Statistical Methods for Analysis with Missing Data

Lecture 5: likelihood-based methods

Mauricio Sadinle

Department of Biostatistics

UNIVERSITY of WASHINGTON

Previous Lectures

Naïve/ad-hoc approaches to handling missing data:

- Complete-case analyses are wasteful and potentially invalid unless MCAR holds
- Imputation methods might be valid for some quantities under MCAR, but:
  - Variances are underestimated $\Longrightarrow$ overconfidence in your results!
  - Invalid results for other quantities, induced biases are not clear!
- R session 1:
  - Simulation study showed mean imputation leads to:
    - Invalid inferences on regression coefficients
    - Underestimation of variances
  - R package VIM implements variants of hot-deck imputation
  - Open question: performance of bootstrap + imputation?

Today's Lecture

Likelihood-based approaches

- General set-up for maximum likelihood estimation
- How did Rubin come up with the MAR assumption?
- The concept of *ignorability*

Reading: pages 50 – 61, Ch. 3, of Davidian and Tsiatis

Outline

Parametric Models

- $Z = (Z_1, \dots, Z_K)$: generic vector of study variables
- Thus far we have written $p(z)$ to represent the probability density function of the distribution of Z
- We now work under a *parametric model* for the distribution of Z

  $$\{p(z \mid \theta)\}_\theta,$$

  with $\theta = (\theta_1, \theta_2, \dots, \theta_d)$
- Model written as $\{p(z; \theta)\}_\theta$ in Davidian and Tsiatis (philosophical difference)

Example of Parametric Model: Bivariate Normal

Suppose that $Y = (Y_1, Y_2)^T$ is bivariate normal

$$Y \sim \mathcal{N}(\mu, \Sigma)\ ,\ \mu = (\mu_1,\ \mu_2)^T,\ \Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}.$$

The probability density of Y is

$$p(y \mid \theta) = \frac{1}{2\pi|\Sigma|^{1/2}} \exp\{-(y-\mu)^T \Sigma^{-1}(y-\mu)/2\},$$

where $\theta = (\mu_1,\ \sigma_1^2,\ \mu_2,\ \sigma_2^2,\ \sigma_{12})^T$.

Our Typical, Idealized Sampling Process

- In practice, we have data z_i, for each $i = 1, \ldots, n$
- We imagine that $z_i = (z_{i1}, \ldots, z_{iK})$ is a realization of a random vector $Z_i = (Z_{i1}, \ldots, Z_{iK})$
- All random vectors $\{Z_i\}_{i=1}^n$ follow the same distribution and are independent of each other – *independent and identically distributed* (i.i.d. or IID)
- Under our parametric model, the joint distribution of $\{Z_i\}_{i=1}^n$ has a density function

$$\prod_{i=1}^n p(z_i \mid \theta)$$

Maximum Likelihood Estimation

- The likelihood function is defined as

$$L(\theta) = \prod_{i=1}^{n} p(z_i \mid \theta),$$

seen as a function of θ

- The maximum likelihood estimator (MLE) is the value $\hat{\theta}$ that maximizes $L(\theta)$

$$\hat{\theta} = \arg\max_{\theta} L(\theta) = \arg\max_{\theta} \log L(\theta)$$

- We take the log because it is usually easier to work with

$$\log L(\theta) = \sum_{i=1}^{n} \log p(z_i \mid \theta)$$

and it leads to the same maximizer

Finding the MLE

- Under some *regularity conditions*, the MLE is the solution to the score equations

$$\sum_{i=1}^{n} S_\theta(z_i;\theta) = \sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log p(z_i \mid \theta) = \mathbf{0}$$

- Where the score vector

$$S_\theta(z;\theta) = \frac{\partial}{\partial\theta} \log p(z \mid \theta) = \begin{pmatrix} \frac{\partial}{\partial\theta_1} \log p(z \mid \theta) \\ \frac{\partial}{\partial\theta_2} \log p(z \mid \theta) \\ \vdots \\ \frac{\partial}{\partial\theta_d} \log p(z \mid \theta) \end{pmatrix}$$

- Solving the score equations might require iterative methods, such as Newton–Raphson

Why MLEs?

Under regularity conditions, including that the model is *correctly specified*, i.e., there really exists θ_0 such that $p(z \mid \theta_0)$ is the true density:

- The MLE is a consistent estimator: $\hat{\theta} \xrightarrow{p} \theta_0$
- We know the MLE's asymptotic distribution:

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(\mathbf{0}, \mathcal{I}(\theta_0)^{-1}),$$

  where $\mathcal{I}(\theta)$ is Fisher's information matrix

$$\mathcal{I}(\theta) = -E\left[\frac{\partial^2}{\partial\theta\partial\theta^T}\log p(Z \mid \theta)\right] = E\left[S_\theta(Z;\theta)S_\theta(Z;\theta)^T\right]$$

- $\mathcal{I}(\theta_0)$ is unknown, but $\mathcal{I}(\hat{\theta}) \xrightarrow{p} \mathcal{I}(\theta_0)$
- Heuristically, we say

$$\hat{\theta} \approx \mathcal{N}(\theta_0, \mathcal{I}(\hat{\theta})^{-1}/n)$$

Why MLEs?

Under regularity conditions, including that the model is *correctly specified*, i.e., there really exists θ_0 such that $p(z \mid \theta_0)$ is the true density:

- The MLE is a consistent estimator: $\hat{\theta} \xrightarrow{p} \theta_0$
- We know the MLE's asymptotic distribution:

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(\mathbf{0}, \mathcal{I}(\theta_0)^{-1}),$$

where $\mathcal{I}(\theta)$ is Fisher's information matrix

$$\mathcal{I}(\theta) = -E\left[\frac{\partial^2}{\partial\theta\partial\theta^T} \log p(Z \mid \theta)\right] = E\left[S_\theta(Z;\theta)S_\theta(Z;\theta)^T\right]$$

- $\mathcal{I}(\theta_0)$ is unknown, but $\mathcal{I}(\hat{\theta}) \xrightarrow{p} \mathcal{I}(\theta_0)$
- Heuristically, we say

$$\hat{\theta} \approx \mathcal{N}(\theta_0, \mathcal{I}(\hat{\theta})^{-1}/n)$$

Why MLEs?

Under regularity conditions, including that the model is *correctly specified*, i.e., there really exists θ_0 such that $p(z \mid \theta_0)$ is the true density:

- The MLE is a consistent estimator: $\hat{\theta} \xrightarrow{p} \theta_0$
- We know the MLE's asymptotic distribution:

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(\mathbf{0}, \mathcal{I}(\theta_0)^{-1}),$$

  where $\mathcal{I}(\theta)$ is Fisher's information matrix

$$\mathcal{I}(\theta) = -E\left[\frac{\partial^2}{\partial\theta\partial\theta^T} \log p(Z \mid \theta)\right] = E\left[S_\theta(Z;\theta)S_\theta(Z;\theta)^T\right]$$

- $\mathcal{I}(\theta_0)$ is unknown, but $\mathcal{I}(\hat{\theta}) \xrightarrow{p} \mathcal{I}(\theta_0)$
- Heuristically, we say

$$\hat{\theta} \approx \mathcal{N}(\theta_0, \mathcal{I}(\hat{\theta})^{-1}/n)$$

Why MLEs?

Under regularity conditions, including that the model is *correctly specified*, i.e., there really exists θ_0 such that $p(z \mid \theta_0)$ is the true density:

- The MLE is a consistent estimator: $\hat{\theta} \xrightarrow{p} \theta_0$
- We know the MLE's asymptotic distribution:

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(\mathbf{0}, \mathcal{I}(\theta_0)^{-1}),$$

  where $\mathcal{I}(\theta)$ is Fisher's information matrix

$$\mathcal{I}(\theta) = -E\left[\frac{\partial^2}{\partial\theta\partial\theta^T}\log p(Z \mid \theta)\right] = E\left[S_\theta(Z;\theta)S_\theta(Z;\theta)^T\right]$$

- $\mathcal{I}(\theta_0)$ is unknown, but $\mathcal{I}(\hat{\theta}) \xrightarrow{p} \mathcal{I}(\theta_0)$
- Heuristically, we say

$$\hat{\theta} \approx \mathcal{N}(\theta_0, \mathcal{I}(\hat{\theta})^{-1}/n)$$

Why MLEs?

