33 Lecture 33: April 25

Last time

- Lack of fit test
- Theoretical background of linear models

Today

- Theoretical background of linear models
- Course evaluation started (current: 2/17)

Additional reference

Course notes by Dr. Hua Zhou
"A Primer on Linear Models" by Dr. John F. Monahan

Estimable function

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

- A parametric function $\mathbf{\Lambda}\mathbf{b}$, $\mathbf{\Lambda} \in \mathbb{R}^{m \times p}$ is said to be (linearly) estimable if there exists an affinely unbiased estimator of $\mathbf{\Lambda}\mathbf{b}$ for all $\mathbf{b} \in \mathbb{R}^p$. That is there exist constants $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $\mathbf{c} \in \mathbb{R}^m$ such that $\mathrm{E}(\mathbf{A}\mathbf{y} + \mathbf{c}) = \mathbf{\Lambda}\mathbf{b}$ for all $\mathbf{b}$.
- Theorem: Assuming the linear mean model, the parametric function $\mathbf{\Lambda}\mathbf{b}$ is (linearly) estimable if and only if $\mathcal{C}(\mathbf{\Lambda}) \subset \mathcal{C}(\mathbf{X}^T)$, or equivalently $\mathcal{N}(\mathbf{X}) \subset \mathcal{N}(\mathbf{\Lambda})$.
  "$\mathbf{\Lambda}\mathbf{b}$ is estimable $\iff$ the row space of $\mathbf{\Lambda}$ is contained in the row space of $\mathbf{X}$ $\iff$ the null space of $\mathbf{X}$ is contained in the null space of $\mathbf{\Lambda}$."
  Proof:
- $\lambda^T\mathbf{b}$ is linearly estimable if and only if $\lambda^T\mathbf{b}$ is a linear combination of the components in $\mu_Y = \mathrm{E}(\mathbf{Y})$
- Corollary: $\mathbf{X}\mathbf{b}$ is estimable.
  "Expected value of any observation $\mathrm{E}(y_i)$ and their linear combinations are estimable."
- Corollary: If $\mathbf{X}$ has full column rank, then any linear combinations of $\mathbf{b}$ are estimable.
- If $\mathbf{\Lambda}\mathbf{b}$ is (linearly) estimable, then its *least squares estimator* $\mathbf{\Lambda}\hat{\mathbf{b}}$ is invariant to the choice of the least squares solution $\hat{\mathbf{b}}$.
  Proof:
- The least squares estimator $\mathbf{\Lambda}\hat{\mathbf{b}}$ is a linearly unbiased estimator of $\mathbf{\Lambda}\mathbf{b}$.
  Proof:

Estimability example: One-way ANOVA model

Consider the following example with one-way ANOVA model.

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij} \quad i = 1, 2, 3,\ j = 1, 2$$

In matrix form:

$$\begin{bmatrix} Y_{11} \\ Y_{21} \\ Y_{31} \\ Y_{12} \\ Y_{22} \\ Y_{32} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} + \begin{bmatrix} \epsilon_{11} \\ \epsilon_{21} \\ \epsilon_{31} \\ \epsilon_{12} \\ \epsilon_{22} \\ \epsilon_{32} \end{bmatrix}$$

Note: replication doesn't help with estimability. What functions of $\lambda^T \mathbf{b}$ are estimable?
Solutions:

Idempotent matrix

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

- A matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ is idempotent if and only if $\mathbf{A}^2 (= \mathbf{AA}) = \mathbf{A}$.
- Any idempotent matrix $\mathbf{A}$ is a generalized inverse of itself.
- The only idempotent matrix of full rank is $\mathbf{I}$.
  Proof. Interpretation: all idempotent matrices are singular except for the identity matrix.
- $\mathbf{A}$ is idempotent if and only if $\mathbf{A}^T$ is idempotent if and only if $\mathbf{I}_n - \mathbf{A}$ is idempotent.
- For a general matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$, the matrices $\mathbf{A}^-\mathbf{A}$ and $\mathbf{AA}^-$ are idempotent and

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

Projection

- A matrix $\mathbf{P} \in \mathbb{R}^{m \times n}$ is a projection onto a vector space $\mathcal{V}$ if and only if
  1. $\mathbf{P}$ is idempotent
  2. $\mathbf{Px} \in \mathcal{V}$ for any $\mathbf{x} \in \mathbb{R}^n$
  3. $\mathbf{Pz} = \mathbf{z}$ for any $\mathbf{z} \in \mathcal{V}$.
- Any idempotent matrix $\mathbf{P}$ is a projection onto its own column space $\mathcal{C}(\mathbf{P})$.
  Proof:
- $\mathbf{AA}^-$ is a projection onto the column space $\mathcal{C}(\mathbf{A})$.
  Proof:

- Proposition: Let $\mathbf{X}, \mathbf{A}, \mathbf{B}$ be matrices, then $\mathbf{X}^T\mathbf{X}\mathbf{A} = \mathbf{X}^T\mathbf{X}\mathbf{B}$ if and only if $\mathbf{X}\mathbf{A} = \mathbf{X}\mathbf{B}$.
  Proof:

- The projection matrix
$$\underset{n\times n}{\mathbf{P}_\mathbf{X}} = \underset{n\times p}{\mathbf{X}}(\underset{p\times p}{\mathbf{X}^T\mathbf{X}})^-\underset{p\times n}{\mathbf{X}^T}$$
  is unique.
  Proof:

- Start with $\mathbf{P}_\mathbf{X}\mathbf{X} = \mathbf{X}$, we have $\mathbf{X}(\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T\mathbf{X} = \mathbf{X}$. Therefore, $(\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T$ is a generalized inverse of $\mathbf{X}$ which is sometimes called the least-squares inverse. And $\mathbf{P}_\mathbf{X}$ is a projection onto $\mathcal{C}(\mathbf{X})$.

- $\underset{n\times n}{\mathbf{P}_\mathbf{X}}\ \underset{n\times p}{\mathbf{X}} = \underset{n\times p}{\mathbf{X}}$
  Proof:

- Predicted values $\hat{\mathbf{Y}} = \mathbf{X}\hat{\mathbf{b}}_{ls}$ are invariant to choice of solution to the normal equation, where
$$\hat{\mathbf{b}}_{ls} = (\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T\mathbf{Y}$$
  is not necessarily unique.
  Proof:

Geometry of least squares

- $\mathbf{P}_\mathbf{X}^2 = \mathbf{P}_\mathbf{X}$ and $\hat{\mathbf{Y}} = \mathbf{P}_\mathbf{X}\mathbf{Y}$ is unique.

- Recall the column space of $\mathbf{X}$ is $\mathcal{C}(\mathbf{X}) = \left\{ \underset{n\times 1}{\mathbf{y}} : \mathbf{y} = \underset{p\times 1}{\mathbf{X}\,\mathbf{b}} \text{ for some } \mathbf{b} \right\}$

- The vector in $\mathcal{C}(\mathbf{X})$ that is closest in terms of squared norm (L_2 norm: $||\mathbf{a}-\mathbf{b}||_2 = \sqrt{(\mathbf{a}-\mathbf{b})^T(\mathbf{a}-\mathbf{b})}$) to $\mathbf{Y}$ is given by $\hat{\mathbf{Y}} = \mathbf{X}\hat{\mathbf{b}}_{ls} = \mathbf{P}_\mathbf{X}\mathbf{Y}$.
  Proof:

- $\hat{\mathbf{Y}} \in \mathcal{C}(\mathbf{X})$

- $\underset{n\times 1}{\hat{\mathbf{e}}} = \mathbf{Y} - \hat{\mathbf{Y}} = (\mathbf{I} - \mathbf{P}_\mathbf{X})\mathbf{Y} \in \mathcal{N}(\mathbf{X}^T)$ where $\mathcal{N}(\mathbf{X}^T) = \left\{ \underset{n\times 1}{\mathbf{v}} : \mathbf{X}^T\mathbf{v} = \mathbf{0} \right\}$ is the null space of $\mathbf{X}^T$.
  Proof:

Normal distribution in scaler case

- A random variable Z has a standard normal distribution, denoted $Z \sim \mathcal{N}(0,1)$, if
$$F_Z(t) = \Pr(Z \leqslant t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz,$$

or equivalently Z has density

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

or equivalently, Z has moment generating function (mgf)

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

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

    $$X = \mu + \sigma Z$$

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

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

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

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

Multivariate normal distribution

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

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

    The mgf is

    $$m_{\mathbf{Z}}(\mathbf{t}) = \prod_{i=1}^p m_{Z_i}(t_1) = \prod_{i=1}^p e^{t_i^2/2} = e^{\mathbf{t}^T\mathbf{t}/2}.$$

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

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

    and the moment generating function is

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

  - $\mathbf{X} \in \mathbb{R}^p$ has a multivariate normal distribution with mean $\boldsymbol{\mu} \in \mathbb{R}^p$ and covariance $\mathbf{V} \in \mathbb{R}^{p \times p}, \mathbf{V} \succeq_{p.s.d.} \mathbf{0}$, denoted $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{V})$, if its mgf takes the form

