

Lecture 4

Multiple imputation

Revision of single random imputation

- Suppose that a variable Y is observed for some individuals but missing for others
- In single random imputation a single value is selected, at random, to replace each missing value of Y
- For example, in stochastic regression imputation we impute

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \dots + \hat{\beta}_p X_{pi} + s\epsilon_i$$

where X_1, \dots, X_p are explanatory variables which are observed for all individuals and $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$ are regression parameters estimated from the complete cases, $\epsilon_i \sim \text{Normal}(0, 1)$ and s^2 is the mean square error from the fitted model

Uncertainty in imputation

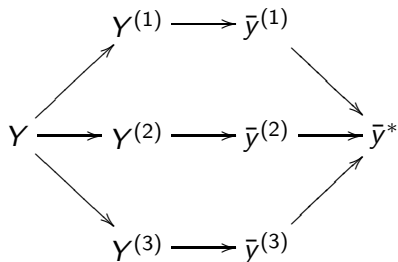
- If we carried out a second single random imputation, with the same data, we would obtain different imputed values
- Does this matter?
- How can we quantify this **imputation uncertainty**?
- Could try a sensitivity analysis, or could use **multiple imputation**...

Multiple imputation - key ideas

- Replace **each missing value** with N imputed values, selected **at random** as in single random imputation where $N \geq 2$
- Consequently have multiple imputed data sets
- We can carry out a **data analysis** of **each** of these imputed data sets and estimate whatever parameters we are interested in (means, variances, correlation, regression coefficients *etc*)
- The **variability** across the N parameter estimates reflects our uncertainty due to the imputation
- How can we combine the N parameter estimates to get a **single** estimate of the parameter of interest?

Schematic of MI

Let $Y = (Y_{obs}, Y_{mis})$ and suppose that we want to estimate the population mean of Y using the sample mean. We will use three imputations



$Y^{(i)}$ denotes the i th imputed data set and $\bar{y}^{(i)}$ the associated sample mean. How do we combine these sample means to get the final estimate \bar{y}^* ?

Example

Suppose that we want to estimate the correlation between variables X and Y . X is complete but some values of Y are missing.

1. Fit a regression model, estimating the regression parameters β_0 and β_1 using the complete cases

$$Y_i = \beta_0 + \beta_1 x_i$$

and the MSE s of the residuals

2. For each missing value **impute**

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + s \epsilon_i$$

where ϵ_i is a random draw from a Normal(0, 1) distribution

3. Repeat step 2 $N \geq 2$ times to obtain **multiple imputations**.

Example continued

For $N = 4$ we might have

X	Y	$j = 1$	$j = 2$	$j = 3$	$j = 4$
6.5	-	2.9	2.2	2.0	3.4
2.9	1.4				
7.6	3.3				
5.9	4.2				
5.1	-0.24				
5.3	2.8				
4.2	-	-2.2	2.0	2.1	-0.79
8.3	4.8				
1.9	1.3				
6.5	-	5.4	2.0	6.5	4.7

From the complete cases $Y_i = -0.30 + 0.53X_i$ with $s = 1.19$.

Example continued

Calculate 4 correlations - one from each imputed data set

j	$\text{corr}(X, Y)$
1	0.62
2	0.63
3	0.60
4	0.68

- We could obtain a single (**point**) estimate of the correlation by taking the mean of our 4 estimates, that is 0.63
- Correlation estimates are approximately unbiased
- However the **estimated standard errors** are **biased downwards** (*i.e.* too small) because they do not take into account multiple imputation
- The variability across imputations can be used to **adjust these standard errors upwards**

How MI works - summary

- Restores variation in missing data and accounts for uncertainty due to imputation
- Original variability obtained by imputing missing values using variables correlated with the missing variables
- Imputation uncertainty accounted for by obtaining several imputed data sets

Have not yet discussed how many imputed data sets we might take, but it turns out that only a very small number are required ($N = 3, 4$ or 5).

Extending MI - missing values in explanatory variables

- So far, we have only discussed what happens if there are missing values in the response. What happens if there are also missing values in the explanatory variables?
- If this is the case we use **Bayesian inference** to sample the missing values from their **posterior predictive distribution**
- The main inference, once we have imputed all the missing values, can be done using **either** a frequentist (maximum likelihood) **or** a Bayesian approach

Review of Bayesian methods

- Alternative way to estimate parameters in a statistical model
- Assume that **parameters are random variables**, *i.e.* have a probability distribution
- Starting point is our **prior belief** about this probability distribution
- Goal is to use the information from the data to update our prior probability distribution
- Resulting distribution is known as the **posterior distribution**

Conditional probability and Bayes Theorem

- Suppose that A and B are two events in a sample space S then the **conditional probability of A given B** is

$$\Pr[A|B] = \frac{\Pr[A \cap B]}{\Pr[B]} = \frac{\Pr[B|A] \Pr[A]}{\Pr[B]}$$

- Now suppose that Y and θ are two random variables and that p denotes a probability density function, then

$$p(\theta|y) = \frac{p(\theta, y)}{p(y)} = \frac{p(y|\theta)p(\theta)}{p(y)}$$

- Now since $p(y)$ **depends on the data only** we have

$$\begin{aligned} p(\theta|y) &\propto p(y|\theta)p(\theta) \\ \text{posterior} &\propto \text{likelihood} \times \text{prior}. \end{aligned}$$