Under regularity conditions, including that the model is *correctly specified*, i.e., there really exists θ_0 such that $p(z \mid \theta_0)$ is the true density:

- The MLE is a consistent estimator: $\hat{\theta} \xrightarrow{p} \theta_0$
- We know the MLE's asymptotic distribution:

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(\mathbf{0}, \mathcal{I}(\theta_0)^{-1}),$$

  where $\mathcal{I}(\theta)$ is Fisher's information matrix

$$\mathcal{I}(\theta) = -E\left[\frac{\partial^2}{\partial\theta\partial\theta^T} \log p(Z \mid \theta)\right] = E\left[S_\theta(Z;\theta)S_\theta(Z;\theta)^T\right]$$

- $\mathcal{I}(\theta_0)$ is unknown, but $\mathcal{I}(\hat{\theta}) \xrightarrow{p} \mathcal{I}(\theta_0)$
- Heuristically, we say

$$\hat{\theta} \approx \mathcal{N}(\theta_0, \mathcal{I}(\hat{\theta})^{-1}/n)$$

Why MLEs?

- Sometimes, computing $\mathcal{I}(\theta)$ can be complicated, so we might instead use the observed information matrix

$$J(\theta) = -\sum_{i=1}^{n} \frac{\partial^2}{\partial\theta\partial\theta^T} \log p(z_i \mid \theta)$$

- We have that $n^{-1} J(\hat{\theta}) \xrightarrow{p} \mathcal{I}(\theta_0)$
- Therefore, we heuristically say

$$\hat{\theta} \approx \mathcal{N}(\theta_0, J(\hat{\theta})^{-1})$$

- This can be used for approximating standard errors for the components of θ and to compute approximately valid confidence intervals
- What if we have missing data? Our observed data are realizations of $(Z_{(R)}, R)$, not realizations of Z!

Why MLEs?

- Sometimes, computing $\mathcal{I}(\theta)$ can be complicated, so we might instead use the observed information matrix

$$J(\theta) = -\sum_{i=1}^{n} \frac{\partial^2}{\partial\theta\partial\theta^T} \log p(z_i \mid \theta)$$

- We have that $n^{-1}J(\hat{\theta}) \xrightarrow{p} \mathcal{I}(\theta_0)$
- Therefore, we heuristically say

$$\hat{\theta} \approx \mathcal{N}(\theta_0, J(\hat{\theta})^{-1})$$

- This can be used for approximating standard errors for the components of θ and to compute approximately valid confidence intervals
- What if we have missing data? Our observed data are realizations of $(Z_{(R)}, R)$, not realizations of Z!

Why MLEs?

- Sometimes, computing $\mathcal{I}(\theta)$ can be complicated, so we might instead use the observed information matrix

$$J(\theta) = -\sum_{i=1}^{n} \frac{\partial^2}{\partial\theta\partial\theta^T} \log p(z_i \mid \theta)$$

- We have that $n^{-1}J(\hat{\theta}) \xrightarrow{p} \mathcal{I}(\theta_0)$

- Therefore, we heuristically say

$$\hat{\theta} \approx \mathcal{N}(\theta_0, J(\hat{\theta})^{-1})$$

- This can be used for approximating standard errors for the components of θ and to compute approximately valid confidence intervals

- What if we have missing data? Our observed data are realizations of $(Z_{(R)}, R)$, not realizations of Z!

Why MLEs?

- Sometimes, computing $\mathcal{I}(\theta)$ can be complicated, so we might instead use the observed information matrix

$$J(\theta) = -\sum_{i=1}^{n} \frac{\partial^2}{\partial\theta\partial\theta^T} \log p(z_i \mid \theta)$$

- We have that $n^{-1}J(\hat{\theta}) \xrightarrow{p} \mathcal{I}(\theta_0)$
- Therefore, we heuristically say

$$\hat{\theta} \approx \mathcal{N}(\theta_0, J(\hat{\theta})^{-1})$$

- This can be used for approximating standard errors for the components of θ and to compute approximately valid confidence intervals
- What if we have missing data? Our observed data are realizations of $(Z_{(R)}, R)$, not realizations of Z!

Why MLEs?

- Sometimes, computing $\mathcal{I}(\theta)$ can be complicated, so we might instead use the observed information matrix

$$J(\theta) = -\sum_{i=1}^{n} \frac{\partial^2}{\partial\theta\partial\theta^T} \log p(z_i \mid \theta)$$

- We have that $n^{-1}J(\hat{\theta}) \xrightarrow{p} \mathcal{I}(\theta_0)$
- Therefore, we heuristically say

$$\hat{\theta} \approx \mathcal{N}(\theta_0, J(\hat{\theta})^{-1})$$

- This can be used for approximating standard errors for the components of θ and to compute approximately valid confidence intervals
- What if we have missing data? Our observed data are realizations of $(Z_{(R)}, R)$, not realizations of Z!

Outline

Factorizations of the Full-Data Distribution

Full-data distribution: joint distribution of (Z, R), with density

$$p(z, r)$$

Not accessible to us, mere humans, even with infinite samples, but we know it can be factorized in different ways

- Selection model factorization:

$$p(z, r) = p(r \mid z)p(z)$$

  - $p(z)$ can come from the parametric model we would use if we had complete data, say $p(z \mid \theta)$
  - $p(r \mid z)$ can come from a model for the response mechanism, $p(r \mid z, \psi)$
- Other factorizations are important and lead to alternative approaches for handling missing data, but they will be covered later in the course

Factorizations of the Full-Data Distribution

Full-data distribution: joint distribution of (Z, R), with density

$$p(z, r)$$

Not accessible to us, mere humans, even with infinite samples, but we know it can be factorized in different ways

- Selection model factorization:

$$p(z, r) = p(r \mid z)p(z)$$

  - $p(z)$ can come from the parametric model we would use if we had complete data, say $p(z \mid \theta)$
  - $p(r \mid z)$ can come from a model for the response mechanism, $p(r \mid z, \psi)$
- Other factorizations are important and lead to alternative approaches for handling missing data, but they will be covered later in the course

Factorizations of the Full-Data Distribution

Full-data distribution: joint distribution of (Z, R), with density

$$p(z, r)$$

Not accessible to us, mere humans, even with infinite samples, but we know it can be factorized in different ways

- Selection model factorization:

$$p(z, r) = p(r \mid z)p(z)$$

  - $p(z)$ can come from the parametric model we would use if we had complete data, say $p(z \mid \theta)$
  - $p(r \mid z)$ can come from a model for the response mechanism, $p(r \mid z, \psi)$
- Other factorizations are important and lead to alternative approaches for handling missing data, but they will be covered later in the course

Parametric Models

- Consider a parametric family for the marginal distribution of Z

$$\{p(z \mid \theta)\}_\theta,$$

  and for the response mechanism

$$\{p(r \mid z, \psi)\}_\psi$$

- We assume *separability* of θ and ψ: knowledge on the value of θ says nothing about the value of ψ, and vice versa
  - All combinations of values of θ and ψ are possible
  - The range of values of θ is the same regardless of ψ, and vice versa

Parametric Models

- Consider a parametric family for the marginal distribution of Z

$$\{p(z \mid \theta)\}_\theta,$$

  and for the response mechanism

$$\{p(r \mid z, \psi)\}_\psi$$

- We assume *separability* of θ and ψ: knowledge on the value of θ says nothing about the value of ψ, and vice versa
  - All combinations of values of θ and ψ are possible
  - The range of values of θ is the same regardless of ψ, and vice versa

Full-Data Sample

In the full-data world:

- Study variables for individual i: $Z_i = (Z_{i1}, \ldots, Z_{iK})$
- Response indicators for individual i: $R_i = (R_{i1}, \ldots, R_{iK})$
- $\{(Z_i, R_i)\}_{i=1}^n$ are independent and identically distributed
- The realized values are $\{(z_i, r_i)\}_{i=1}^n$
- This leads to a *full-data likelihood* function

$$L_{full}(\theta, \psi) = \prod_{i=1}^n p(r_i \mid z_i, \psi) p(z_i \mid \theta)$$

Clearly, we cannot work with $L_{full}(\theta, \psi)$, as it depends on missing data!

Full-Data Sample

In the full-data world:

- Study variables for individual i: $Z_i = (Z_{i1}, \ldots, Z_{iK})$
- Response indicators for individual i: $R_i = (R_{i1}, \ldots, R_{iK})$
- $\{(Z_i, R_i)\}_{i=1}^n$ are independent and identically distributed
- The realized values are $\{(z_i, r_i)\}_{i=1}^n$
- This leads to a *full-data likelihood* function

$$L_{full}(\theta, \psi) = \prod_{i=1}^{n} p(r_i \mid z_i, \psi) p(z_i \mid \theta)$$

Clearly, we cannot work with $L_{full}(\theta, \psi)$, as it depends on missing data!

Full-Data Sample

In the full-data world:

- Study variables for individual i: $Z_i = (Z_{i1}, \ldots, Z_{iK})$
- Response indicators for individual i: $R_i = (R_{i1}, \ldots, R_{iK})$
- $\{(Z_i, R_i)\}_{i=1}^n$ are independent and identically distributed
- The realized values are $\{(z_i, r_i)\}_{i=1}^n$
- This leads to a *full-data likelihood* function

$$L_{full}(\theta, \psi) = \prod_{i=1}^{n} p(r_i \mid z_i, \psi) p(z_i \mid \theta)$$

Clearly, we cannot work with $L_{full}(\theta, \psi)$, as it depends on missing data!

