\sum_{i=1}^{n-s} w_i^2} \leq \frac{1}{n} \sum_{i=1}^{\ell} \xi_i \right),$$

which provides a good approximation of the point mass at 0 of the null distribution of LRT.

- Same derivation can be carried out for the restricted (residual) LRT, in which case

$$\text{RLRT} \quad =^{\mathcal{D}} \quad \sup_{\lambda \geq 0} (n-s) \ln \frac{\sum_{i=1}^{n-s} w_i^2}{\sum_{i=1}^{k} \frac{w_i^2}{1+\lambda\mu_i} + \sum_{i=k+1}^{n-s} w_i^2} - \sum_{i=1}^{l} \ln(1+\lambda\xi_i)$$

under the null $\lambda = 0$.

(Rao) score test with one variance component

- Again we consider the variance component model

$$\boldsymbol{Y} \sim N_n(\boldsymbol{X}\boldsymbol{b}, \boldsymbol{V}),$$

where

$$\boldsymbol{V} = \sigma_0^2 \boldsymbol{I}_n + \sigma_1^2 \boldsymbol{V}_1.$$

- We develop an exact score test for $H_0 : \sigma_1^2 = 0$ vs $H_A : \sigma_1^2 > 0$.

Score test avoids the maximization step in simulating the null distribution of test statistic.

- The log-likelihood function is

$$L(\boldsymbol{b}, \sigma_0^2, \lambda) = -\frac{1}{2}\ln\det(\boldsymbol{V}) - \frac{1}{2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{b})^T\boldsymbol{V}^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{b}).$$

  and the first derivative with respect to σ_1^2 is

$$\frac{\partial}{\partial\sigma_1^2}L = -\frac{1}{2}\mathrm{tr}(\boldsymbol{V}^{-1}\boldsymbol{V}_1) + \frac{1}{2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{b})^T\boldsymbol{V}^{-1}\boldsymbol{V}_1\boldsymbol{V}^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{b}).$$

  The information matrix relevant to variance components are

$$\begin{aligned}
\mathrm{E}\left(-\frac{\partial^2}{\partial\sigma_0^2\partial\sigma_0^2}L\right) &= \frac{1}{2}\mathrm{tr}(\boldsymbol{V}^{-2}) \\
\mathrm{E}\left(-\frac{\partial^2}{\partial\sigma_0^2\partial\sigma_1^2}L\right) &= \mathrm{E}\left(-\frac{\partial^2}{\partial\sigma_1^2\partial\sigma_0^2}L\right) = \frac{1}{2}\mathrm{tr}(\boldsymbol{V}^{-2}\boldsymbol{V}_1) \\
\mathrm{E}\left(-\frac{\partial^2}{\partial\sigma_1^2\partial\sigma_1^2}L\right) &= \frac{1}{2}\mathrm{tr}(\boldsymbol{V}^{-1}\boldsymbol{V}_1\boldsymbol{V}^{-1}\boldsymbol{V}_1).
\end{aligned}$$

- The (Rao) score statistic is based on

$$\boldsymbol{I}^{-1}_{\sigma_1^2,\sigma_1^2}\left(\frac{\partial}{\partial\sigma_1^2}L\right)^2$$

  evaluated at the MLE under the null.

- We evaluate the partial derivatives at the MLE under the null

$$\hat{\boldsymbol{b}} = (\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{y}, \quad \hat{\sigma}_0^2 = \frac{\boldsymbol{y}^T(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{y}}{n}.$$

  That is

$$\begin{aligned}
D_1 &:= \frac{\partial}{\partial\sigma_1^2}L(\hat{\boldsymbol{b}}, \hat{\sigma}_0^2) \\
&= -\frac{n\mathrm{tr}(\boldsymbol{V}_1)}{2\boldsymbol{y}^T(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{y}} + \frac{n^2\boldsymbol{y}^T(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{V}_1(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{y}}{2[\boldsymbol{y}^T(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{y}]^2} \\
&= \frac{-n\mathrm{tr}(\boldsymbol{V}_1)[\boldsymbol{y}^T(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{y}] + n^2\boldsymbol{y}^T(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{V}_1(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{y}}{2[\boldsymbol{y}^T(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{y}]^2}
\end{aligned}$$

$$\begin{aligned}
J_{00} &:= \mathrm{E}\left(-\frac{\partial^2}{\partial\sigma_0^2\partial\sigma_0^2}L(\hat{\boldsymbol{b}}, \hat{\sigma}_0^2)\right) = \frac{n^3}{2[\boldsymbol{y}^T(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{y}]^2} \\
J_{01} &= J_{10} := \mathrm{E}\left(-\frac{\partial^2}{\partial\sigma_0^2\partial\sigma_1^2}L(\hat{\boldsymbol{b}}, \hat{\sigma}_0^2)\right) = \frac{n^2\mathrm{tr}(\boldsymbol{V}_1)}{2[\boldsymbol{y}^T(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{y}]^2} \\
J_{11} &:= \mathrm{E}\left(-\frac{\partial^2}{\partial\sigma_1^2\partial\sigma_1^2}L(\hat{\boldsymbol{b}}, \hat{\sigma}_0^2)\right) = \frac{n^2\mathrm{tr}(\boldsymbol{V}_1^2)}{2[\boldsymbol{y}^T(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{y}]^2},
\end{aligned}$$

