Missing data lecture 15: Imputation methods

Imputation of missing data

We've covered why imputing each data point with a single value is wrong. Just to review, this is akin to maximizing the likelihood function:

$$L(\theta, y_{(1)} \mid Y_{(0)} = \tilde{y}_{(0)})$$

Instead of the correct:

$$L(\theta \mid Y_{(0)} = \tilde{y}_{(0)}) \propto \int_{\mathcal{Y}_{(1)}} f_Y(Y_{(0)} = \tilde{y}_{(0)}, Y_{(1)} = y_{(1)} \mid \theta) dy_{(1)}$$

We covered two examples in lecture 10. In one example, the $y_i, i = 1, \ldots, n$ were normally distributed with unknown mean μ and variance σ^2 where we imputed $\bar{y}$ for $n - r$ missing data points; this led to a mean estimator that matched the mean estimator which maximized $L(\mu, \sigma^2 \mid Y_{(0)} = \tilde{y}_{(0)})$, but a biased and inconsistent variance estimator of $\hat{\sigma}^2 = \sum_{i=1}^{r}(y_i - \bar{y})/n$.

In another, we had $y_i, i = 1, \ldots, n$ exponentially distributed with unknown rate λ that were censored at a specified time c. Maximizing $L(\lambda, y_{(1)} \mid Y_{(0)} = \tilde{y}_{(0)})$ led to a biased and inconsistent estimator for λ:

$$\hat{\lambda} = (\sum_{i=1}^{r} y_i + (n - r)c)/n$$

which, when divided by the correct estimator, was off by a factor of r/n. If you had a sample where 80% of your observations were censored, your estimate would be 20% of the true estimator.

The key idea is that imputing unknown missing values with a deterministic value will lead to biased, inconsistent estimators. This is true when you're using likelihood-based inference or other inference techniques. This is because we're using a strategy that ignores the variability inherent in these imputations.

An alternative is to fill in values of $y_{(1)}$ with random draws from a distribution. Which distribution you use is dependent on whether the data are MCAR, MAR, or MNAR. One way to think about this is via approximate Bayesian modeling.

Stochastic imputation

Suppose we have a model $f_Y(Y = y \mid \theta)$ which we can rewrite as $f_Y(Y_{(0)} = y_{(0)}, Y_{(1)} = y_{(1)} \mid \theta)$ for a given missing data pattern. We'll assume our missingness process is ignorable, which means we're in an MCAR or a MAR setting. If we want to compute the posterior for θ under a prior $p(\theta)$ and our likelihood $f_Y(Y_{(0)} = y_{(0)}, Y_{(1)} = y_{(1)} \mid \theta)$, the posterior is:

$$\begin{aligned} p(\theta \mid Y_{(0)} = \tilde{y}_{(0)}) &= \int_{\mathcal{Y}_{(1)}} p(\theta, Y_{(1)} = y_{(1)} \mid Y_{(0)} = \tilde{y}_{(0)}) dy_{(1)} \\ &= \int_{\mathcal{Y}_{(1)}} p(\theta \mid Y_{(1)} = y_{(1)}, Y_{(0)} = \tilde{y}_{(0)}) f_{Y_1|Y_0}(Y_{(1)} = y_{(1)} \mid Y_{(0)} = \tilde{y}_{(0)}) dy_{(1)} \end{aligned}$$

This suggests the following approximate scheme:

1. Draw a random set of values $y_{(1)}^{(s)}$ from the distribution $Y_{(1)} \mid Y_{(0)} = \tilde{y}_{(0)}$
2. Draw a value $\theta^{(s)}$ from the posterior $p(\theta \mid Y_{(1)} = y_{(1)}^{(s)}, Y_{(0)} = \tilde{y}_{(0)})$

If we're instead interested in getting posterior moments, like a mean, we could instead compute:

$$\mathbb{E}\left[\theta \mid Y_{(1)} = y_{(1)}^{(s)}, Y_{(0)} = \tilde{y}_{(0)}\right].$$

If we're using a flat prior and we approximate the posterior for $\theta \mid Y_{(1)} = y_{(1)}^{(s)}, Y_{(0)} = \tilde{y}_{(0)}$ as multivariate normal, the approximate value for $\mathbb{E}\left[\theta \mid Y_{(1)} = y_{(1)}^{(s)}, Y_{(0)} = \tilde{y}_{(0)}\right]$ will be $\hat{\theta}_{\text{MLE}}$.