The Observed-Data Distribution

As mentioned in Lecture 2, given that R is random, the *observed data* are obtained as realizations of

$$(Z_{(R)}, R)$$

We can think of the generative process

$$Z \implies R \implies (Z_{(R)}, R)$$

The distribution of $(Z_{(R)}, R)$ is referred to as the *observed-data distribution*, and it has a probability density denoted by

$$p(z_{(r)}, r)$$

The Observed-Data Distribution

As mentioned in Lecture 2, given that R is random, the *observed data* are obtained as realizations of

$$(Z_{(R)}, R)$$

We can think of the generative process

$$Z \implies R \implies (Z_{(R)}, R)$$

The distribution of $(Z_{(R)}, R)$ is referred to as the *observed-data distribution*, and it has a probability density denoted by

$$p(z_{(r)}, r)$$

The Observed-Data Distribution

As mentioned in Lecture 2, given that R is random, the *observed data* are obtained as realizations of

$$(Z_{(R)}, R)$$

We can think of the generative process

$$Z \implies R \implies (Z_{(R)}, R)$$

The distribution of $(Z_{(R)}, R)$ is referred to as the *observed-data distribution*, and it has a probability density denoted by

$$p(z_{(r)}, r)$$

The Observed-Data Distribution

To derive $p(z_{(r)}, r)$, we need to integrate $p(z, r)$ over the possible missing values $z_{(\bar{r})}$, denoted $\mathcal{Z}_{(\bar{r})}$

$$
\begin{aligned}
p(z_{(r)}, r) &= \int_{\mathcal{Z}_{(\bar{r})}} p(z, r)\ \mu(dz_{(\bar{r})}) \\
&= \int_{\mathcal{Z}_{(\bar{r})}} p(r \mid z) p(z)\ \mu(dz_{(\bar{r})}) \\
&= \begin{cases} \int_{\mathcal{Z}_{(\bar{r})}} p(r \mid z, \psi) p(z \mid \theta) dz_{(\bar{r})} & \text{if } Z \text{ is continuous} \\ \sum_{\mathcal{Z}_{(\bar{r})}} p(r \mid z, \psi) p(z \mid \theta) & \text{if } Z \text{ is discrete} \end{cases}
\end{aligned}
$$

From now on, we'll write $\int_{\mathcal{Z}_{(\bar{r})}} p(z, r) dz_{(\bar{r})}$ instead of $\int_{\mathcal{Z}_{(\bar{r})}} p(z, r) \mu(dz_{(\bar{r})})$

The Observed-Data Distribution

To derive $p(z_{(r)}, r)$, we need to integrate $p(z, r)$ over the possible missing values $z_{(\bar{r})}$, denoted $\mathcal{Z}_{(\bar{r})}$

$$
\begin{aligned}
p(z_{(r)}, r) &= \int_{\mathcal{Z}_{(\bar{r})}} p(z, r)\ \mu(dz_{(\bar{r})}) \\
&= \int_{\mathcal{Z}_{(\bar{r})}} p(r \mid z) p(z)\ \mu(dz_{(\bar{r})}) \\
&= \begin{cases} \int_{\mathcal{Z}_{(\bar{r})}} p(r \mid z, \psi) p(z \mid \theta) dz_{(\bar{r})} & \text{if } Z \text{ is continuous} \\ \sum_{\mathcal{Z}_{(\bar{r})}} p(r \mid z, \psi) p(z \mid \theta) & \text{if } Z \text{ is discrete} \end{cases}
\end{aligned}
$$

From now on, we'll write $\int_{\mathcal{Z}_{(\bar{r})}} p(z, r) dz_{(\bar{r})}$ instead of $\int_{\mathcal{Z}_{(\bar{r})}} p(z, r) \mu(dz_{(\bar{r})})$

The Observed-Data Distribution

To derive $p(z_{(r)}, r)$, we need to integrate $p(z, r)$ over the possible missing values $z_{(\bar{r})}$, denoted $\mathcal{Z}_{(\bar{r})}$

$$\begin{aligned} p(z_{(r)}, r) &= \int_{\mathcal{Z}_{(\bar{r})}} p(z, r)\ \mu(dz_{(\bar{r})}) \\ &= \int_{\mathcal{Z}_{(\bar{r})}} p(r \mid z) p(z)\ \mu(dz_{(\bar{r})}) \\ &= \begin{cases} \int_{\mathcal{Z}_{(\bar{r})}} p(r \mid z, \psi) p(z \mid \theta) dz_{(\bar{r})} & \text{if } Z \text{ is continuous} \\ \sum_{\mathcal{Z}_{(\bar{r})}} p(r \mid z, \psi) p(z \mid \theta) & \text{if } Z \text{ is discrete} \end{cases} \end{aligned}$$

From now on, we'll write $\int_{\mathcal{Z}_{(\bar{r})}} p(z, r) dz_{(\bar{r})}$ instead of $\int_{\mathcal{Z}_{(\bar{r})}} p(z, r) \mu(dz_{(\bar{r})})$

The Observed-Data Distribution

To derive $p(z_{(r)}, r)$, we need to integrate $p(z, r)$ over the possible missing values $z_{(\bar{r})}$, denoted $\mathcal{Z}_{(\bar{r})}$

$$
\begin{aligned}
p(z_{(r)}, r) &= \int_{\mathcal{Z}_{(\bar{r})}} p(z, r)\ \mu(dz_{(\bar{r})}) \\
&= \int_{\mathcal{Z}_{(\bar{r})}} p(r \mid z) p(z)\ \mu(dz_{(\bar{r})}) \\
&= \begin{cases} \int_{\mathcal{Z}_{(\bar{r})}} p(r \mid z, \psi) p(z \mid \theta) dz_{(\bar{r})} & \text{if } Z \text{ is continuous} \\ \sum_{\mathcal{Z}_{(\bar{r})}} p(r \mid z, \psi) p(z \mid \theta) & \text{if } Z \text{ is discrete} \end{cases}
\end{aligned}
$$

From now on, we'll write $\int_{\mathcal{Z}_{(\bar{r})}} p(z, r) dz_{(\bar{r})}$ instead of $\int_{\mathcal{Z}_{(\bar{r})}} p(z, r) \mu(dz_{(\bar{r})})$

Example of Observed-Data Distribution

HW2: problem 6 of HW1 continued: say $K = 2$, $Z_1 \in \{1, 2\}$, $Z_2 \in \{A, B\}$, $R \in \{0, 1\}^2$.

- Write down all the elements of the sample space of $(Z_{(R)}, R)$
- Say the full-data probability density is given by

$$p(z, r) \equiv p(z_1, z_2, r_1, r_2) \equiv \pi_{z_1, z_2, r_1, r_2}$$

  Derive $p(z_{(r)}, r)$ for all elements $(z_{(r)}, r)$ in the sample space of $(Z_{(R)}, R)$

Example of Observed-Data Distribution

HW2: say $K = 2$, $(Z_1, Z_2)^T \sim \mathcal{N}(\mu, \Sigma)$, $R \in \{0,1\}^2$.

- Describe the sample space of $(Z_{(R)}, R)$ (problem 7 of HW1)
- Say $p(r \mid z) = p(r)$. Derive $p(z_{(r)}, r)$ for all $r \in \{0,1\}^2$
- Say $R_1 \perp\!\!\!\perp R_2 \mid Z$,

$$\text{logit } p(R_j = 1 \mid z) = \beta_{j0} + \beta_{j1} z_1 + \beta_{j2} z_2, \quad j = 1, 2.$$

Derive $p(z_{(r)}, r)$ for all $r \in \{0,1\}^2$

Example of Observed-Data Distribution

HW2: say $K = 2$, $(Z_1, Z_2)^T \sim \mathcal{N}(\mu, \Sigma)$, $R \in \{0, 1\}^2$.

- Describe the sample space of $(Z_{(R)}, R)$ (problem 7 of HW1)
- Say $p(r \mid z) = p(r)$. Derive $p(z_{(r)}, r)$ for all $r \in \{0, 1\}^2$
- Say $R_1 \perp\!\!\!\perp R_2 \mid Z$,

$$\text{logit } p(R_j = 1 \mid z) = \beta_{j0} + \beta_{j1} z_1 + \beta_{j2} z_2, \quad j = 1, 2.$$

Derive $p(z_{(r)}, r)$ for all $r \in \{0, 1\}^2$

Likelihood-Based Set-Up

- The random sample we are actually working with is

$$\{(Z_{i(R_i)}, R_i)\}_{i=1}^n$$

- The realized values are actually

$$\{(z_{i(r_i)}, r_i)\}_{i=1}^n$$

- As before, we can think of the generative process, for each i:

$$Z_i \implies R_i \implies (Z_{i(R_i)}, R_i)$$

- What is the *observed-data likelihood* function?