from which we form the score statistic

$$\begin{aligned} T &= \begin{cases} (J_{11} - J_{10}J_{00}^{-1}J_{01})^{-1}D_1^2 & D_1 \geq 0 \\ 0 & D_1 < 0 \end{cases} \\ &= \begin{cases} \frac{\left[-n\mathrm{tr}(\boldsymbol{V}_1)+n^2\frac{\boldsymbol{y}^T(\boldsymbol{I}-\boldsymbol{P_X})\boldsymbol{V}_1(\boldsymbol{I}-\boldsymbol{P_X})\boldsymbol{y}}{\boldsymbol{y}^T(\boldsymbol{I}-\boldsymbol{P_X})\boldsymbol{y}}\right]^2}{2[n^2\mathrm{tr}(\boldsymbol{V}_1^2)-n\mathrm{tr}(\boldsymbol{V}_1)^2]} & \frac{\boldsymbol{y}^T(\boldsymbol{I}-\boldsymbol{P_X})\boldsymbol{V}_1(\boldsymbol{I}-\boldsymbol{P_X})\boldsymbol{y}}{\boldsymbol{y}^T(\boldsymbol{I}-\boldsymbol{P_X})\boldsymbol{y}} \geq \frac{\mathrm{tr}(\boldsymbol{V}_1)}{n} \\ 0 & \frac{\boldsymbol{y}^T(\boldsymbol{I}-\boldsymbol{P_X})\boldsymbol{V}_1(\boldsymbol{I}-\boldsymbol{P_X})\boldsymbol{y}}{\boldsymbol{y}^T(\boldsymbol{I}-\boldsymbol{P_X})\boldsymbol{y}} < \frac{\mathrm{tr}(\boldsymbol{V}_1)}{n} \end{cases}. \end{aligned}$$

Essentially the score test rejects when

$$T' = \max\left\{\frac{\boldsymbol{y}^T(\boldsymbol{I}-\boldsymbol{P_X})\boldsymbol{V}_1(\boldsymbol{I}-\boldsymbol{P_X})\boldsymbol{y}}{\boldsymbol{y}^T(\boldsymbol{I}-\boldsymbol{P_X})\boldsymbol{y}}, \frac{\mathrm{tr}(\boldsymbol{V}_1)}{n}\right\}$$

is large.

- Let the eigen-decomposition of $(\boldsymbol{I}-\boldsymbol{P_X})\boldsymbol{V}_1(\boldsymbol{I}-\boldsymbol{P_X})$ be

$$(\boldsymbol{I}-\boldsymbol{P_X})\boldsymbol{V}_1(\boldsymbol{I}-\boldsymbol{P_X}) = \boldsymbol{Q}_1\mathrm{diag}(\mu_1,\ldots,\mu_k)\boldsymbol{Q}_1^T,$$

$\boldsymbol{Q}_2$ be an orthonormal basis of $\mathcal{C}(\boldsymbol{X},\boldsymbol{Q}_1)^\perp$, and $\boldsymbol{Q} = (\boldsymbol{Q}_1,\boldsymbol{Q}_2) \in \mathbb{R}^{n\times(n-s)}$. Then

$$\begin{aligned} T' &= \max\left\{\frac{\boldsymbol{y}^T\boldsymbol{Q}\mathrm{diag}(\mu_1,\ldots,\mu_k,0,\ldots,0)\boldsymbol{Q}^T\boldsymbol{y}}{\boldsymbol{y}^T\boldsymbol{Q}\boldsymbol{Q}^T\boldsymbol{y}}, \frac{\mathrm{tr}(\boldsymbol{V}_1)}{n}\right\} \\ &=^{\mathcal{D}} \max\left\{\frac{\sum_{i=1}^k \mu_k w_i^2}{\sum_{i=1}^{n-s} w_i^2}, \frac{\mathrm{tr}(\boldsymbol{V}_1)}{n}\right\}, \end{aligned}$$

where w_i are $n-s$ independent standard normals.

- The null distribution can be obtained from computer simulation:

  Given $\boldsymbol{y},\boldsymbol{X},\boldsymbol{V}_1$, and simulation replicates B
  Eigen-decomposition: $\boldsymbol{V}_1 = \boldsymbol{U}_\ell\mathrm{diag}(\xi_1,\ldots,\xi_l)\boldsymbol{U}_\ell^T$
  Regress $\boldsymbol{U}_\ell\mathrm{diag}(\sqrt{\xi_1},\ldots,\sqrt{\xi_\ell})$ on $\boldsymbol{X}$ to obtain residuals $(\boldsymbol{I}-\boldsymbol{P_X})\boldsymbol{V}_1^{1/2}$
  Obtain eigenvalues $(\mu_1,\ldots,\mu_k)$ of $(\boldsymbol{I}-\boldsymbol{P_X})\boldsymbol{V}_1(\boldsymbol{I}-\boldsymbol{P_X})$
  for $b = 1$ **to** B **do**
    Simulate $(w_1,\ldots,w_{n-s})$ independent standard normals
    Record the value

$$\max\left\{\frac{\sum_{i=1}^k \mu_k w_i^2}{\sum_{i=1}^{n-s} w_i^2}, \frac{\mathrm{tr}(\boldsymbol{V}_1)}{n}\right\}$$

end for

- The probability mass of T' at lower boundary $\text{tr}(\boldsymbol{V}_1)/n$ is

$$\text{Prob}\left(\frac{\sum_{i=1}^{k}\mu_i w_i^2}{\sum_{i=1}^{n-s} w_i^2} \leq \frac{\text{tr}(\boldsymbol{V}_1)}{n}\right),$$

  same as the point mass at 0 of the null distribution of eLRT.

Wald test with one variance component

TODO.

