Maximum Likelihood Estimation w/ missing data

MLE is a method to estimate the parameters of a probabilistic model given data by finding the likeliest parameter values for those observations given the model.

Suppose we have a set of iid* data (observations) $x_1, x_2, \ldots, x_n$.

Our model is $Pr(x_1, x_2, \ldots, x_n | \theta)$, the joint probability density given a set of parameters θ.

→ since iid, this reduces to

$$Pr(x_1, \ldots, x_n | \theta) = Pr(x_1|\theta) Pr(x_2|\theta) \cdots Pr(x_n|\theta) = \prod_{i=1}^{n} Pr(x_i|\theta)$$

→ This lets us say how probable the $\{x_i\}$ are, given $Pr(\cdot)$ and θ.
→ We can reverse this to ask: how likely is θ, given the $Pr(\cdot)$ and the $\{x_i\}$.

$$\Rightarrow \mathcal{L}(\theta; \underbrace{x_1, \ldots, x_n}_{\text{"parameters" of } \mathcal{L}}) = \prod_{i=1}^{n} Pr(x_i|\theta)$$

"variable" of $\mathcal{L}$

Often one works w/ the "log-likelihood" $\ell \equiv \ln \mathcal{L} = \sum_{i=1}^{n} \ln Pr(x_i|\theta)$

Now we can "plug in" different θ's into $\mathcal{L}$ to see which is best. (actually optimize θ to maximize $\mathcal{L}$ either analytically or numerically)

<u>Example</u>: binary data $x_i = 0$ or $x_i = 1$
model $Pr(x_i = 1|\theta) = \theta$ ← $\theta \equiv$ prob of a "1".

Likelihood: $\mathcal{L}(\theta|\{x_i\}) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i}$ $\qquad \theta_{MLE} = \frac{\sum_{i=1}^{n} x_i}{n}$

<u>Missing Data</u> ML is fairly convenient at handling missing values.

Suppose we have <u>two</u> variables x_i, y_i for each observation. Then

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} Pr(x_i, y_i|\theta) \quad \text{as before.}$$

Now suppose x_i is missing for some observations:

→ First m observations have both x_i and y_i

→ Remaining $n-m$ observations have y_i <u>only</u>.

For an i w/ missing data, the probability for y_i only can be computed from its <u>marginal</u> distribution:

$$f(y_i|\theta) = Pr(\cdot, y_i|\theta) = \sum_x Pr(x, y_i|\theta) \quad \text{if } x \text{ discrete}$$

sum over all possible values of x. (integrate if x is continuous)

Likelihood of θ given entire sample is now:

$$\mathcal{L}(\theta) = \underbrace{\prod_{i=1}^{m} Pr(x_i, y_i|\theta)}_{\text{complete observations}} \underbrace{\prod_{i=m+1}^{n} f(y_i|\theta)}_{\text{incomplete observations}}$$

ML strengths: effective for MCAR, MAR! Software available for linear models.
weakness: Requires a model for joint distribution of all variables w/ missing data. Often, assume multivariate normal distribution, but this is not feasible for all problems.

Imputation

→ Fill in the blanks → Replace the missing observations w/ (hopefully) reasonable estimates.

Marginal Mean Imputation

→ Replace any missing values for variable X_j w/ the mean value of the non-missing X_j entries.

Conditional Mean Imputation:

If only one variable X_j has missing values, learn a model for it from the other variables:

$$X_j = g(X_1, X_2, \ldots, X_{j-1}, X_{j+1}, \ldots, X_p)$$

and replace the missing entries x_{ij} with $g(x_{i1}, x_{i2}, \ldots x_{i(j-1)}, x_{i(j+1)}, \ldots x_{ip})$

→ Gets really complicated when multiple X_j's have missing values → which is almost always true.

Last observation Carried Forward

- specific to repeated observations taken over time (longitudinal studies)

unit	observation time t_1	t_2	t_3	...
1	1.7	2.0	?	? ?
2	4.1	3.7	4.0	...
3				
⋮				

⇒

unit	observation time t_1	t_2	t_3	...
1	1.7	2.0	2.0	2.0 ...
2				
⋮				
3				

Logical Rules and Relatedness

Suppose yearly income is missing from a survey, but, sometimes a related question — # months unemployed — is given. If 12 months unemployed, reasonable to impute yearly income = 0.

Simple random imputation

For each missing value for variable j, replace it with a randomly chosen value from the observations where j is observed. Sample w/ replacement if multiple observations are missing variable j.

→ OK if few observations are missing (but then you can just use listwise deletion) but this uses none of the information available regarding how p relates to other variables.

Random (regression) imputation (RI)

Perform conditional mean imputation (typically w/ linear model) but then simulate uncertainty in each imputed measurement by drawing a random noise value ϵ and adding it onto the imputed value.

$$X_{ij} = g(x_{i1}, x_{i2}, \ldots, x_{i(j-1)}, x_{i(j+1)}, \ldots x_p) + \epsilon$$

where typically $\epsilon \sim N(0, \sigma^2)$ and σ^2 is appropriately estimated from the same procedure used to fit g.

Multiple Imputation

Next step up from random imputation.

→ want to reflect uncertainty both in sampling variation (like Random Imputation) but also uncertainty in g itself.

Idea: take M different models $g_1, g_2, \ldots, g_m$, perform R.I. on each, to make M different complete datasets $X_1, \ldots, X_M$. (typically $M = 5-10$)

→ complete observations are duplicated across X's, missing values introduce a new source of variability.

Then estimate $\hat{f}_i$ on each, and pool these together in some way.

Pooling:

$Y = \alpha_i + \beta_i X_i \ , \ i = 1, M$

Suppose each $\hat{f}_i$ is a simple linear regression. Pooling here means combining the estimated regression coefficients

$$\hat{\beta} = \frac{1}{M} \sum_{i=1}^{M} \hat{\beta}_m$$

Where this becomes tricky is we also have uncertainties standard errors associated with each $\hat{\beta}$ and these need to be pooled as well. (variation both within and between imputations is needed)