- We need to integrate the full-data likelihood $L_{full}(\theta, \psi)$ over the possible values of each $z_{i(\bar{r}_i)}$

Likelihood-Based Set-Up

- The random sample we are actually working with is

$$\{(Z_{i(R_i)}, R_i)\}_{i=1}^{n}$$

- The realized values are actually

$$\{(z_{i(r_i)}, r_i)\}_{i=1}^{n}$$

- As before, we can think of the generative process, for each i:

$$Z_i \implies R_i \implies (Z_{i(R_i)}, R_i)$$

- What is the *observed-data likelihood* function?

- We need to integrate the full-data likelihood $L_{full}(\theta, \psi)$ over the possible values of each $z_{i(\bar{r}_i)}$

Likelihood-Based Set-Up

- The random sample we are actually working with is

$$\{(Z_{i(R_i)}, R_i)\}_{i=1}^n$$

- The realized values are actually

$$\{(z_{i(r_i)}, r_i)\}_{i=1}^n$$

- As before, we can think of the generative process, for each i:

$$Z_i \implies R_i \implies (Z_{i(R_i)}, R_i)$$

- What is the *observed-data likelihood* function?
- We need to integrate the full-data likelihood $L_{full}(\theta, \psi)$ over the possible values of each $z_{i(\bar{r}_i)}$

Likelihood-Based Set-Up

- The random sample we are actually working with is

$$\{(Z_{i(R_i)}, R_i)\}_{i=1}^{n}$$

- The realized values are actually

$$\{(z_{i(r_i)}, r_i)\}_{i=1}^{n}$$

- As before, we can think of the generative process, for each i:

$$Z_i \implies R_i \implies (Z_{i(R_i)}, R_i)$$

- What is the *observed-data likelihood* function?

- We need to integrate the full-data likelihood $L_{full}(\theta, \psi)$ over the possible values of each $z_{i(\bar{r}_i)}$

Likelihood-Based Set-Up

- The random sample we are actually working with is

$$\{(Z_{i(R_i)}, R_i)\}_{i=1}^n$$

- The realized values are actually

$$\{(z_{i(r_i)}, r_i)\}_{i=1}^n$$

- As before, we can think of the generative process, for each i:

$$Z_i \implies R_i \implies (Z_{i(R_i)}, R_i)$$

- What is the *observed-data likelihood* function?
- We need to integrate the full-data likelihood $L_{full}(\theta, \psi)$ over the possible values of each $z_{i(\bar{r}_i)}$

Likelihood-Based Set-Up

- Since we are assuming i.i.d. data, let's focus on a generic term of the full-data likelihood

$$\ell_{full}(\theta, \psi) = p(r \mid z, \psi)p(z \mid \theta)$$

  to facilitate the notation

- We cannot work with $\ell_{full}(\theta, \psi)$ since we don't observe a complete realization z, but rather $z_{(r)}$

- We need to integrate over the missing data to derive the *observed-data likelihood*

$$\ell_{obs}(\theta, \psi) = \int_{\mathcal{Z}_{(\bar{r})}} p(r \mid z, \psi)p(z \mid \theta) \; dz_{(\bar{r})}$$

- $\ell_{obs}(\theta, \psi)$ does not depend on missing data

- To obtain likelihood-based inferences on θ, it seems we need to pass through the specification of $p(r \mid z, \psi)$

- Typically, $p(r \mid z, \psi)$ is not of scientific interest so it can be seen as a nuisance

Likelihood-Based Set-Up

- Since we are assuming i.i.d. data, let's focus on a generic term of the full-data likelihood

$$\ell_{full}(\theta, \psi) = p(r \mid z, \psi) p(z \mid \theta)$$

  to facilitate the notation

- We cannot work with $\ell_{full}(\theta, \psi)$ since we don't observe a complete realization z, but rather $z_{(r)}$

- We need to integrate over the missing data to derive the *observed-data likelihood*

$$\ell_{obs}(\theta, \psi) = \int_{\mathcal{Z}_{(\bar{r})}} p(r \mid z, \psi) p(z \mid \theta) \, dz_{(\bar{r})}$$

- $\ell_{obs}(\theta, \psi)$ does not depend on missing data

- To obtain likelihood-based inferences on θ, it seems we need to pass through the specification of $p(r \mid z, \psi)$

- Typically, $p(r \mid z, \psi)$ is not of scientific interest so it can be seen as a nuisance

Likelihood-Based Set-Up

- Since we are assuming i.i.d. data, let's focus on a generic term of the full-data likelihood
$$\ell_{full}(\theta, \psi) = p(r \mid z, \psi) p(z \mid \theta)$$
to facilitate the notation

- We cannot work with $\ell_{full}(\theta, \psi)$ since we don't observe a complete realization z, but rather $z_{(r)}$

- We need to integrate over the missing data to derive the *observed-data likelihood*
$$\ell_{obs}(\theta, \psi) = \int_{\mathcal{Z}_{(\bar{r})}} p(r \mid z, \psi) p(z \mid \theta) \; dz_{(\bar{r})}$$

- $\ell_{obs}(\theta, \psi)$ does not depend on missing data

- To obtain likelihood-based inferences on θ, it seems we need to pass through the specification of $p(r \mid z, \psi)$

- Typically, $p(r \mid z, \psi)$ is not of scientific interest so it can be seen as a nuisance

Likelihood-Based Set-Up

- Since we are assuming i.i.d. data, let's focus on a generic term of the full-data likelihood
$$\ell_{full}(\theta, \psi) = p(r \mid z, \psi)p(z \mid \theta)$$
to facilitate the notation

- We cannot work with $\ell_{full}(\theta, \psi)$ since we don't observe a complete realization z, but rather $z_{(r)}$

- We need to integrate over the missing data to derive the *observed-data likelihood*
$$\ell_{obs}(\theta, \psi) = \int_{\mathcal{Z}_{(\bar{r})}} p(r \mid z, \psi)p(z \mid \theta)\ dz_{(\bar{r})}$$

- $\ell_{obs}(\theta, \psi)$ does not depend on missing data

- To obtain likelihood-based inferences on θ, it seems we need to pass through the specification of $p(r \mid z, \psi)$

- Typically, $p(r \mid z, \psi)$ is not of scientific interest so it can be seen as a nuisance

Likelihood-Based Set-Up

- Since we are assuming i.i.d. data, let's focus on a generic term of the full-data likelihood
$$\ell_{full}(\theta, \psi) = p(r \mid z, \psi) p(z \mid \theta)$$
to facilitate the notation

- We cannot work with $\ell_{full}(\theta, \psi)$ since we don't observe a complete realization z, but rather $z_{(r)}$

- We need to integrate over the missing data to derive the *observed-data likelihood*
$$\ell_{obs}(\theta, \psi) = \int_{\mathcal{Z}_{(\bar{r})}} p(r \mid z, \psi) p(z \mid \theta) \; dz_{(\bar{r})}$$

- $\ell_{obs}(\theta, \psi)$ does not depend on missing data

- To obtain likelihood-based inferences on θ, it seems we need to pass through the specification of $p(r \mid z, \psi)$

- Typically, $p(r \mid z, \psi)$ is not of scientific interest so it can be seen as a nuisance

Likelihood-Based Set-Up

- Since we are assuming i.i.d. data, let's focus on a generic term of the full-data likelihood
$$\ell_{full}(\theta, \psi) = p(r \mid z, \psi) p(z \mid \theta)$$
to facilitate the notation

- We cannot work with $\ell_{full}(\theta, \psi)$ since we don't observe a complete realization z, but rather $z_{(r)}$

- We need to integrate over the missing data to derive the *observed-data likelihood*
$$\ell_{obs}(\theta, \psi) = \int_{\mathcal{Z}_{(\bar{r})}} p(r \mid z, \psi) p(z \mid \theta) \; dz_{(\bar{r})}$$

- $\ell_{obs}(\theta, \psi)$ does not depend on missing data

- To obtain likelihood-based inferences on θ, it seems we need to pass through the specification of $p(r \mid z, \psi)$

- Typically, $p(r \mid z, \psi)$ is not of scientific interest so it can be seen as a nuisance

Developing the Missing at Random (MAR) Assumption

Rubin's (1976, Biometrika) fundamental motivation:

- How can we get rid of this nuisance $p(r \mid z, \psi)$?
- When are inferences for θ based on $\int_{\mathcal{Z}_{(\bar{r})}} p(z \mid \theta)\ dz_{(\bar{r})}$ valid?