24 Lecture 24: Dec 4

Announcement

- HW8 due today.
- Course evaluation: `https://classeval.ncsu.edu/`
- No OH from next week. Reachable by email.
- Final exam: Fiday, Dec 13 @ 8A-11A, SAS 5270.

Last time

- Variance component testing: eLRT, eRLRT, score test.

Today

- Variance component model: testing with two or more variance components.
- (Last topic!) Prediction: BLP and BLUP.

Testing with two or more variance components

- First we extend the test with one variance component to a slightly more general case $\boldsymbol{y} \sim N(\boldsymbol{X}\boldsymbol{b}, \boldsymbol{V})$ with

$$\boldsymbol{V} = \sigma_0^2 \boldsymbol{V}_0 + \sigma_1^2 \boldsymbol{V}_1, \tag{5}$$

  where $\boldsymbol{V}_0 \in \mathbb{R}^{n \times n}$ is a known psd matrix. Let $r = \text{rank}(\boldsymbol{V}_0)$.

  Given eigen-decomposition $\boldsymbol{V}_0 = \boldsymbol{U}_r \boldsymbol{D}_r \boldsymbol{U}_r^T$, define $\boldsymbol{T} = \boldsymbol{D}_r^{-1/2} \boldsymbol{U}_r^T \in \mathbb{R}^{r \times n}$. Then

$$\boldsymbol{T}\boldsymbol{Y} \sim N(\boldsymbol{T}\boldsymbol{X}\boldsymbol{b}, \sigma_0^2 \boldsymbol{I}_r + \sigma_1^2 \boldsymbol{T}\boldsymbol{V}_1\boldsymbol{T}^T)$$

  and the eLRT, eRLRT or score test can be applied to $\boldsymbol{T}\boldsymbol{Y}$.

- Now we consider the linear model with two variance components

$$\boldsymbol{Y} \sim N_n(\boldsymbol{X}\boldsymbol{b}, \boldsymbol{V}),$$

  where

$$\boldsymbol{V} = \sigma_0^2 \boldsymbol{I}_n + \sigma_1^2 \boldsymbol{V}_1 + \sigma_2^2 \boldsymbol{V}_2.$$

- We are interested in testing $H_0 : \sigma_2^2 = 0$ vs $H_A : \sigma_2^2 > 0$.
- The idea is to massage the problem to the case (5) above.
- First perform QR (Gram-Schmidt) on the matrix $(\boldsymbol{X}, \boldsymbol{V}_1, \boldsymbol{V}_2, \boldsymbol{I}_n)$ to obtain an orthonormal basis $(\boldsymbol{Q}_0, \boldsymbol{Q}_1, \boldsymbol{Q}_2, \boldsymbol{Q}_3)$ of $\mathbb{R}^n$, where
  - $\boldsymbol{Q}_0$ is an orthonormal basis of $\mathcal{C}(\boldsymbol{X})$
  - $\boldsymbol{Q}_1$ is an orthonormal basis of $\mathcal{C}(\boldsymbol{X}, \boldsymbol{V}_1) - \mathcal{C}(\boldsymbol{X})$
  - $\boldsymbol{Q}_2$ is an orthonormal basis of $\mathcal{C}(\boldsymbol{X}, \boldsymbol{V}_1, \boldsymbol{V}_2) - \mathcal{C}(\boldsymbol{X}, \boldsymbol{V}_1)$
  - $\boldsymbol{Q}_3$ is an orthonormal basis of $\mathcal{C}(\boldsymbol{X}, \boldsymbol{V}_1, \boldsymbol{V}_2)^\perp$.
- If $\text{rank}(\boldsymbol{Q}_2) > 0$, that is $\mathcal{C}(\boldsymbol{X}, \boldsymbol{V}_1) \subsetneq \mathcal{C}(\boldsymbol{X}, \boldsymbol{V}_1, \boldsymbol{V}_2)$, then

  $$\boldsymbol{Q}_2^T \boldsymbol{Y} \sim N(\boldsymbol{0}, \sigma_0^2 \boldsymbol{I} + \sigma_2^2 \boldsymbol{Q}_2^T \boldsymbol{V}_2 \boldsymbol{Q}_2)$$

  and the exact F test, eLRT, eRLRT and score test can be applied to $\boldsymbol{Q}_2^T \boldsymbol{y}$.

  Is this beneficial to add the $\boldsymbol{Q}_3^T \boldsymbol{Y}$ component???
- If $\text{rank}(\boldsymbol{Q}_2) = 0$, that is $\mathcal{C}(\boldsymbol{X}, \boldsymbol{V}_1) = \mathcal{C}(\boldsymbol{X}, \boldsymbol{V}_1, \boldsymbol{V}_2)$, then we would like to choose λ and matrix $\boldsymbol{K}$ such that

  $$\boldsymbol{Q}_1^T \boldsymbol{Y} + \boldsymbol{K} \boldsymbol{Q}_3^T \boldsymbol{Y} \sim N(\boldsymbol{0}, (\sigma_1^2 + \sigma_0^2/\lambda) \boldsymbol{Q}_1^T \boldsymbol{V}_1 \boldsymbol{Q}_1 + \sigma_2^2 \boldsymbol{Q}_1^T \boldsymbol{V}_2 \boldsymbol{Q}_1),$$

  which is of form (5). We consider following situations.
  - If $\boldsymbol{Q}_1^T \boldsymbol{V}_2 = \boldsymbol{0}$, e.g., when $\mathcal{C}(\boldsymbol{V}_2) \subset \mathcal{C}(\boldsymbol{X})$, then this test cannot be performed.
  - If $\boldsymbol{Q}_1^T \boldsymbol{V}_1 \boldsymbol{Q}_1 = \lambda \boldsymbol{I}$, note

