Maximum Likelihood Estimation

STA 721: Lecture 2

Merlise Clyde (clyde@duke.edu)
Duke University

Outline

- Likelihood Function
- Projections
- Maximum Likelihood Estimates

Readings: Christensen Chapter 1-2, Appendix A, and Appendix B

Normal Model

Take an random vector $\mathbf{Y} \in \mathbb{R}^n$ which is observable and decompose

$$\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\epsilon}$$

- $\boldsymbol{\mu} \in \mathbb{R}^n$ (unknown, fixed)
- $\boldsymbol{\epsilon} \in \mathbb{R}^n$ unobservable error vector (random)

Usual assumptions?

- $E[\epsilon_i] = 0 \; \forall i \Leftrightarrow \mathsf{E}[\boldsymbol{\epsilon}] = \mathbf{0} \quad \Rightarrow \mathsf{E}[\mathbf{Y}] = \boldsymbol{\mu}$ (mean vector)
- ϵ_i independent with $\mathsf{Var}(\epsilon_i) = \sigma^2$ and $\mathsf{Cov}(\epsilon_i, \epsilon_j) = 0$
- Matrix version
  $\mathsf{Cov}[\boldsymbol{\epsilon}] \equiv [(\mathsf{E}\,[(\epsilon_i - \mathsf{E}[\epsilon_i])(\epsilon_j - \mathsf{E}[\epsilon_j])]]_{ij} = \sigma^2 \mathbf{I}_n \quad \Rightarrow \mathsf{Cov}[\mathbf{Y}] = \sigma^2 \mathbf{I}_n$ (errors are uncorrelated)
- $\boldsymbol{\epsilon}_i \stackrel{\text{iid}}{\sim} \mathsf{N}(0, \sigma^2)$ implies that $Y_i \stackrel{\text{ind}}{\sim} \mathsf{N}(\mu_i, \sigma^2)$

Likelihood Function

The likelihood function for $\boldsymbol{\mu}, \sigma^2$ is proportional to the sampling distribution of the data

$$
\begin{aligned}
\mathcal{L}(\boldsymbol{\mu}, \sigma^2) &\propto \prod_{i=1}^{n} \frac{1}{\sqrt{(2\pi\sigma^2)}} \exp - \frac{1}{2}\left\{ \frac{(Y_i - \mu_i)^2}{\sigma^2} \right\} \\
&\propto (2\pi\sigma^2)^{-n/2} \exp\left\{ - \frac{1}{2} \frac{\sum_i (Y_i - \mu_i)^2)}{\sigma^2} \right\} \\
&\propto (\sigma^2)^{-n/2} \exp\left\{ - \frac{1}{2} \frac{\|\mathbf{Y} - \boldsymbol{\mu}\|^2}{\sigma^2} \right\} \\
&\propto (2\pi)^{-n/2} |\mathbf{I}_n \sigma^2|^{-1/2} \exp\left\{ - \frac{1}{2} \frac{\|\mathbf{Y} - \boldsymbol{\mu}\|^2}{\sigma^2} \right\}
\end{aligned}
$$

Last line is the density of $\mathbf{Y} \sim \mathsf{N}_n\left(\boldsymbol{\mu}, \sigma^2 \mathbf{I}_n\right)$

MLEs

Find values of $\hat{\boldsymbol{\mu}}$ and $\hat{\sigma}^2$ that maximize the likelihood $\mathcal{L}(\boldsymbol{\mu}, \sigma^2)$ for $\boldsymbol{\mu} \in \mathbb{R}^n$ and $\boldsymbol{\sigma}^2 \in \mathbb{R}^+$

$$\mathcal{L}(\boldsymbol{\mu}, \sigma^2) \propto (\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2}\frac{\|\mathbf{Y} - \boldsymbol{\mu}\|^2}{\sigma^2}\right\}$$

$$\log(\mathcal{L}(\boldsymbol{\mu}, \sigma^2)) \propto -\frac{n}{2}\log(\sigma^2) - \frac{1}{2}\frac{\|\mathbf{Y} - \boldsymbol{\mu}\|^2}{\sigma^2}$$

or equivalently the log likelihood

- Clearly, $\hat{\boldsymbol{\mu}} = \mathbf{Y}$ but $\hat{\sigma}^2 = 0$ is outside the parameter space
- If $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$, can show that $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ is the MLE/OLS estimator of $\boldsymbol{\beta}$ and $\hat{\boldsymbol{\mu}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ if $\mathbf{X}$ is full column rank.
- show via projections

Projections

take any point $\mathbf{y} \in \mathbb{R}^n$ and "project" it onto $C(\mathbf{X}) = \boldsymbol{\mathcal{M}}$