That suggests another approximation to get the posterior mean

1. Draw a random set of values $y_{(1)}^{(s)}$ from the distribution $Y_{(1)} \mid Y_{(0)} = \tilde{y}_{(0)}$
2. Compute $\hat{\theta}_{\text{MLE}}^{(s)}$ from the log-liklehood $\ell(\theta \mid Y_{(1)} = y_{(1)}^{(s)}, Y_{(0)} = \tilde{y}_{(0)})$
3. Compute $\widehat{\text{Cov}}(\theta)^{(s)}$ from the log-liklehood $\ell(\theta \mid Y_{(1)} = y_{(1)}^{(s)}, Y_{(0)} = \tilde{y}_{(0)})$ as $i(\hat{\theta}_{\text{MLE}}^{(s)}) = \sum_{i=1}^{n} -\nabla_\theta^2 \ell(\theta \mid Y_{(1)} = y_{(1)}^{(s)}, Y_{(0)} = \tilde{y}_{(0)}) \mid_{\theta = \hat{\theta}_{\text{MLE}}^{(s)}}$
4. After running S imputations, compute

$$\begin{aligned} \bar{\theta} = \mathbb{E}\left[\theta \mid Y_{(0)} = \tilde{y}_{(0)}\right] &\approx \frac{1}{S} \sum_{s=1}^{S} \hat{\theta}_{\text{MLE}}^{(s)} \\ \text{Cov}(\theta \mid Y_{(0)} = \tilde{y}_{(0)}) &\approx \frac{1}{S} \sum_{s=1}^{S} i(\hat{\theta}_{\text{MLE}}^{(s)}) + \frac{1}{S-1} \sum_{s=1}^{S} (\hat{\theta}_{\text{MLE}}^{(s)} - \bar{\theta})(\hat{\theta}_{\text{MLE}}^{(s)} - \bar{\theta})^T \end{aligned}$$

where the last step comes from the formula for total covariance:

$$\text{Cov}(X \mid Z) = \mathbb{E}\left[\text{Cov}(X \mid Y, Z) \mid Z\right] + \text{Cov}(\mathbb{E}\left[X \mid Y, Z\right] \mid Z)$$

This is all well and good, but the key issue for us is going to be how to generate draws from the conditional distribution:

$$Y_{(1)} \mid Y_{(0)} = \tilde{y}_{(0)}.$$

In fact, the above procedure doesn't have to be run many times, and $S = 1$ is a valid imputation procedure, though it will have higher variance than if $S > 1$.

Stochastic regression imputation

Suppose we have p continuous predictors arranged into an $n \times p$ matrix X, and a response variable y. Let's say we have one predictor, X_p that is missing values. Suppose we have the first r rows have no missing values, and the last $n - r$ values do have missing values.

Then we could run the following regression using the values for the complete units:

$$X_{[1:r,p]} = X_{[1:r,1:(p-1)]}\beta + Y_{[1:r]}\beta_y + \epsilon_{ip}$$

Then we could fill in the $n - r$ missing values with the following values:

$$\hat{x}_{ip} = \beta_0 + \sum_{j=1}^{p-1} \beta_j x_{ij} + \beta_y y_i + z_{ip}$$

where z_{ip} is a normal random variable with variance equal to the residual variance in the regression model above.

This might be odd that we're including values of the response in the regression for the missing X values. This doesn't lead to any issues, however. In fact, this would be the exact distribution we would use in say, an EM algorithm, if all of our data were jointly MV normal and we were computing the Q step.