Stare at the observed-data likelihood:

$$\ell_{obs}(\theta, \psi) = \int_{\mathcal{Z}_{(\bar{r})}} p(r \mid z, \psi) p(z \mid \theta)\ dz_{(\bar{r})}$$

Developing the Missing at Random (MAR) Assumption

Rubin's (1976, Biometrika) fundamental motivation:

- How can we get rid of this nuisance $p(r \mid z, \psi)$?
- When are inferences for θ based on $\int_{\mathcal{Z}_{(\bar{r})}} p(z \mid \theta)\, dz_{(\bar{r})}$ valid?

Stare at the observed-data likelihood:

$$\ell_{obs}(\theta, \psi) = \int_{\mathcal{Z}_{(\bar{r})}} p(r \mid z, \psi) p(z \mid \theta)\, dz_{(\bar{r})}$$

Developing the Missing at Random (MAR) Assumption

Rubin's (1976, Biometrika) fundamental motivation:

- How can we get rid of this nuisance $p(r \mid z, \psi)$?
- When are inferences for θ based on $\int_{\mathcal{Z}_{(\bar{r})}} p(z \mid \theta)\, dz_{(\bar{r})}$ valid?

Stare at the observed-data likelihood:

$$\ell_{obs}(\theta, \psi) = \int_{\mathcal{Z}_{(\bar{r})}} p(r \mid z, \psi) p(z \mid \theta)\, dz_{(\bar{r})}$$

The Missing at Random (MAR) Assumption

The MAR assumption, in terms of $p(r \mid z, \psi)$ says

$$p(r \mid z, \psi) = p(r \mid z_{(r)}, \psi)$$

(we'll soon talk about the formal definition)

Ignorability Under MAR

- Under the MAR assumption:

$$\begin{aligned}\ell_{obs}(\theta, \psi) &= \int_{\mathcal{Z}_{(\bar{r})}} p(r \mid z, \psi) p(z \mid \theta)\ dz_{(\bar{r})} \\ &\overset{\text{MAR}}{=} \int_{\mathcal{Z}_{(\bar{r})}} p(r \mid z_{(r)}, \psi) p(z \mid \theta)\ dz_{(\bar{r})} \\ &= p(r \mid z_{(r)}, \psi) \int_{\mathcal{Z}_{(\bar{r})}} p(z \mid \theta)\ dz_{(\bar{r})} \\ &= p(r \mid z_{(r)}, \psi)\ p(z_{(r)} \mid \theta)\end{aligned}$$

- Under MAR, likelihood-based inference can be based on

$$\ell_{obs}(\theta) = p(z_{(r)} \mid \theta) = \int_{\mathcal{Z}_{(\bar{r})}} p(z \mid \theta)\ dz_{(\bar{r})}$$

- Missingness mechanism is *ignorable* since there's no need to specify $p(r \mid z, \psi)$ if we only care about θ

Ignorability Under MAR

- Under the MAR assumption:

$$\begin{aligned}
\ell_{obs}(\theta, \psi) &= \int_{\mathcal{Z}_{(\bar{r})}} p(r \mid z, \psi) p(z \mid \theta) \; dz_{(\bar{r})} \\
&\overset{\text{MAR}}{=} \int_{\mathcal{Z}_{(\bar{r})}} p(r \mid z_{(r)}, \psi) p(z \mid \theta) \; dz_{(\bar{r})} \\
&= p(r \mid z_{(r)}, \psi) \int_{\mathcal{Z}_{(\bar{r})}} p(z \mid \theta) \; dz_{(\bar{r})} \\
&= p(r \mid z_{(r)}, \psi) \; p(z_{(r)} \mid \theta)
\end{aligned}$$

- Under MAR, likelihood-based inference can be based on

$$\ell_{obs}(\theta) = p(z_{(r)} \mid \theta) = \int_{\mathcal{Z}_{(\bar{r})}} p(z \mid \theta) \; dz_{(\bar{r})}$$

- Missingness mechanism is *ignorable* since there's no need to specify $p(r \mid z, \psi)$ if we only care about θ

Ignorability Under MAR

- Under the MAR assumption:

$$
\begin{aligned}
\ell_{obs}(\theta, \psi) &= \int_{\mathcal{Z}_{(\bar{r})}} p(r \mid z, \psi) p(z \mid \theta) \; dz_{(\bar{r})} \\
&\overset{\text{MAR}}{=} \int_{\mathcal{Z}_{(\bar{r})}} p(r \mid z_{(r)}, \psi) p(z \mid \theta) \; dz_{(\bar{r})} \\
&= p(r \mid z_{(r)}, \psi) \int_{\mathcal{Z}_{(\bar{r})}} p(z \mid \theta) \; dz_{(\bar{r})} \\
&= p(r \mid z_{(r)}, \psi) \; p(z_{(r)} \mid \theta)
\end{aligned}
$$

- Under MAR, likelihood-based inference can be based on

$$
\ell_{obs}(\theta) = p(z_{(r)} \mid \theta) = \int_{\mathcal{Z}_{(\bar{r})}} p(z \mid \theta) \; dz_{(\bar{r})}
$$

- Missingness mechanism is *ignorable* since there's no need to specify $p(r \mid z, \psi)$ if we only care about θ

Ignorability Under MAR

- Under the MAR assumption:

$$\begin{aligned}
\ell_{obs}(\theta, \psi) &= \int_{\mathcal{Z}_{(\bar{r})}} p(r \mid z, \psi) p(z \mid \theta) \; dz_{(\bar{r})} \\
&\overset{\text{MAR}}{=} \int_{\mathcal{Z}_{(\bar{r})}} p(r \mid z_{(r)}, \psi) p(z \mid \theta) \; dz_{(\bar{r})} \\
&= p(r \mid z_{(r)}, \psi) \int_{\mathcal{Z}_{(\bar{r})}} p(z \mid \theta) \; dz_{(\bar{r})} \\
&= p(r \mid z_{(r)}, \psi) \; p(z_{(r)} \mid \theta)
\end{aligned}$$

- Under MAR, likelihood-based inference can be based on

$$\ell_{obs}(\theta) = p(z_{(r)} \mid \theta) = \int_{\mathcal{Z}_{(\bar{r})}} p(z \mid \theta) \; dz_{(\bar{r})}$$

- Missingness mechanism is *ignorable* since there's no need to specify $p(r \mid z, \psi)$ if we only care about θ

Ignorability Under MAR

- Under the MAR assumption:

$$\begin{aligned}
\ell_{obs}(\theta, \psi) &= \int_{\mathcal{Z}_{(\bar{r})}} p(r \mid z, \psi) p(z \mid \theta)\ dz_{(\bar{r})} \\
&\overset{\text{MAR}}{=} \int_{\mathcal{Z}_{(\bar{r})}} p(r \mid z_{(r)}, \psi) p(z \mid \theta)\ dz_{(\bar{r})} \\
&= p(r \mid z_{(r)}, \psi) \int_{\mathcal{Z}_{(\bar{r})}} p(z \mid \theta)\ dz_{(\bar{r})} \\
&= p(r \mid z_{(r)}, \psi)\ p(z_{(r)} \mid \theta)
\end{aligned}$$

- Under MAR, likelihood-based inference can be based on

$$\ell_{obs}(\theta) = p(z_{(r)} \mid \theta) = \int_{\mathcal{Z}_{(\bar{r})}} p(z \mid \theta)\ dz_{(\bar{r})}$$

- Missingness mechanism is *ignorable* since there's no need to specify $p(r \mid z, \psi)$ if we only care about θ

Ignorability

From Little & Rubin (2002, Definition 6.4):

The missing-data mechanism is ignorable for likelihood inference if:

(a) MAR holds

(b) The parameters θ and ψ are separable

Maximum-Likelihood Estimation

The MLE for θ is obtained from maximizing

$$L_{obs}(\theta, \psi) = \prod_{i=1}^{n} \int_{\mathcal{Z}_{(\bar{r}_i)}} p(r_i \mid z_i, \psi) p(z_i \mid \theta) \, dz_{i(\bar{r}_i)}$$

$$\overset{\text{MAR}}{=} \underbrace{\left[\prod_{i=1}^{n} p(r_i \mid z_{i(r_i)}, \psi) \right]}_{\text{Can be ignored}} \underbrace{\left[\prod_{i=1}^{n} \int_{\mathcal{Z}_{(\bar{r}_i)}} p(z_i \mid \theta) \, dz_{i(\bar{r}_i)} \right]}_{\text{Provides MLE of } \theta \text{ under MAR}}$$

- It might be difficult to work with these expressions, even under MAR; the EM algorithm might help! (next class)
- Note that the MLE is the same whether we assume MAR, MCAR, or anything that satisfies ignorability!