    $$\begin{aligned} \boldsymbol{Q}_1^T \boldsymbol{Y} &\sim N(\boldsymbol{0}, \sigma_0^2 \boldsymbol{I} + \sigma_1^2 \boldsymbol{Q}_1^T \boldsymbol{V}_1 \boldsymbol{Q}_1 + \sigma_2^2 \boldsymbol{Q}_1^T \boldsymbol{V}_2 \boldsymbol{Q}_1) \\ &= N(\boldsymbol{0}, (\sigma_0^2 + \lambda \sigma_1^2) \boldsymbol{I} + \sigma_2^2 \boldsymbol{Q}_1^T \boldsymbol{V}_2 \boldsymbol{Q}_1). \end{aligned}$$

    Then the ordinary tests (F test, eLRT, eRLRT, score) for one variance component can be applied without using the $\boldsymbol{K} \boldsymbol{Q}_3^T \boldsymbol{y}$ piece as long as $\boldsymbol{Q}_1^T \boldsymbol{V}_2 \neq \boldsymbol{0}$.
  - In general, $\boldsymbol{Q}_1^T \boldsymbol{V}_1 \boldsymbol{Q}_1 \neq \lambda \boldsymbol{I}$, then the test requires the $\boldsymbol{K} \boldsymbol{Q}_3^T \boldsymbol{y}$ term, which has distribution

    $$\boldsymbol{K} \boldsymbol{Q}_3^T \boldsymbol{Y} \sim N(\boldsymbol{0}, \sigma_0^2 \boldsymbol{K} \boldsymbol{K}^T).$$

Note that

$$\boldsymbol{Q}_1^T \boldsymbol{V} \boldsymbol{Q}_3 = \boldsymbol{Q}_1^T (\sigma_0^2 \boldsymbol{I}_n + \sigma_1^2 \boldsymbol{V}_1 + \sigma_2^2 \boldsymbol{V}_2) \boldsymbol{Q}_3 = \boldsymbol{0}.$$

Therefore $\boldsymbol{Q}_1^T \boldsymbol{Y} \perp \boldsymbol{K} \boldsymbol{Q}_3^T \boldsymbol{Y}$. We simply pick $\boldsymbol{K}$ such that

$$\boldsymbol{K} \boldsymbol{Q}_3^T \boldsymbol{Y} \sim N(\boldsymbol{0}, \sigma_0^2 (\lambda^{-1} \boldsymbol{Q}_1^T \boldsymbol{V}_2 \boldsymbol{Q}_1 - \boldsymbol{I})).$$

That is

$$\boldsymbol{K} \boldsymbol{K}^T = \lambda^{-1} \boldsymbol{Q}_1^T \boldsymbol{V}_2 \boldsymbol{Q}_1 - \boldsymbol{I}.$$

Apparently we need to choose λ such that $\lambda^{-1} \boldsymbol{Q}_1^T \boldsymbol{V}_2 \boldsymbol{Q}_1 - \boldsymbol{I}$ is psd. Let

$$\boldsymbol{Q}_1^T \boldsymbol{V}_2 \boldsymbol{Q}_1 = \boldsymbol{W} \boldsymbol{\Lambda} \boldsymbol{W}^T = \boldsymbol{W} \mathrm{diag}(\lambda_i) \boldsymbol{W}^T$$

be the eigendecomposition. Setting λ to be the smallest eigenvalue yields

$$\boldsymbol{K} = \boldsymbol{W} \mathrm{diag}(\sqrt{\lambda_i / \lambda - 1}).$$

Best linear prediction (BLP)

- Given data $(y, x_1, \ldots, x_p)$, regression can be thought of one way to predict y from $x_1, \ldots, x_p$.
- A reasonable criterion is to choose predictor $f(\boldsymbol{x})$ such that the mean squared error

  $$\mathrm{MSE} = \mathrm{E}[y - f(\boldsymbol{x})]^2$$

  is minimized. Here the expectation is wrt the joint distribution of $(y, \boldsymbol{x})$.
- Let $m(\boldsymbol{x}) = \mathrm{E}(y|\boldsymbol{x})$. Then for any other predictor $f(\boldsymbol{x})$,

  $$\mathrm{E}[y - m(\boldsymbol{x})]^2 \leq \mathrm{E}[y - f(\boldsymbol{x})]^2.$$

  That is $m(\boldsymbol{x})$ is the best predictor of y.

  Proof.

  $$\begin{aligned} & \mathrm{E}[y - f(\boldsymbol{x})]^2 \\ = \; & \mathrm{E}[y - m(\boldsymbol{x}) + m(\boldsymbol{x}) - f(\boldsymbol{x})]^2 \\ = \; & \mathrm{E}[y - m(\boldsymbol{x})]^2 + \mathrm{E}[m(\boldsymbol{x}) - f(\boldsymbol{x})]^2 + 2\mathrm{E}[y - m(\boldsymbol{x})][m(\boldsymbol{x}) - f(\boldsymbol{x})]. \end{aligned}$$