Stochastic Hot-Deck imputation

The downside of using regression imputation is that we need a model for imputation. One alternative is to use the data itself to provide filled-in values for missing data. The simplest case occurs for univariate missingness: Let y_i be measurements, of which r are observed, and let the mean be $\bar{y}_R$. For the missing values of y_i, we draw a simple random sample with replacement from the donors, which are the completely observed data, $\{y_1, \ldots, y_r\}$.

The mean of the final dataset is:

$$\bar{y}_{\text{HD}} = \frac{1}{n}(r\bar{y}_R + (n-r)\bar{y}_{HD})$$

where $\bar{y}_{\text{HD}} = \frac{1}{n-r}\sum_{i=1}^{r} H_i y_i$ and H_i represents the number of times the value y_i is chosen.

The mean and variance of the hot-deck estimator can be derived from the properties of the multinomial distribution with sampling proportions $1/r, \ldots, 1/r$, which is the distribution for $\{H_1, \ldots, H_r\}$:

$$\mathbb{E}\left[\bar{y}_{\text{HD}} \mid Y_{(0)}\right] = \bar{y}_R$$
$$\text{Var}(\bar{y}_{\text{HD}} \mid Y_{(0)}) = (1 - r^{-1})(1 - r/n)s_R^2/n$$

If one has completely observed covariates $x_1, \ldots, x_p$ for each i, you can define a measure of distance based on the covariates for unit i and j, $d(i, j)$ and randomly choose a donor for a missing unit i from pool of donors that have have $d(i, j) < d_0$.

One good approach to this metric is to use something called predictive mean matching, whereby one fits a regresson to the complete cases, and picks donors for a missing value for the i^{th} participant using the metric $(\hat{y}(x_i) - \hat{y}(x_j))^2$, where $\hat{y}(x)$ is the predicted mean value from the complete case regression applied to a predictor vector x.

Multiple imputation

We still haven't quite figured out a way to draw from this distribution

$$Y_{(1)} \mid Y_{(0)} = \tilde{y}_{(0)}.$$

Our textbook lists several ideas:

1. Improper MI: generate draws $y_{(1)}^{(s)}$ from $Y_{(1)} \mid Y_{(0)} = \tilde{y}_{(0)}, \theta$ where θ is some estimate of θ maybe from complete units. The reason this is called improper is because it doesn't propagate uncertainty in the estimated θ.
2. MI with asymptotic MLE from complete units: generate draws $y_{(1)}^{(s)}$ from $Y_{(1)} \mid Y_{(0)} = \tilde{y}_{(0)}, \theta$ where $\theta \sim \text{N}(\hat{\theta}, i(\hat{\theta})^{-1})$.
3. Chained equation multiple imputation

This is the most flexible option, and the one that is most widely implemented in software packages that do multiple imputation.

The idea is that if you have many variables, say $Y_1, \ldots, Y_p$, and completely observed covariates Z, one could formulate a sequence of conditional models, just like in Gibbs sampling:

$$\begin{aligned}
&f_{Y_1}(y_1 \mid y_2, \ldots, y_p, z, \psi_1) \\
&\vdots \\
&f_{Y_j}(y_j \mid y_1, \ldots, y_{j-1}, y_{j+1}, \ldots, y_p, z, \psi_j) \\
&f_{Y_p}(y_p \mid y_1, \ldots, y_{p-1}, z, \psi_p)
\end{aligned}$$

let $y_{j(0)}$ be the observed values in the j^{th} column of Y and let $y_{j(1)}$ be the missing values. Let the following matrices be the observed information for given MLEs $\hat{\psi}_1, \dots, \hat{\psi}_p$ and observed data $\tilde{y}_{(0)1}, \dots, \tilde{y}_{(0)p}$:

$$
\begin{aligned}
H_{\psi_1,\psi_1}(\hat{\psi}_1) &= -\nabla^2_{\psi_1} \log f_{Y_1}(\tilde{y}_{1(0)} \mid y_2, \dots, y_p, z, \psi_1) \mid_{\psi_1=\hat{\psi}_1} \\
&\vdots \\
H_{\psi_j,\psi_j}(\hat{\psi}_j) &= -\nabla^2_{\psi_j} \log f_{Y_j}(\tilde{y}_{j(0)} \mid y_1, \dots, y_{j-1}, y_{j+1}, \dots, y_p, z, \psi_j) \mid_{\psi_j=\hat{\psi}_j} \\
&\vdots \\
H_{\psi_p,\psi_p}(\hat{\psi}_p) &= -\nabla^2_{\psi_p} \log f_{Y_p}(\tilde{y}_{p(0)} \mid y_1, \dots, y_{p-1}, z, \psi_p) \mid_{\psi_j=\hat{\psi}_j}
\end{aligned}
$$