Maximum-Likelihood Estimation

The MLE for θ is obtained from maximizing

$$L_{obs}(\theta, \psi) = \prod_{i=1}^{n} \int_{\mathcal{Z}_{(\bar{r}_i)}} p(r_i \mid z_i, \psi) p(z_i \mid \theta) \, dz_{i(\bar{r}_i)}$$

$$\stackrel{\text{MAR}}{=} \underbrace{\left[\prod_{i=1}^{n} p(r_i \mid z_{i(r_i)}, \psi) \right]}_{\text{Can be ignored}} \underbrace{\left[\prod_{i=1}^{n} \int_{\mathcal{Z}_{(\bar{r}_i)}} p(z_i \mid \theta) \, dz_{i(\bar{r}_i)} \right]}_{\text{Provides MLE of } \theta \text{ under MAR}}$$

- It might be difficult to work with these expressions, even under MAR; the EM algorithm might help! (next class)
- Note that the MLE is the same whether we assume MAR, MCAR, or anything that satisfies ignorability!

Maximum-Likelihood Estimation

The MLE for θ is obtained from maximizing

$$L_{obs}(\theta, \psi) = \prod_{i=1}^{n} \int_{\mathcal{Z}_{(\bar{r}_i)}} p(r_i \mid z_i, \psi) p(z_i \mid \theta) \; dz_{i(\bar{r}_i)}$$

$$\overset{\text{MAR}}{=} \underbrace{\left[\prod_{i=1}^{n} p(r_i \mid z_{i(r_i)}, \psi) \right]}_{\text{Can be ignored}} \underbrace{\left[\prod_{i=1}^{n} \int_{\mathcal{Z}_{(\bar{r}_i)}} p(z_i \mid \theta) \; dz_{i(\bar{r}_i)} \right]}_{\text{Provides MLE of } \theta \text{ under MAR}}$$

- It might be difficult to work with these expressions, even under MAR; the EM algorithm might help! (next class)
- Note that the MLE is the same whether we assume MAR, MCAR, or anything that satisfies ignorability!

Maximum-Likelihood Estimation

The MLE for θ is obtained from maximizing

$$L_{obs}(\theta, \psi) = \prod_{i=1}^{n} \int_{\mathcal{Z}_{(\bar{r}_i)}} p(r_i \mid z_i, \psi) p(z_i \mid \theta) \; dz_{i(\bar{r}_i)}$$

$$\stackrel{\text{MAR}}{=} \underbrace{\left[\prod_{i=1}^{n} p(r_i \mid z_{i(r_i)}, \psi) \right]}_{\text{Can be ignored}} \underbrace{\left[\prod_{i=1}^{n} \int_{\mathcal{Z}_{(\bar{r}_i)}} p(z_i \mid \theta) \; dz_{i(\bar{r}_i)} \right]}_{\text{Provides MLE of } \theta \text{ under MAR}}$$

- It might be difficult to work with these expressions, even under MAR; the EM algorithm might help! (next class)
- Note that the MLE is the same whether we assume MAR, MCAR, or anything that satisfies ignorability!

Observed-Data Score Vector and Fisher Information

- Davidian and Tsiatis, in pages 60–61, present expressions equivalent to the following
  - The score vector

$$S_\theta(r, z_{(r)}; \theta) = \frac{\partial}{\partial \theta} \log p(z_{(r)} \mid \theta)$$

  - The Fisher's information matrix

$$\mathcal{I}(\theta) = -E\left[\frac{\partial^2}{\partial\theta\partial\theta^T} \log p(Z_{(R)} \mid \theta)\right]$$

  - The observed-information matrix

$$J(\theta) = -\sum_{i=1}^{n} \frac{\partial^2}{\partial\theta\partial\theta^T} \log p(z_{i(r_i)} \mid \theta)$$

- Davidian and Tsiatis provide alternative expressions for these quantities that require some algebraic manipulations (check on your own)
- Note that while we can ignore $p(r \mid z_{(r)}, \psi)$ to compute the MLE, the expectation to obtain $\mathcal{I}(\theta)$ is over $(R, Z_{(R)})$
- If response mechanism is not ignorable, these quantities need to be derived from $L_{obs}(\theta, \psi)$!

Observed-Data Score Vector and Fisher Information

- Davidian and Tsiatis, in pages 60–61, present expressions equivalent to the following
  - The score vector
  $$S_\theta(r, z_{(r)}; \theta) = \frac{\partial}{\partial\theta} \log p(z_{(r)} \mid \theta)$$
  - The Fisher's information matrix
  $$\mathcal{I}(\theta) = -E\left[\frac{\partial^2}{\partial\theta\partial\theta^T} \log p(Z_{(R)} \mid \theta)\right]$$
  - The observed-information matrix
  $$J(\theta) = -\sum_{i=1}^{n} \frac{\partial^2}{\partial\theta\partial\theta^T} \log p(z_{i(r_i)} \mid \theta)$$
- Davidian and Tsiatis provide alternative expressions for these quantities that require some algebraic manipulations (check on your own)
- Note that while we can ignore $p(r \mid z_{(r)}, \psi)$ to compute the MLE, the expectation to obtain $\mathcal{I}(\theta)$ is over $(R, Z_{(R)})$
- If response mechanism is not ignorable, these quantities need to be derived from $L_{obs}(\theta, \psi)$!

Observed-Data Score Vector and Fisher Information

- Davidian and Tsiatis, in pages 60–61, present expressions equivalent to the following
  - The score vector

$$S_\theta(r, z_{(r)}; \theta) = \frac{\partial}{\partial \theta} \log p(z_{(r)} \mid \theta)$$

  - The Fisher's information matrix

$$\mathcal{I}(\theta) = -E\left[\frac{\partial^2}{\partial\theta\partial\theta^T} \log p(Z_{(R)} \mid \theta)\right]$$

  - The observed-information matrix

$$J(\theta) = -\sum_{i=1}^{n} \frac{\partial^2}{\partial\theta\partial\theta^T} \log p(z_{i(r_i)} \mid \theta)$$

- Davidian and Tsiatis provide alternative expressions for these quantities that require some algebraic manipulations (check on your own)
- Note that while we can ignore $p(r \mid z_{(r)}, \psi)$ to compute the MLE, the expectation to obtain $\mathcal{I}(\theta)$ is over $(R, Z_{(R)})$
- If response mechanism is not ignorable, these quantities need to be derived from $L_{obs}(\theta, \psi)$!

Observed-Data Score Vector and Fisher Information

- Davidian and Tsiatis, in pages 60–61, present expressions equivalent to the following
  - The score vector
    $$S_\theta(r, z_{(r)}; \theta) = \frac{\partial}{\partial\theta} \log p(z_{(r)} \mid \theta)$$
  - The Fisher's information matrix
    $$\mathcal{I}(\theta) = -E\left[\frac{\partial^2}{\partial\theta\partial\theta^T} \log p(Z_{(R)} \mid \theta)\right]$$
  - The observed-information matrix
    $$J(\theta) = -\sum_{i=1}^{n} \frac{\partial^2}{\partial\theta\partial\theta^T} \log p(z_{i(r_i)} \mid \theta)$$
- Davidian and Tsiatis provide alternative expressions for these quantities that require some algebraic manipulations (check on your own)
- Note that while we can ignore $p(r \mid z_{(r)}, \psi)$ to compute the MLE, the expectation to obtain $\mathcal{I}(\theta)$ is over $(R, Z_{(R)})$
- If response mechanism is not ignorable, these quantities need to be derived from $L_{obs}(\theta, \psi)$!

Outline

Discussion on What the MAR Assumption Says

- Rubin (1976, Biometrika) introduced a slightly different the idea of MAR
- People use and understand something else – the difference is subtle
- Does it matter?