But the cross term vanishes

$$\begin{aligned} & \mathrm{E}[y - m(\boldsymbol{x})][m(\boldsymbol{x}) - f(\boldsymbol{x})] \\ = \; & \mathrm{E}\left\{\mathrm{E}[y - m(\boldsymbol{x})][m(\boldsymbol{x}) - f(\boldsymbol{x})] \mid \boldsymbol{x}\right\} \\ = \; & \mathrm{E}\left\{[m(\boldsymbol{x}) - f(\boldsymbol{x})]\mathrm{E}[y - m(\boldsymbol{x}) \mid \boldsymbol{x}]\right\} \\ = \; & \mathrm{E}\left\{[m(\boldsymbol{x}) - f(\boldsymbol{x})]0\right\} \\ = \; & 0. \end{aligned}$$

Therefore $\mathrm{E}[y - f(\boldsymbol{x})]^2 \geq \mathrm{E}[y - m(\boldsymbol{x})]^2$ for any $f(\boldsymbol{x})$. □

- In order to use this result, we need to know the joint distribution of $(y, \boldsymbol{x})$. This is often unrealistic ☹ If only the first two moments (means, variances, and covariances) are known, then we can find the *best linear predictor* (BLP) of y.

- Assume

$$\mathrm{E}\begin{pmatrix} y \\ \boldsymbol{x} \end{pmatrix} = \begin{pmatrix} \mu_y \\ \boldsymbol{\mu}_{\boldsymbol{x}} \end{pmatrix}, \quad \mathrm{Cov}\begin{pmatrix} \sigma_{yy} & \boldsymbol{\Sigma}_{\boldsymbol{x}y}^T \\ \boldsymbol{\Sigma}_{\boldsymbol{x}y} & \boldsymbol{\Sigma}_{\boldsymbol{x}\boldsymbol{x}} \end{pmatrix}.$$

Let $\boldsymbol{\beta}^*$ be a solution of $\boldsymbol{\Sigma}_{\boldsymbol{x}\boldsymbol{x}}\boldsymbol{\beta} = \boldsymbol{\Sigma}_{\boldsymbol{x}y}$. Then

$$\widehat{\mathrm{E}}(y|\boldsymbol{x}) := \mu_y + (\boldsymbol{x} - \boldsymbol{\mu}_{\boldsymbol{x}})^T\boldsymbol{\beta}^*$$

is *the* best linear predictor (BLP) of y.

Proof. Let $f(\boldsymbol{x}) = \alpha + \boldsymbol{x}^T\boldsymbol{\beta}$ be an arbitrary linear predictor. Then we find $\alpha, \boldsymbol{\beta}$ by minimizing the MSE

$$\begin{aligned} & \mathrm{E}[y - f(\boldsymbol{x})]^2 \\ = \; & \mathrm{E}(y - \alpha - \boldsymbol{x}^T\boldsymbol{\beta})^2 \\ = \; & \mathrm{E}(y - \alpha)^2 + \boldsymbol{\beta}^T\mathrm{E}(\boldsymbol{x}\boldsymbol{x}^T)\boldsymbol{\beta} - 2\mathrm{E}[(y - \alpha)(\boldsymbol{x}^T\boldsymbol{\beta})] \\ = \; & \mathrm{E}(y - \alpha)^2 + \boldsymbol{\beta}^T\mathrm{E}(\boldsymbol{x}\boldsymbol{x}^T)\boldsymbol{\beta} - 2\mathrm{E}(y\boldsymbol{x}^T)\boldsymbol{\beta} + 2\alpha\mathrm{E}(\boldsymbol{x}^T)\boldsymbol{\beta} \\ = \; & \mathrm{E}(y - \alpha)^2 + \boldsymbol{\beta}^T\mathrm{E}(\boldsymbol{x}\boldsymbol{x}^T)\boldsymbol{\beta} - 2\mathrm{E}(y\boldsymbol{x})^T\boldsymbol{\beta} + 2\alpha\boldsymbol{\mu}_{\boldsymbol{x}}^T\boldsymbol{\beta}. \end{aligned}$$

Setting derivatives to 0 gives

$$\begin{aligned} \frac{\partial}{\partial\alpha}\mathrm{E}[y - f(\boldsymbol{x})]^2 &= 2(\alpha - \mu_y) + 2\boldsymbol{\mu}_{\boldsymbol{x}}^T\boldsymbol{\beta} = 0 \\ \nabla_{\boldsymbol{\beta}}\mathrm{E}[y - f(\boldsymbol{x})]^2 &= 2\mathrm{E}(\boldsymbol{x}\boldsymbol{x}^T)\boldsymbol{\beta} - 2\mathrm{E}(\boldsymbol{x}y) + 2\alpha\boldsymbol{\mu}_{\boldsymbol{x}} = \mathbf{0}_p. \end{aligned}$$

From the first equation, $\alpha = \mu_y - \boldsymbol{\mu}_{\boldsymbol{x}}^T \boldsymbol{\beta}$. Substitution into the second equation yields

$$(\mathrm{E}\boldsymbol{x}\boldsymbol{x}^T - \boldsymbol{\mu}_{\boldsymbol{x}}\boldsymbol{\mu}_{\boldsymbol{x}}^T)\boldsymbol{\beta} = \mathrm{E}(\boldsymbol{x}y) - \boldsymbol{\mu}_{\boldsymbol{x}}\mu_y.$$

That is

$$\boldsymbol{\Sigma}_{\boldsymbol{x}\boldsymbol{x}}\boldsymbol{\beta} = \boldsymbol{\Sigma}_{\boldsymbol{x}y}.$$

Therefore the optimal α is $\alpha^* = \mu_y - \boldsymbol{\mu}_{\boldsymbol{x}}^T \beta^*$, where $\boldsymbol{\beta}^*$ is any solution to above equation. And the BLP is

$$\alpha^* + \boldsymbol{x}^T\boldsymbol{\beta}^* = \mu_y + (\boldsymbol{x} - \boldsymbol{\mu}_{\boldsymbol{x}})^T\boldsymbol{\beta}^*.$$

Because the criterion function is a convex function, the stationary condition is both necessary and sufficient for the global minima. Therefore any BLP must be of this form. □

Best linear unbiased prediction (BLUP)

- Consider random variables $y_0, y_1, \ldots, y_n$, and we are interested predicting y_0 given data $y_1, \ldots, y_n$. If we know the mean, variances, and covariances, then we can use above theory to find the BLP of y_0.