- any point already in $\boldsymbol{\mathcal{M}}$ stays the same
- so if $\mathbf{P_X}$ is a projection onto the column space of $\mathbf{X}$ then for $\mathbf{m} \in C(\mathbf{X})$ $\mathbf{P_X m} = \mathbf{m}$
- $\mathbf{P_X}$ is a linear transformation from $\mathbb{R}^n \to \mathbb{R}^n$
- maps vectors in $\mathbb{R}^n$ into $C(\mathbf{X})$
- if $\mathbf{z} \in \mathbb{R}^n$ then $\mathbf{P_X z} = \mathbf{Xa} \in C(\mathbf{X})$ for some $\mathbf{a} \in \mathbb{R}^p$

Example

For $\mathbf{X} \in \mathbb{R}^{n \times p}$, rank p, $\mathbf{P_X} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}$ is a projection onto the p dimensional subspace $\boldsymbol{\mathcal{M}} = C(\mathbf{X})$

Idempotent Matrix

What if we project a projection?

- $\mathbf{P}_{\mathbf{X}}\mathbf{z} = \mathbf{X}\mathbf{a} \in C(\mathbf{X})$
- $\mathbf{P}_{\mathbf{X}}\mathbf{X}\mathbf{a} = \mathbf{X}\mathbf{a}$
- since $\mathbf{P}_{\mathbf{X}}^2\mathbf{z} = \mathbf{P}_{\mathbf{X}}\mathbf{z}$ for all $\mathbf{z} \in \mathbb{R}^n$ we have $\mathbf{P}_{\mathbf{X}}^2 = \mathbf{P}_{\mathbf{X}}$

▼ **Definition:** Projection

For a matrix $\mathbf{P}$ in $\mathbb{R}^{n \times n}$ is a projection matrix if $\mathbf{P}^2 = \mathbf{P}$. That is all projections $\mathbf{P}$ are idempotent matrix.

Exercise

For $\mathbf{X} \in \mathbb{R}^{n \times p}$, rank p, if $\mathbf{P}_{\mathbf{X}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}$ use the definition to show that it is a projection onto the p dimensional subspace $\boldsymbol{\mathcal{M}} = C(\mathbf{X})$

Null Space

Definition: Orthogonal Complement

The set of all vectors that are orthogonal to a given subspace $\boldsymbol{\mathcal{M}}$ is called the *orthogonal complement* of the subspace denoted as $\boldsymbol{\mathcal{M}}^{\perp}$. Under the usual inner product, $\boldsymbol{\mathcal{M}}^{\perp} \equiv \{\mathbf{n} \in \mathbb{R}^n \ni \mathbf{m}^T\mathbf{n} = 0 \text{ for } \mathbf{m} \in \boldsymbol{\mathcal{M}}\}$

Definition: Null Space

For a matrix $\mathbf{A}$, the *null space* of $\mathbf{A}$ is defined as $N(\mathbf{A}) = \{\mathbf{n} \ni \mathbf{A}\mathbf{n} = \mathbf{0}\}$

Exercise

Show that $C(\mathbf{X})^{\perp}$ (the *orthogonal complement* of $C(\mathbf{X})$) is the *null space* of $\mathbf{X}^T$, $N(\mathbf{X}^T)$.

Orthogonal Projection

▼ **Definition:** Orthogonal Projections

For a vector space $\mathcal{V}$ with an inner product $\langle \mathbf{x}, \mathbf{y} \rangle$ for $\mathbf{x}, \mathbf{y} \in \mathcal{V}$, $\mathbf{x}$ and $\mathbf{y}$ are orthogonal if $\langle \mathbf{x}, \mathbf{y} \rangle = 0$. A projection $\mathbf{P}$ is an *orthogonal projection* onto a subspace $\boldsymbol{\mathcal{M}}$ of $\mathcal{V}$ if for any $\mathbf{m} \in \mathcal{V}$, $\mathbf{Pm} = \mathbf{m}$ and for any $\mathbf{n} \in \boldsymbol{\mathcal{M}}^{\perp}$, $\mathbf{Pn} = \mathbf{0}$.

The null space of $\mathbf{P}$ is the orthogonal complement of $\boldsymbol{\mathcal{M}}$

For $\mathbb{R}^N$ with the inner product, $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^T\mathbf{y}$, $\mathbf{P}$ is an orthogonal projection onto $\boldsymbol{\mathcal{M}}$ if $\mathbf{P}$ is a projection ($\mathbf{P}^2 = \mathbf{P}$) and it is symmetric ($\mathbf{P} = \mathbf{P}^T$)

Exercise

Show that $\mathbf{P}_{\mathbf{X}}$ is an orthogonal projection on $C(\mathbf{X})$.