The Original MAR Assumption

Rubin (1976, Biometrika):

- $\mathbf{r}$: response indicators for your entire dataset, realized, fixed
- $\mathbf{z}_{(\mathbf{r})}$: observed values for entire dataset, realized, fixed
- Rubin's original definition says:

  Missing data $\mathbf{z}_{(\bar{\mathbf{r}})}$ are MAR if

  $$p(\mathbf{r} \mid \mathbf{z}_{(\mathbf{r})}, \mathbf{z}_{(\bar{\mathbf{r}})}, \phi) = p(\mathbf{r} \mid \mathbf{z}_{(\mathbf{r})}, \mathbf{z}'_{(\bar{\mathbf{r}})}, \phi)$$

  for all possible values $\mathbf{z}_{(\bar{\mathbf{r}})}$, $\mathbf{z}'_{(\bar{\mathbf{r}})}$ and ϕ
- This doesn't say anything about other $\mathbf{r}' \neq \mathbf{r}$ or other $\mathbf{z}'_{(\mathbf{r})} \neq \mathbf{z}_{(\mathbf{r})}$
- It's an assumption on the probability of observing what I observed, not about what I could have observed

The Original MAR Assumption

Rubin (1976, Biometrika):

- $\mathbf{r}$: response indicators for your entire dataset, realized, fixed
- $\mathbf{z}_{(\mathbf{r})}$: observed values for entire dataset, realized, fixed
- Rubin's original definition says:

  Missing data $\mathbf{z}_{(\bar{\mathbf{r}})}$ are MAR if

  $$p(\mathbf{r} \mid \mathbf{z}_{(\mathbf{r})}, \mathbf{z}_{(\bar{\mathbf{r}})}, \phi) = p(\mathbf{r} \mid \mathbf{z}_{(\mathbf{r})}, \mathbf{z}'_{(\bar{\mathbf{r}})}, \phi)$$

  for all possible values $\mathbf{z}_{(\bar{\mathbf{r}})}$, $\mathbf{z}'_{(\bar{\mathbf{r}})}$ and ϕ

- This doesn't say anything about other $\mathbf{r}' \neq \mathbf{r}$ or other $\mathbf{z}'_{(\mathbf{r})} \neq \mathbf{z}_{(\mathbf{r})}$
- It's an assumption on the probability of observing what I observed, not about what I could have observed

The Original MAR Assumption

Rubin (1976, Biometrika):

- $\mathbf{r}$: response indicators for your entire dataset, realized, fixed
- $\mathbf{z}_{(\mathbf{r})}$: observed values for entire dataset, realized, fixed
- Rubin's original definition says:

  Missing data $\mathbf{z}_{(\bar{\mathbf{r}})}$ are MAR if

  $$p(\mathbf{r} \mid \mathbf{z}_{(\mathbf{r})}, \mathbf{z}_{(\bar{\mathbf{r}})}, \phi) = p(\mathbf{r} \mid \mathbf{z}_{(\mathbf{r})}, \mathbf{z}'_{(\bar{\mathbf{r}})}, \phi)$$

  for all possible values $\mathbf{z}_{(\bar{\mathbf{r}})}$, $\mathbf{z}'_{(\bar{\mathbf{r}})}$ and ϕ
- This doesn't say anything about other $\mathbf{r}' \neq \mathbf{r}$ or other $\mathbf{z}'_{(\mathbf{r})} \neq \mathbf{z}_{(\mathbf{r})}$
- It's an assumption on the probability of observing what I observed, not about what I could have observed

The Original MAR Assumption

Rubin (1976, Biometrika):

- $\mathbf{r}$: response indicators for your entire dataset, realized, fixed
- $\mathbf{z}_{(\mathbf{r})}$: observed values for entire dataset, realized, fixed
- Rubin's original definition says:

  Missing data $\mathbf{z}_{(\bar{\mathbf{r}})}$ are MAR if

  $$p(\mathbf{r} \mid \mathbf{z}_{(\mathbf{r})}, \mathbf{z}_{(\bar{\mathbf{r}})}, \phi) = p(\mathbf{r} \mid \mathbf{z}_{(\mathbf{r})}, \mathbf{z}'_{(\bar{\mathbf{r}})}, \phi)$$

  for all possible values $\mathbf{z}_{(\bar{\mathbf{r}})}$, $\mathbf{z}'_{(\bar{\mathbf{r}})}$ and ϕ
- This doesn't say anything about other $\mathbf{r}' \neq \mathbf{r}$ or other $\mathbf{z}'_{(\mathbf{r})} \neq \mathbf{z}_{(\mathbf{r})}$
- It's an assumption on the probability of observing what I observed, not about what I could have observed

Example: the Original MAR Assumption

Example: let's say I try to measure Gender, Age, and Income on two individuals

- $r_1 = 110, \quad z_1 = (F, 29, 100K), \quad r_2 = 010, \quad z_2 = (M, 40, 80K)$
- Missing data: $\quad z_{1(\bar{r}_1)} = (100K), \quad z_{2(\bar{r}_2)} = (M, 80K)$
- In Rubin's original definition, the missing data are MAR if

$$p\left(R_1 = 110, R_2 = 010 \mid Z_1 = (F, 29, 100K), Z_2 = (M, 40, 80K)\right) =$$

$$p\left(R_1 = 110, R_2 = 010 \mid Z_1 = (F, 29, a), Z_2 = (b, 40, c)\right),$$

for any values of a, b, c

- Rubin's original MAR assumption doesn't say anything about

$$p(R_1 = r_1', R_2 = r_2' \mid z_1', z_2')$$

for $r_1' \neq 110$, or $r_2' \neq 010$, or $z'_{1(r_1)} \neq (F, 29)$ or $z'_{2(r_2)} \neq (40)$

Example: the Original MAR Assumption

Example: let's say I try to measure Gender, Age, and Income on two individuals

- $r_1 = 110, \quad z_1 = (F, 29, 100K), \quad r_2 = 010, \quad z_2 = (M, 40, 80K)$
- Missing data: $\quad z_{1(\bar{r}_1)} = (100K), \quad z_{2(\bar{r}_2)} = (M, 80K)$
- In Rubin's original definition, the missing data are MAR if

$$p\left(R_1 = 110, R_2 = 010 \mid Z_1 = (F, 29, 100K), Z_2 = (M, 40, 80K)\right) =$$

$$p\left(R_1 = 110, R_2 = 010 \mid Z_1 = (F, 29, a), Z_2 = (b, 40, c)\right),$$

  for any values of a, b, c

- Rubin's original MAR assumption doesn't say anything about

$$p(R_1 = r_1', R_2 = r_2' \mid z_1', z_2')$$

  for $r_1' \neq 110$, or $r_2' \neq 010$, or $z'_{1(r_1)} \neq (F, 29)$ or $z'_{2(r_2)} \neq (40)$

Example: the Original MAR Assumption

Example: let's say I try to measure Gender, Age, and Income on two individuals

- $r_1 = 110, \quad z_1 = (F, 29, 100K), \quad r_2 = 010, \quad z_2 = (M, 40, 80K)$
- Missing data: $\quad z_{1(\bar{r}_1)} = (100K), \quad z_{2(\bar{r}_2)} = (M, 80K)$
- In Rubin's original definition, the missing data are MAR if

$$p\left(R_1 = 110, R_2 = 010 \mid Z_1 = (F, 29, 100K), Z_2 = (M, 40, 80K)\right) =$$

$$p\left(R_1 = 110, R_2 = 010 \mid Z_1 = (F, 29, a), Z_2 = (b, 40, c)\right),$$

  for any values of a, b, c

- Rubin's original MAR assumption doesn't say anything about

$$p(R_1 = r_1', R_2 = r_2' \mid z_1', z_2')$$

  for $r_1' \neq 110$, or $r_2' \neq 010$, or $z'_{1(r_1)} \neq (F, 29)$ or $z'_{2(r_2)} \neq (40)$

Example: the Original MAR Assumption

Example: let's say I try to measure Gender, Age, and Income on two individuals

- $r_1 = 110, \quad z_1 = (F, 29, 100K), \quad r_2 = 010, \quad z_2 = (M, 40, 80K)$
- Missing data: $\quad z_{1(\bar{r}_1)} = (100K), \quad z_{2(\bar{r}_2)} = (M, 80K)$
- In Rubin's original definition, the missing data are MAR if

$$p\left(R_1 = 110, R_2 = 010 \mid Z_1 = (F, 29, 100K), Z_2 = (M, 40, 80K)\right) =$$

$$p\left(R_1 = 110, R_2 = 010 \mid Z_1 = (F, 29, a), Z_2 = (b, 40, c)\right),$$

for any values of a, b, c

- Rubin's original MAR assumption doesn't say anything about

$$p(R_1 = r_1', R_2 = r_2' \mid z_1', z_2')$$

for $r_1' \neq 110$, or $r_2' \neq 010$, or $z'_{1(r_1)} \neq (F, 29)$ or $z'_{2(r_2)} \neq (40)$

Example: the Original MAR Assumption

Example: let's say I try to measure Gender, Age, and Income on two individuals

- $r_1 = 110, \quad z_1 = (F, 29, 100K), \quad r_2 = 010, \quad z_2 = (M, 40, 80K)$
- Missing data: $\quad z_{1(\bar{r}_1)} = (100K), \quad z_{2(\bar{r}_2)} = (M, 80K)$
- In Rubin's original definition, the missing data are MAR if

$$p\left(R_1 = 110, R_2 = 010 \mid Z_1 = (F, 29, 100K), Z_2 = (M, 40, 80K)\right) =$$

$$p\left(R_1 = 110, R_2 = 010 \mid Z_1 = (F, 29, a), Z_2 = (b, 40, c)\right),$$

  for any values of a, b, c

- Rubin's original MAR assumption doesn't say anything about

$$p(R_1 = r_1', R_2 = r_2' \mid z_1', z_2')$$

  for $r_1' \neq 110$, or $r_2' \neq 010$, or $z'_{1(r_1)} \neq (F, 29)$ or $z'_{2(r_2)} \neq (40)$

The MAR Assumption Today

- Today, most authors interpret the MAR assumption as

$$p(\mathbf{r} \mid \mathbf{z}_{(\mathbf{r})}, \mathbf{z}_{(\bar{\mathbf{r}})}, \phi) = p(\mathbf{r} \mid \mathbf{z}_{(\mathbf{r})}, \mathbf{z}'_{(\bar{\mathbf{r}})}, \phi)$$

  for all possible values $\mathbf{r}$, $\mathbf{z}_{(\mathbf{r})}$, $\mathbf{z}_{(\bar{\mathbf{r}})}$, $\mathbf{z}'_{(\bar{\mathbf{r}})}$ and ϕ

  - Equivalently,

$$p(\mathbf{r} \mid \mathbf{z}, \phi) = p(\mathbf{r} \mid \mathbf{z}_{(\mathbf{r})}, \phi)$$

    for all possible values $\mathbf{r}$, $\mathbf{z}$, and ϕ

- Mealli & Rubin (2015, Biometrika) call this *missing always at random* – MAAR (see also Seaman et al. (2013, Stat. Sci.))