- In practice, we don't know the means $\mu_0 = \mathrm{E}y_0, \boldsymbol{\mu}_{\boldsymbol{y}} = \mathrm{E}\boldsymbol{y}$ most of time ☹
  Let's impose a linear (Aitken) model for the means μ_i

$$\mathrm{E}\begin{pmatrix}\boldsymbol{y}\\ y_0\end{pmatrix} = \begin{pmatrix}\boldsymbol{X}\boldsymbol{b}\\ \boldsymbol{x}_0^T\boldsymbol{b}\end{pmatrix}, \quad \mathrm{Cov}\begin{pmatrix}\boldsymbol{y}\\ y_0\end{pmatrix} = \begin{pmatrix}\boldsymbol{V} & \boldsymbol{V}_{\boldsymbol{y}y_0}\\ \boldsymbol{V}_{\boldsymbol{y}y_0}^T & V_{y_0y_0}\end{pmatrix}. \tag{6}$$

- If $\boldsymbol{b}$ is known, then the BLP of y_0 is

$$\boldsymbol{x}_0^T\boldsymbol{b} + (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b})^T\boldsymbol{\beta}^*,$$

  where $\boldsymbol{\beta}^*$ is a solution to $\boldsymbol{V}\boldsymbol{\beta} = \boldsymbol{V}_{\boldsymbol{y}y_0}$.

- If $\boldsymbol{b}$ is unknown, then the hope is to find the *best linear unbiased predictor* (BLUP).

  We call a linear predictor $f(\boldsymbol{y}) = a_0 + \boldsymbol{a}^T\boldsymbol{y}$ of y_0 BLUP if

1. it is unbiased, i.e.,

$$\mathrm{E}\, f(\boldsymbol{y}) = a_0 + \boldsymbol{a}^T \boldsymbol{X}\boldsymbol{b} = \boldsymbol{x}_0^T \boldsymbol{b} = \mathrm{E} y_0$$

   for all $\boldsymbol{b}$, and

2. for any other linear unbiased predictor $b_0 + \boldsymbol{b}^T \boldsymbol{y}$,

$$\mathrm{E}(y_0 - a_0 - \boldsymbol{a}^T \boldsymbol{y})^2 \le \mathrm{E}(y_0 - b_0 - \boldsymbol{b}^T \boldsymbol{y})^2.$$

- Theorem: Under the Aitken model (6) and assume $\boldsymbol{V}_{\boldsymbol{y}y_0} \in \mathcal{C}(\boldsymbol{V}, \boldsymbol{X})$ and $\boldsymbol{x}_0 \in \mathcal{C}(\boldsymbol{X}^T)$, the BLUP of y_0 is

$$\boldsymbol{x}_0^T \hat{\boldsymbol{b}}_{\mathrm{GLS}} + (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{b}}_{\mathrm{GLS}})^T \boldsymbol{\beta}_*,$$

  where $\boldsymbol{\beta}_*$ is a solution of $(\boldsymbol{V} + \boldsymbol{X}\boldsymbol{X}^T)\boldsymbol{\beta} = \boldsymbol{V}_{\boldsymbol{y}y_0}$ and $\boldsymbol{X}\hat{\boldsymbol{b}}_{\mathrm{GLS}}$ is the BLUE of $\boldsymbol{X}\boldsymbol{b}$.

  Remark 1: We don't assume $\boldsymbol{V}$ is nonsingular. For nonsingular $\boldsymbol{V}$, we can take $\boldsymbol{\beta}_* = \boldsymbol{V}^{-1}\boldsymbol{V}_{\boldsymbol{y}y_0}$.

  Remark 2: Both $\hat{\boldsymbol{b}}_{\mathrm{GLS}}$ and $\boldsymbol{\beta}^*$ depend crucially on $\boldsymbol{V}$.

  Proof. Let $a_0 + \boldsymbol{a}^T \boldsymbol{y}$ be arbitrary linear predictor of y_0. Unbiasedness requires

$$a_0 + \boldsymbol{a}^T \boldsymbol{X}\boldsymbol{b} = \boldsymbol{x}_0^T \boldsymbol{b}$$

  for all $\boldsymbol{b}$. Thus $a_0 = 0$ and $\boldsymbol{a}^T \boldsymbol{X} = \boldsymbol{x}_0^T$. We need to solve the constrained optimization problem

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}\mathrm{E}(y_0 - \boldsymbol{a}^T \boldsymbol{y})^2 = \frac{1}{2}\boldsymbol{a}^T \mathrm{E}(\boldsymbol{y}\boldsymbol{y}^T)\boldsymbol{a} - \mathrm{E}(y_0 \boldsymbol{a}^T \boldsymbol{y}) + \frac{1}{2}\mathrm{E} y_0^2 \\ \text{subject to} & \boldsymbol{X}^T \boldsymbol{a} = \boldsymbol{x}_0. \end{array}$$