Likelihoods and posteriors

- Notation : you might see the likelihood written as $L(y)$, $L(y|\theta)$, $f(y)$ or $f(y|\theta)$. Similarly the prior and posterior might be written as $\pi(\theta)$ and $\pi(\theta|y)$.
- Usually the posteriors distribution will be summarised with a **point estimate**, e.g. the posterior mode, median or mean (median and mean are easier to calculate)
- Uncertainty in this point estimate is measured using a **credibility interval**
- Bayesian inference accounts both for **sampling uncertainty** (through the likelihood) and **parameter uncertainty** (by putting a probability distribution on the parameter, rather than treating it as fixed but unknown)

Priors 1

- One criticism of Bayesian inference is that it requires the **subjective choice** of the prior distribution $p(\theta)$
- Different people may think that different priors are appropriate, and these priors might lead to different estimates
- Choice of prior might be for ease of **mathematical computations** (sometimes these are **improper**) or **based on historical data**
- Key is to choose a prior that is as **uninformative** as possible, unless there is genuine evidence to suggest a particular form, e.g. from historical data

Priors 2

- A **non-informative prior** is one with a very large variance compared to the range of parameter values supported by the likelihood
- It might be uniform over a large range of values so that, at least approximately

$$p(\theta|y) \propto L(\theta)$$

- Such a prior is referred to as a **flat prior** and many results from frequentist statistics are the same as using a flat prior, although there is a difference in interpretation difference

Priors 3

- Jeffreys' Prior is taken to be the square root of the determinant of the information matrix $I(\theta)$
- Suppose that the data are a random sample from a Normal(μ, σ^2) distribution, then Jeffreys' prior for $\theta = (\mu, \sigma^2)$ is

$$p(\theta) \propto \frac{1}{\sigma^2}, \quad -\infty < \mu < \infty, \sigma^2 > 0$$

- This is an improper distribution as it has an infinite integral over the parameter space
- The posterior distributions are proper and give results equivalent to those obtained under classical inference

Generally, should test sensitivity to the choice of prior.

Bayesian MI

Assume that $Y = (Y_{obs}, Y_{mis})$ has a parametric model characterised by the likelihood $L(y|\theta)$, that θ has prior distribution $p(\theta)$ and that **data are MAR**

1. First draw a sample of the parameter at random from the **posterior distribution** $p(\theta|y)$. Call the sample θ^* .
2. **Given** θ^* , draw at random the missing values from the **conditional predictive distribution**

$$Y_{mis}^* \sim p(y_{mis}|y_{obs}, \theta^*)$$

We illustrate the technique with a couple of examples.

Binomial example

Suppose that $Y \sim \text{Binomial}(n, \theta)$ and that θ has a $\text{Beta}(1, 1)$ prior (equivalently is $\text{Uniform}(0,1)$)

- Assume that $Y^* \sim \text{Binomial}(n, \theta)$ and that Y is independent of Y^* , but that Y^* is unobserved - can we use Y to impute Y^* ?
- **Posterior distribution** for θ is $\text{Beta}(y + 1, n - y + 1)$
- To impute the value of Y^* first draw θ^* at random from the $\text{Beta}(y + 1, n - y + 1)$ posterior distribution, then draw y^* at random from the $\text{Binomial}(n, \theta^*)$ distribution
- The value y^* sampled in this way is **a draw from the predictive distribution for Y^***
- We may draw N independent pairs of (θ^*, y^*) to give N MI's.

Normal example

Suppose that we have $Y = (Y_{obs}, Y_{mis})$ where $Y_{obs} = (Y_1, \dots, Y_a)$ and $Y_{mis} = (Y_{a+1}, \dots, Y_n)$. Assume that the data are MAR, that the Y_i 's are **independent** and that $Y_i \sim \text{Normal}(\mu, \sigma^2)$. **How can we create MI's for Y_{mis} ?**

Assume Jeffrey's prior, *i.e.* $p(\mu, \sigma^2) \propto 1/\sigma^2$ then the **posterior distributions** of μ and σ^2 are

$$\begin{aligned}\mu | \sigma^2, y_{obs} &\sim \text{Normal}(\bar{y}_{obs}, \sigma^2/a) \\ \sigma^2 | y_{obs} &\sim (a-1)s_{obs}^2 / \chi_{a-1}^2\end{aligned}$$

where \bar{y}_{obs} and s_{obs}^2 are the sample mean and variance of the observed data.

Normal example continued

To generate MI's for Y_{mis} repeat the following algorithm for $j = 1, \dots, N$

- Using the observed data, generate a random sample $\sigma_{(j)}^2$ from the posterior distribution $\sigma^2 | y_{obs}$
- Using the observed data and the randomly sampled σ^2 from step 1, generate a random sample $\mu_{(j)}$ from the posterior distribution $\mu | \sigma^2, y_{obs}$
- Using the randomly samples $\mu_{(j)}$ and $\sigma_{(j)}^2$ from steps 1 and 2 draw $Y_{a+1}^{(j)}, \dots, Y_n^{(j)}$ independently from a $\text{Normal}(\mu_{(j)}, \sigma_{(j)}^2)$ distribution.