- However, we don't really use the original definition of MAR; for example, nobody says *"I will assume MAR if I obtain* $\mathbf{r}$ *and* $\mathbf{z}_{(\mathbf{r})}$*, but not if I obtain* $\mathbf{r}'$ *or* $\mathbf{z}'_{(\mathbf{r})}$*"*

- Here we'll use the common interpretation of MAR (MAAR). With i.i.d. data, it corresponds to assuming

$$p(r \mid z, \phi) = p(r \mid z_{(r)}, \phi),$$

  for a generic observation, for all possible values r, z, and ϕ

The MAR Assumption Today

- Today, most authors interpret the MAR assumption as

$$p(\mathbf{r} \mid \mathbf{z}_{(\mathbf{r})}, \mathbf{z}_{(\bar{\mathbf{r}})}, \phi) = p(\mathbf{r} \mid \mathbf{z}_{(\mathbf{r})}, \mathbf{z}'_{(\bar{\mathbf{r}})}, \phi)$$

  for all possible values $\mathbf{r}$, $\mathbf{z}_{(\mathbf{r})}$, $\mathbf{z}_{(\bar{\mathbf{r}})}$, $\mathbf{z}'_{(\bar{\mathbf{r}})}$ and ϕ

  - Equivalently,

  $$p(\mathbf{r} \mid \mathbf{z}, \phi) = p(\mathbf{r} \mid \mathbf{z}_{(\mathbf{r})}, \phi)$$

    for all possible values $\mathbf{r}$, $\mathbf{z}$, and ϕ

- Mealli & Rubin (2015, Biometrika) call this *missing always at random* – MAAR (see also Seaman et al. (2013, Stat. Sci.))

- However, we don't really use the original definition of MAR; for example, nobody says *"I will assume MAR if I obtain* $\mathbf{r}$ *and* $\mathbf{z}_{(\mathbf{r})}$*, but not if I obtain* $\mathbf{r}'$ *or* $\mathbf{z}'_{(\mathbf{r})}$*"*

- Here we'll use the common interpretation of MAR (MAAR). With i.i.d. data, it corresponds to assuming

$$p(r \mid z, \phi) = p(r \mid z_{(r)}, \phi),$$

  for a generic observation, for all possible values r, z, and ϕ

The MAR Assumption Today

- Today, most authors interpret the MAR assumption as

$$p(\mathbf{r} \mid \mathbf{z}_{(\mathbf{r})}, \mathbf{z}_{(\bar{\mathbf{r}})}, \phi) = p(\mathbf{r} \mid \mathbf{z}_{(\mathbf{r})}, \mathbf{z}'_{(\bar{\mathbf{r}})}, \phi)$$

  for all possible values $\mathbf{r}$, $\mathbf{z}_{(\mathbf{r})}$, $\mathbf{z}_{(\bar{\mathbf{r}})}$, $\mathbf{z}'_{(\bar{\mathbf{r}})}$ and ϕ

  - Equivalently,

$$p(\mathbf{r} \mid \mathbf{z}, \phi) = p(\mathbf{r} \mid \mathbf{z}_{(\mathbf{r})}, \phi)$$

    for all possible values $\mathbf{r}$, $\mathbf{z}$, and ϕ

- Mealli & Rubin (2015, Biometrika) call this *missing always at random* – MAAR (see also Seaman et al. (2013, Stat. Sci.))

- However, we don't really use the original definition of MAR; for example, nobody says *"I will assume MAR if I obtain* $\mathbf{r}$ *and* $\mathbf{z}_{(\mathbf{r})}$*, but not if I obtain* $\mathbf{r}'$ *or* $\mathbf{z}'_{(\mathbf{r})}$*"*

- Here we'll use the common interpretation of MAR (MAAR). With i.i.d. data, it corresponds to assuming

$$p(r \mid z, \phi) = p(r \mid z_{(r)}, \phi),$$

  for a generic observation, for all possible values r, z, and ϕ

The MAR Assumption Today

- Today, most authors interpret the MAR assumption as

$$p(\mathbf{r} \mid \mathbf{z}_{(\mathbf{r})}, \mathbf{z}_{(\bar{\mathbf{r}})}, \phi) = p(\mathbf{r} \mid \mathbf{z}_{(\mathbf{r})}, \mathbf{z}'_{(\bar{\mathbf{r}})}, \phi)$$

  for all possible values $\mathbf{r}$, $\mathbf{z}_{(\mathbf{r})}$, $\mathbf{z}_{(\bar{\mathbf{r}})}$, $\mathbf{z}'_{(\bar{\mathbf{r}})}$ and ϕ

  - Equivalently,

$$p(\mathbf{r} \mid \mathbf{z}, \phi) = p(\mathbf{r} \mid \mathbf{z}_{(\mathbf{r})}, \phi)$$

    for all possible values $\mathbf{r}$, $\mathbf{z}$, and ϕ

- Mealli & Rubin (2015, Biometrika) call this *missing always at random* – MAAR (see also Seaman et al. (2013, Stat. Sci.))

- However, we don't really use the original definition of MAR; for example, nobody says "*I will assume MAR if I obtain* $\mathbf{r}$ *and* $\mathbf{z}_{(\mathbf{r})}$*, but not if I obtain* $\mathbf{r}'$ *or* $\mathbf{z}'_{(\mathbf{r})}$"

- Here we'll use the common interpretation of MAR (MAAR). With i.i.d. data, it corresponds to assuming

$$p(r \mid z, \phi) = p(r \mid z_{(r)}, \phi),$$

  for a generic observation, for all possible values r, z and ϕ

The MAR Assumption Today

- Today, most authors interpret the MAR assumption as

$$p(\mathbf{r} \mid \mathbf{z}_{(\mathbf{r})}, \mathbf{z}_{(\bar{\mathbf{r}})}, \phi) = p(\mathbf{r} \mid \mathbf{z}_{(\mathbf{r})}, \mathbf{z}'_{(\bar{\mathbf{r}})}, \phi)$$

  for all possible values $\mathbf{r}$, $\mathbf{z}_{(\mathbf{r})}$, $\mathbf{z}_{(\bar{\mathbf{r}})}$, $\mathbf{z}'_{(\bar{\mathbf{r}})}$ and ϕ

  - Equivalently,

$$p(\mathbf{r} \mid \mathbf{z}, \phi) = p(\mathbf{r} \mid \mathbf{z}_{(\mathbf{r})}, \phi)$$

    for all possible values $\mathbf{r}$, $\mathbf{z}$, and ϕ

- Mealli & Rubin (2015, Biometrika) call this *missing always at random* – MAAR (see also Seaman et al. (2013, Stat. Sci.))

- However, we don't really use the original definition of MAR; for example, nobody says "*I will assume MAR if I obtain* $\mathbf{r}$ *and* $\mathbf{z}_{(\mathbf{r})}$, *but not if I obtain* $\mathbf{r}'$ *or* $\mathbf{z}'_{(\mathbf{r})}$"

- Here we'll use the common interpretation of MAR (MAAR). With i.i.d. data, it corresponds to assuming

$$p(r \mid z, \phi) = p(r \mid z_{(r)}, \phi),$$

  for a generic observation, for all possible values r, z, and ϕ

Outline

Summary

Main take-aways from today's lecture:

- In general, likelihood-based inference requires positing a model for the study variables and for the response mechanism
- Under ignorability (MAR + separability), we don't need to explicitly write the response mechanism
- Original MAR definition has mutated over the years

Next lecture:

- The EM algorithm!

Summary

Main take-aways from today's lecture:

- In general, likelihood-based inference requires positing a model for the study variables and for the response mechanism
- Under ignorability (MAR + separability), we don't need to explicitly write the response mechanism
- Original MAR definition has mutated over the years

Next lecture:

- The EM algorithm!