  Setting the gradient of the Lagrangian

$$L(\boldsymbol{a}, \boldsymbol{\lambda}) = \frac{1}{2}\boldsymbol{a}^T \mathrm{E}(\boldsymbol{y}\boldsymbol{y}^T)\boldsymbol{a} - \mathrm{E}(y_0 \boldsymbol{a}^T \boldsymbol{y}) + \frac{1}{2}\mathrm{E} y_0^2 + \boldsymbol{\lambda}^T (\boldsymbol{X}^T \boldsymbol{a} - \boldsymbol{x}_0)$$

  to zero yields equations

$$\begin{aligned} \boldsymbol{E}(\boldsymbol{y}\boldsymbol{y}^T)\boldsymbol{a} - \boldsymbol{E}(\boldsymbol{y} y_0) + \boldsymbol{X}\boldsymbol{\lambda} &= \boldsymbol{0}_n \\ \boldsymbol{X}^T \boldsymbol{a} &= \boldsymbol{x}_0. \end{aligned}$$

Adding and subtracting $(\mathrm{E}\boldsymbol{y}\mathrm{E}\boldsymbol{y}^T)\boldsymbol{a} = \boldsymbol{X}\boldsymbol{b}\boldsymbol{b}^T\boldsymbol{X}^T\boldsymbol{a} = \boldsymbol{X}\boldsymbol{b}\boldsymbol{b}^T\boldsymbol{x}_0 = \mathrm{E}\boldsymbol{y}\mathrm{E}y_0$ to the first equation shows $\boldsymbol{V}\boldsymbol{a} - \boldsymbol{V}_{\boldsymbol{y}y_0} + \boldsymbol{X}\boldsymbol{\lambda} = \boldsymbol{0}_n$. In matrix notation,

$$\begin{pmatrix} \boldsymbol{V} & \boldsymbol{X} \\ \boldsymbol{X}^T & \boldsymbol{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{a} \\ \boldsymbol{\lambda} \end{pmatrix} = \begin{pmatrix} \boldsymbol{V}_{\boldsymbol{y}y_0} \\ \boldsymbol{x}_0 \end{pmatrix}.$$

By HW4 5(d), solution for the optimal $\boldsymbol{a}$ is

$$\boldsymbol{a}_* = [\boldsymbol{V}_0^- - \boldsymbol{V}_0^-\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{V}_0^-\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{V}_0^-]\boldsymbol{V}_{\boldsymbol{y}y_0} + \boldsymbol{V}_0^-\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{V}_0^-\boldsymbol{X})^-\boldsymbol{x}_0,$$

where $\boldsymbol{V}_0 = \boldsymbol{V} + \boldsymbol{X}\boldsymbol{X}^T$. Thus the BLUP is

$$\begin{aligned} \boldsymbol{a}_*^T\boldsymbol{y} &= \boldsymbol{V}_{\boldsymbol{y}y_0}^T[\boldsymbol{V}_0^- - \boldsymbol{V}_0^-\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{V}_0^-\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{V}_0^-]\boldsymbol{y} + \boldsymbol{x}_0^T(\boldsymbol{X}^T\boldsymbol{V}_0^-\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{V}_0^-\boldsymbol{y} \\ &= \boldsymbol{V}_{\boldsymbol{y}y_0}^T\boldsymbol{V}_0^-(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{b}}_{\mathrm{GLS}}) + \boldsymbol{x}_0^T\hat{\boldsymbol{b}}_{\mathrm{GLS}} \\ &= \boldsymbol{\beta}_*^T(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{b}}_{\mathrm{GLS}}) + \boldsymbol{x}_0^T\hat{\boldsymbol{b}}_{\mathrm{GLS}}. \end{aligned}$$

□

- The prediction variance of BLUP is (TODO)

$$\mathrm{E}(y_0 - \boldsymbol{a}_*^T\boldsymbol{y})^2 = V_{y_0y_0} - 2\boldsymbol{a}_*^T\boldsymbol{V}_{\boldsymbol{y}y_0} + \boldsymbol{a}_*^T\boldsymbol{V}\boldsymbol{a}_*. \quad ???$$

- Example (BLUP in Gauss-Markov linear model): $\boldsymbol{V} = \sigma^2\boldsymbol{I}$ and $\boldsymbol{V}_{\boldsymbol{y}y_0} = \boldsymbol{0}_p$. Thus $\boldsymbol{\beta}_* = \boldsymbol{0}_p$ and the BLUP for y_0 is $\boldsymbol{x}_0^T\hat{\boldsymbol{b}}$, which is also the BLUE of $\boldsymbol{x}_0^T\boldsymbol{b}$.

Mixed model equation (MME)

- Consider the mixed effects model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e}$$

  - $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ is a design matrix for fixed effects $\boldsymbol{b} \in \mathbb{R}^p$.
  - $\boldsymbol{Z} \in \mathbb{R}^{n\times q}$ is a design matrix for random effects $\boldsymbol{u} \in \mathbb{R}^q$.
  - The most general assumption is $\boldsymbol{e} \in N(\boldsymbol{0}_n, \boldsymbol{R})$, $\boldsymbol{u} \in N(\boldsymbol{0}_q, \boldsymbol{G})$, and $\boldsymbol{e}$ is independent of $\boldsymbol{u}$. That is

$$\begin{pmatrix} \boldsymbol{u} \\ \boldsymbol{e} \end{pmatrix} \sim N\left(\boldsymbol{0}_{q+n}, \begin{pmatrix} \boldsymbol{G} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{R} \end{pmatrix}\right).$$

    Assume $\boldsymbol{G}$ and $\boldsymbol{R}$ are nonsingular.