Decompsition

- For any $\mathbf{y} \in \mathbb{R}^n$, we can write it uniquely as a vector

$$\mathbf{y} = \mathbf{m} + \mathbf{n}, \quad \mathbf{m} \in \boldsymbol{\mathcal{M}} \quad \mathbf{n} \in \boldsymbol{\mathcal{M}}^{\perp}$$

- write $\mathbf{y} = \mathbf{P}\mathbf{y} + (\mathbf{y} - \mathbf{P}\mathbf{y}) = \mathbf{P}\mathbf{y} + (\mathbf{I} - \mathbf{P})\mathbf{y}$
- claim that if $\mathbf{P}$ is an orthogonal projection, $(\mathbf{I} - \mathbf{P})$ is an orthogonal projection onto $\boldsymbol{\mathcal{M}}^{\perp}$
- if $\mathbf{n} \in \boldsymbol{\mathcal{M}}^{\perp}$, then $(\mathbf{I} - \mathbf{P})\mathbf{n} = \mathbf{n} - \mathbf{P}\mathbf{n} = \mathbf{n}$

Back to MLEs

- $\mathbf{Y} \sim \mathsf{N}(\boldsymbol{\mu}, \sigma^2 \mathbf{I}_n)$ with $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$ and $\mathbf{X}$ full column rank
- Claim: Maximum Likelihood Estimator (MLE) of $\boldsymbol{\mu}$ is $\mathbf{P}_{\mathbf{X}}\mathbf{Y}$
- Log Likelihood:

$$\log \mathcal{L}(\boldsymbol{\mu}, \sigma^2) = -\frac{n}{2}\log(\sigma^2) - \frac{1}{2}\frac{\|\mathbf{Y} - \boldsymbol{\mu}\|^2}{\sigma^2}$$

- Decompose $\mathbf{Y} = \mathbf{P}_{\mathbf{X}}\mathbf{Y} + (\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{Y}$
- Use $\mathbf{P}_{\mathbf{X}}\boldsymbol{\mu} = \boldsymbol{\mu}$
- Simplify $\|\mathbf{Y} - \boldsymbol{\mu}\|^2$

Expand

$$
\begin{aligned}
\|\mathbf{Y} - \boldsymbol{\mu}\|^2 &= \|(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{Y} + \mathbf{P}_x\mathbf{Y} - \boldsymbol{\mu}\|^2 \\
&= \|(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{Y} + \mathbf{P}_x\mathbf{Y} - \mathbf{P}_{\mathbf{X}}\boldsymbol{\mu}\|^2 \\
&= \|(\mathbf{I} - \mathbf{P}_{\mathbf{x}})\mathbf{Y} + \mathbf{P}_{\mathbf{X}}(\mathbf{Y} - \boldsymbol{\mu})\|^2 \\
&= \|(\mathbf{I} - \mathbf{P}_{\mathbf{x}})\mathbf{Y}\|^2 + \|\mathbf{P}_{\mathbf{X}}(\mathbf{Y} - \boldsymbol{\mu})\|^2 + 2(\mathbf{Y} - \boldsymbol{\mu})^T\mathbf{P}_{\mathbf{X}}^T(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{Y} \\
&= \|(\mathbf{I} - \mathbf{P}_{\mathbf{x}})\mathbf{Y}\|^2 + \|\mathbf{P}_{\mathbf{X}}(\mathbf{Y} - \boldsymbol{\mu})\|^2 + 0 \\
&= \|(\mathbf{I} - \mathbf{P}_{\mathbf{x}})\mathbf{Y}\|^2 + \|\mathbf{P}_{\mathbf{X}}\mathbf{Y} - \boldsymbol{\mu}\|^2
\end{aligned}
$$

Crossproduct term is zero:

$$
\begin{aligned}
\mathbf{P}_{\mathbf{X}}^T(\mathbf{I} - \mathbf{P}_{\mathbf{X}}) &= \mathbf{P}_{\mathbf{X}}(\mathbf{I} - \mathbf{P}_{\mathbf{X}}) \\
&= \mathbf{P}_{\mathbf{X}} - \mathbf{P}_{\mathbf{X}}\mathbf{P}_{\mathbf{X}} \\
&= \mathbf{P}_{\mathbf{X}} - \mathbf{P}_{\mathbf{X}} \\
&= \mathbf{0}
\end{aligned}
$$

Log Likelihood

Substitute decomposition into log likelihood

$$
\begin{aligned}
\log \mathcal{L}(\boldsymbol{\mu}, \sigma^2) &= -\frac{n}{2}\log(\sigma^2) - \frac{1}{2}\frac{\|\mathbf{Y} - \boldsymbol{\mu}\|^2}{\sigma^2} \\
&= -\frac{n}{2}\log(\sigma^2) - \frac{1}{2}\left(\frac{\|(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{Y}\|^2}{\sigma^2} + \frac{\|\mathbf{P}_{\mathbf{X}}\mathbf{Y} - \boldsymbol{\mu}\|^2}{\sigma^2}\right) \\
&= \underbrace{-\frac{n}{2}\log(\sigma^2) - \frac{1}{2}\frac{\|(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{Y}\|^2}{\sigma^2}}_{} + \underbrace{-\frac{1}{2}\frac{\|\mathbf{P}_{\mathbf{X}}\mathbf{Y} - \boldsymbol{\mu}\|^2}{\sigma^2}}_{} \\
&= \text{constant with respect to } \boldsymbol{\mu} \qquad \leq 0
\end{aligned}
$$