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

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

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

- (Any affine transform of normal is normal) If $\mathbf{X} \in \mathbb{R}^p, \mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{V})$ and $\mathbf{Y} = \mathbf{a} + \mathbf{B}\mathbf{X}$, where $\mathbf{a} \in \mathbb{R}^q$ and $\mathbf{B} \in \mathbb{R}^{q \times p}$, then $\mathbf{Y} \sim \mathcal{N}(\mathbf{a} + \mathbf{B}\boldsymbol{\mu}, \mathbf{B}\mathbf{V}\mathbf{B}^T)$.
- (Marginal of normal is normal) If $\mathbf{X} \in \mathbb{R}^p, \mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{V})$, then any subvector of $\mathbf{X}$ is normal too.
- A convenient fact about normal random variables/vectors is that zero correlation/covariance implies independence.
  If $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{V})$ and is partitioned as

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

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

Independence and Cochran's theorem

- (Independence between two linear forms of a multivariate normal) Let $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{V})$, $\mathbf{Y}_1 = \mathbf{a}_1 + \mathbf{B}_1\mathbf{X}$ and $\mathbf{Y}_2 = \mathbf{a}_2 + \mathbf{B}_2\mathbf{X}$. Then $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are independent if and only if $\mathbf{B}_1\mathbf{V}\mathbf{B}_2^T = 0$.
  Proof:
- Consider the normal linear model $\mathbf{y} \sim \mathcal{N}(\mathbf{X}\mathbf{b}, \sigma^2\mathbf{I}_n)$
  - Using $\mathbf{A} = (1/\sigma^2)(\mathbf{I} - \mathbf{P}_{\mathbf{X}})$, we have

$$SSE/\sigma^2 = ||\hat{\boldsymbol{\epsilon}}||_2^2/\sigma^2 = \mathbf{y}^T\mathbf{A}\mathbf{y} \sim \chi^2_{n-r},$$

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

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

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

$$SSR/\sigma^2 = ||\hat{\mathbf{y}}||_2^2/\sigma^2 = \mathbf{y}^T\mathbf{A}\mathbf{y} \sim \chi^2_r(\phi),$$

    with the noncentrality parameter

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

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

$$\begin{bmatrix} \hat{\mathbf{y}} \\ \hat{\boldsymbol{\epsilon}} \end{bmatrix} = \begin{bmatrix} \mathbf{P_X} \\ \mathbf{I}_n - \mathbf{P_X} \end{bmatrix} \mathbf{y} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{Xb} \\ \mathbf{0}_n \end{bmatrix}, \begin{bmatrix} \sigma^2 \mathbf{P_X} & \mathbf{0} \\ \mathbf{0} & \sigma^2(\mathbf{I} - \mathbf{P_X}) \end{bmatrix} \right).$$

So $\hat{\mathbf{y}}$ is independent of $\hat{\boldsymbol{\epsilon}}$. Thus $||\hat{\mathbf{y}}||_2^2/\sigma^2$ is independent of $||\hat{\boldsymbol{\epsilon}}||_2^2/\sigma^2$ and

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

- (Independence between linear and quadratic forms of a multivariate normal) Let $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{V})$. Let $\mathbf{A}$ be symmetric with rank s. Then $\mathbf{BX}$ and $\mathbf{X}^T\mathbf{AX}$ are independent if $\mathbf{BVA} = \mathbf{0}$.
  Proof:

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

- (Cochran's theorem) Let $\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2\mathbf{I}_n)$ and $\mathbf{A}_i$, $i = 1, \ldots, k$ be symmetric idempotent matrix with rank s_i. If $\sum_{i=1}^k \mathbf{A}_i = \mathbf{I}_n$, then $(1/\sigma^2)\mathbf{y}^T\mathbf{A}_i\mathbf{y}$ are independent $\chi^2_{s_i}(\phi_i)$, with $\phi_i = \frac{1}{2\sigma^2}\boldsymbol{\mu}^T\mathbf{A}_i\boldsymbol{\mu}$ and $\sum_{i=1}^k s_i = n$.
  Proof:

- Application to the one-way ANOVA: $y_{ij} = \mu + \alpha_i + \epsilon_{ij}$. We have the classical ANOVA table

Source	df	Projection	SS	Noncentrality
Mean	1	$\mathbf{P_1}$	$SSM = n\bar{y}^2$	$\frac{1}{2\sigma^2}n(\mu + \bar{\alpha})^2$
Group	$a-1$	$\mathbf{P_X} - \mathbf{P_1}$	$SSA = \sum_{i=1}^a n_i\bar{y}_i^2 - n\bar{y}^2$	$\frac{1}{2\sigma^2}\sum_{i=1}^a n_i(\alpha_i - \bar{\alpha})^2$
Error	$n-a$	$\mathbf{I} - \mathbf{P_X}$	$SSE = \sum_{i=1}^a \sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2$	0
Total	n	$\mathbf{I}$	$SST = \sum_i \sum_j y_{ij}^2$	$\frac{1}{\sigma^2}\sum_{i=1}^a n_i(\mu + \alpha_i)^2$