- We already know that the BLUE of $\boldsymbol{X}\boldsymbol{b}$ is

$$\boldsymbol{X}\hat{\boldsymbol{b}}_{\text{GLS}} = \boldsymbol{X}[\boldsymbol{X}^T(\boldsymbol{R}+\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^T)^{-1}\boldsymbol{X}]^-\boldsymbol{X}^T(\boldsymbol{R}+\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^T)^{-1}\boldsymbol{y}.$$

- We can apply the previous theorem to derive the BLUP of $\boldsymbol{u}$. Note

$$\text{E}\begin{pmatrix}\boldsymbol{y}\\ \boldsymbol{u}\end{pmatrix} = \begin{pmatrix}\boldsymbol{X}\boldsymbol{b}\\ \boldsymbol{0}_q\end{pmatrix}, \quad \text{Cov}\begin{pmatrix}\boldsymbol{y}\\ \boldsymbol{u}\end{pmatrix} = \begin{pmatrix}\boldsymbol{R}+\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^T & \boldsymbol{Z}\boldsymbol{G}\\ \boldsymbol{G}\boldsymbol{Z}^T & \boldsymbol{R}\end{pmatrix}.$$

  Therefore the BLUP for $\boldsymbol{u}$ is

$$\boldsymbol{G}\boldsymbol{Z}^T(\boldsymbol{R}+\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^T)^{-1}(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{b}}_{\text{GLS}}).$$

- It turns out both BLUE of $\boldsymbol{X}\boldsymbol{b}$ and BLUP of $\boldsymbol{u}$ can be obtained simultaneously by solving a so-called mixed model equation (MME).

  Mixed model equation (MME) defined as

$$\begin{pmatrix}\boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{Z}\\ \boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{G}^{-1}+\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z}\end{pmatrix}\begin{pmatrix}\boldsymbol{b}\\ \boldsymbol{u}\end{pmatrix} = \begin{pmatrix}\boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{y}\\ \boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{y}\end{pmatrix}$$

  is a generalization of the normal equation for fixed effects model.

- Theorem: Let $(\hat{\boldsymbol{b}}, \hat{\boldsymbol{u}})$ be a solution to MME. Then $\boldsymbol{X}\hat{\boldsymbol{b}}$ is the BLUE of $\boldsymbol{X}\boldsymbol{b}$ and $\hat{\boldsymbol{u}}$ is the BLUP of $\boldsymbol{u}$.

  Proof. Let

$$\boldsymbol{V} = \text{Cov}(\boldsymbol{y}) = \boldsymbol{R}+\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^T.$$

  Then $\boldsymbol{X}\hat{\boldsymbol{b}}$ is a BLUE of $\boldsymbol{X}\boldsymbol{b}$ if $\hat{\boldsymbol{b}}$ is a solution to

$$\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}\boldsymbol{b} = \boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{y}.$$

  By the binomial inversion formula (HW1), we have

$$\boldsymbol{V}^{-1} = \boldsymbol{R}^{-1} - \boldsymbol{R}^{-1}\boldsymbol{Z}(\boldsymbol{G}^{-1}+\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{R}^{-1}.$$

  The MME says

$$\begin{aligned}\boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{X}\boldsymbol{b}+\boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{Z}\boldsymbol{u} &= \boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{y}\\ \boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{X}\boldsymbol{b}+(\boldsymbol{G}^{-1}+\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z})\boldsymbol{u} &= \boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{y}.\end{aligned}$$

From the second equation we solve for

$$\boldsymbol{u} = (\boldsymbol{G}^{-1} + \boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{R}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b})$$

and substituting into the first equation shows

$$\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}\hat{\boldsymbol{b}} = \boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{y}.$$

Thus $\hat{\boldsymbol{b}}$ is a generalized least squares solution and $\boldsymbol{X}\hat{\boldsymbol{b}}$ is the BLUE of $\boldsymbol{X}\boldsymbol{b}$.

To show that $\hat{\boldsymbol{u}}$ is BLUP, only need to show that

$$(\boldsymbol{G}^{-1} + \boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{R}^{-1} \;\; = \;\; \boldsymbol{G}\boldsymbol{Z}^T(\boldsymbol{R} + \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^T)^{-1},$$

or equivalently

$$\boldsymbol{Z}^T\boldsymbol{R}^{-1}(\boldsymbol{R} + \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^T) = (\boldsymbol{G}^{-1} + \boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z})\boldsymbol{G}\boldsymbol{Z}^T,$$

which is obvious. □

References

Casella, G. and Berger, R. (2001). *Statistical Inference*. Duxbury Resource Center.

Christensen, R. (1996). Exact tests for variance components. *Biometrics*, 52(1):309–314.

Crainiceanu, C. M. and Ruppert, D. (2004). Likelihood ratio tests in linear mixed models with one variance component. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 66(1):165–185.

Davies, R. B. (1980). The distribution of linear combinations of χ^2 random variables. *Applied Statistics*, 29:323–333.

Öfversten, J. (1993). Exact tests for variance components in unbalanced mixed linear models. *Biometrics*, 49(1):45–57.

Seely, J. F. and El Bassiouni, Y. (1983). Applying Wald's variance component test. *Ann. Statist.*, 11(1):197–201.

Teets, D. and Whitehead, K. (1999). The discovery of Ceres: how Gauss became famous. *Math. Mag.*, 72(2):83–93.