- Maximize with respect to $\boldsymbol{\mu}$ for each σ^2
- RHS is largest when $\boldsymbol{\mu} = \mathbf{P}_{\mathbf{X}}\mathbf{Y}$ for any choice of σ^2

$$
\therefore \quad \hat{\boldsymbol{\mu}} = \mathbf{P}_{\mathbf{X}}\mathbf{Y}
$$

is the MLE of $\boldsymbol{\mu}$ (fitted values $\hat{\mathbf{Y}} = \mathbf{P}_{\mathbf{X}}\mathbf{Y}$)

MLE of β

$$\mathcal{L}(\boldsymbol{\mu}, \sigma^2) = -\frac{n}{2}\log(\sigma^2) - \frac{1}{2}\left(\frac{\|(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{Y}\|^2}{\sigma^2} + \frac{\|\mathbf{P}_{\mathbf{X}}\mathbf{Y} - \boldsymbol{\mu}\|^2}{\sigma^2}\right)$$

Rewrite as likeloood function for $\boldsymbol{\beta}, \sigma^2$:

$$\mathcal{L}(\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2}\log(\sigma^2) - \frac{1}{2}\left(\frac{\|(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{Y}\|^2}{\sigma^2} + \frac{\|\mathbf{P}_{\mathbf{X}}\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2}{\sigma^2}\right)$$

- Similar argument to show that RHS is maximized by minimizing

$$\|\mathbf{P}_{\mathbf{X}}\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2$$

- Therefore $\hat{\boldsymbol{\beta}}$ is a MLE of $\boldsymbol{\beta}$ if and only if satisfies

$$\mathbf{P}_{\mathbf{X}}\mathbf{Y} = \mathbf{X}\hat{\boldsymbol{\beta}}$$

- If $\mathbf{X}^T\mathbf{X}$ is full rank, the MLE of $\boldsymbol{\beta}$ is $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = \hat{\boldsymbol{\beta}}$

MLE of σ^2

- Plug-in MLE of $\hat{\boldsymbol{\mu}}$ for $\boldsymbol{\mu}$

$$\log \mathcal{L}(\hat{\boldsymbol{\mu}}, \sigma^2) = -\frac{n}{2} \log \sigma^2 - \frac{1}{2} \frac{\|(\mathbf{I} - \mathbf{P_X})\mathbf{Y}\|^2}{\sigma^2}$$

- Differentiate with respect to σ^2

$$\frac{\partial \log \mathcal{L}(\hat{\boldsymbol{\mu}}, \sigma^2)}{\partial \sigma^2} = -\frac{n}{2} \frac{1}{\sigma^2} + \frac{1}{2} \|(\mathbf{I} - \mathbf{P_X})\mathbf{Y}\|^2 \left(\frac{1}{\sigma^2}\right)^2$$

- Set derivative to zero and solve for MLE

$$\begin{aligned} 0 &= -\frac{n}{2} \frac{1}{\hat{\sigma}^2} + \frac{1}{2} \|(\mathbf{I} - \mathbf{P_X})\mathbf{Y}\|^2 \left(\frac{1}{\hat{\sigma}^2}\right)^2 \\ \frac{n}{2} \hat{\sigma}^2 &= \frac{1}{2} \|(\mathbf{I} - \mathbf{P_X})\mathbf{Y}\|^2 \\ \hat{\sigma}^2 &= \frac{\|(\mathbf{I} - \mathbf{P_X})\mathbf{Y}\|^2}{n} \end{aligned}$$

MLE Estimate of σ^2

Maximum Likelihood Estimate of σ^2

$$
\begin{aligned}
\hat{\sigma}^2 &= \frac{\|(\mathbf{I} - \mathbf{P_X})\mathbf{Y}\|^2}{n} \\
&= \frac{\mathbf{Y}^T(\mathbf{I} - \mathbf{P_X})^T(\mathbf{I} - \mathbf{P_X})\mathbf{Y}}{n} \\
&= \frac{\mathbf{Y}^T(\mathbf{I} - \mathbf{P_X})\mathbf{Y}}{n} \\
&= \frac{\mathbf{e}^T\mathbf{e}}{n}
\end{aligned}
$$

where $\mathbf{e} = (\mathbf{I} - \mathbf{P_X})\mathbf{Y}$ are the *residuals* from the regression of $\mathbf{Y}$ on $\mathbf{X}$