Then we run the following algorithm:

1. Generate a set of initial values: $y^{(0)}_{1(1)}, y^{(0)}_{2(1)}, \dots, y^{(0)}_{p(1)}$
2. At step (t) of the algorithm generate draws sequentially

$$
\begin{aligned}
\hat{\psi}_1 &= \text{argmax}_{\psi_1} \log f_{Y_1}(y_{1(0)} \mid y_{2(0)}, y^{(t)}_{2(1)} \dots, y_{p(0)}, y^{(t)}_{p(1)}, z, \psi_1) \\
\hat{\psi}^{(t+1)}_1 &\sim \text{Normal}(\hat{\psi}_1, H_{\psi_1,\psi_1}(\hat{\psi}_1)^{-1}) \\
y^{(t+1)}_{1(1)} &\sim f_{Y_1}(y_{1(1)} \mid y_{2(0)}, y^{(t)}_{2(1)} \dots, y_{p(0)}, y^{(t)}_{p(1)}, z, \hat{\psi}^{(t+1)}_1) \\
&\vdots \\
\hat{\psi}_j &= \text{argmax}_{\psi_j} \log f_{Y_j}(y_{j(0)} \mid y_{1(0)}, y^{(t+1)}_{2(1)} \dots, y_{(j-1)(0)}, y^{(t+1)}_{(j-1)(1)}, \dots y_{p(0)}, y^{(t)}_{p(1)}, z, \psi_j) \\
\hat{\psi}^{(t+1)}_j &\sim \text{Normal}(\hat{\psi}_j, H_{\psi_j,\psi_j}(\hat{\psi}_j)^{-1}) \\
y^{(t+1)}_{j(1)} &\sim f_{Y_j}(y_{j(1)} \mid y_{1(0)}, y^{(t+1)}_{2(1)} \dots, y_{(j-1)(0)}, y^{(t+1)}_{(j-1)(1)}, \dots y_{p(0)}, y^{(t)}_{p(1)}, z, \hat{\psi}^{(t+1)}_j) \\
&\vdots \\
\hat{\psi}_p &= \text{argmax}_{\psi_j} \log f_{Y_p}(y_{p(0)} \mid y_{1(0)}, y^{(t+1)}_{2(1)} \dots, y_{(p-1)(0)}, y^{(t+1)}_{(p-1)(1)}, z, \psi_p) \\
\hat{\psi}^{(t+1)}_p &\sim \text{Normal}(\hat{\psi}_p, H_{\psi_p,\psi_p}(\hat{\psi}_p)^{-1}) \\
y^{(t+1)}_{p(1)} &\sim f_{Y_p}(y_{p(1)} \mid y_{1(0)}, y^{(t+1)}_{2(1)} \dots, y_{(p-1)(0)}, y^{(t+1)}_{(p-1)(1)}, z, \hat{\psi}^{(t+1)}_p)
\end{aligned}
$$

This is a pseudo Bayesian algorithm, where we approximate the marginal posterior predictive distribution:

$$y^{(t+1)}_{1(1)} \sim f_{Y_1}(y_{1(1)} \mid y_{1(0)}, y_{2(0)}, y^{(t)}_{2(1)} \dots, y_{p(0)}, y^{(t)}_{p(1)}, z)$$

with a normal approximation to the posterior for ψ_1 and a draw from the predictive density for $y_{1(1)}$ given all other variables, predictors and the draw from the approximate posterior at step $t+1$, $\hat{\psi}^{(t+1)}_1$. It is possible to do full Bayesian inference, but this would be very computationally intensive.