Bayesian Methods: Part 2

A. Colin Cameron
Univ. of Calif. - Davis

. .

May 2021

1. Introduction

- Consider extensions of Bayesian methods.
- 1. Code up a Metropolis example in Mata.
- 2. Code up a Gibbs example in Mata.
- 3. Multiple Imputation.

Outline

2. Metropolis Algorithm for probit in Mata

- Consider probit model.
- The likelihood is

$$L(\mathbf{y}|\boldsymbol{\beta},\mathbf{X}) = \prod_{i=1}^{N} \Phi(\mathbf{x}_i'\boldsymbol{\beta})^{y_i}(1-\Phi(\mathbf{x}_i'\boldsymbol{\beta}))^{1-y_i}$$

- Use an uninformative prior (all values of $\boldsymbol{\beta}$ equally likely)

$$\pi(\boldsymbol{\beta}) \propto 1$$

  - even though improper the posterior will be proper.
- The posterior is

$$\begin{aligned} p(\boldsymbol{\beta}|\mathbf{y},\mathbf{X}) &\propto L(\mathbf{y}|\boldsymbol{\beta},\mathbf{X}) \times \pi(\boldsymbol{\beta}) \\ &\propto \prod_{i=1}^{N} \Phi(\mathbf{x}_i'\boldsymbol{\beta})^{y_i}(1-\Phi(\mathbf{x}_i'\boldsymbol{\beta}))^{1-y_i} \times 1 \\ &\propto \prod_{i=1}^{N} \Phi(\mathbf{x}_i'\boldsymbol{\beta})^{y_i}(1-\Phi(\mathbf{x}_i'\boldsymbol{\beta}))^{1-y_i} \end{aligned}$$

  - Note: we know $p(\boldsymbol{\beta}|\mathbf{y},\mathbf{X})$ only up to a scale factor.

Random walk Metropolis draws

- We use Metropolis algorithm to make draws from this posterior.
- The random walk MH uses a draw from $\mathcal{N}[\boldsymbol{\beta}^{(s-1)}, c\mathbf{I}]$ where c is set.
  - So we draw $\boldsymbol{\beta}^* = \boldsymbol{\beta}^{(s-1)} + \mathbf{v}$ where $\mathbf{v}$ is draw from $\mathcal{N}[\mathbf{0}, c\mathbf{I}]$
- For $u \sim \text{uniform}[0,1]$ draw and acceptance probability $paccept = p(\boldsymbol{\beta}^*)/p(\boldsymbol{\beta}^{(s-1)})$
  - set $\boldsymbol{\beta}^{(s)} = \boldsymbol{\beta}^*$ if $u < paccept$
  - set $\boldsymbol{\beta}^{(s)} = \boldsymbol{\beta}^{(s-1)}$ if $u > paccept$
- Taking logs, equivalent to
  - $\boldsymbol{\beta}^{(s)} = \boldsymbol{\beta}^*$ if $\ln u < \ln(paccept)$ where
  - $\ln(paccept) = [\sum_i y_i \ln \Phi(\mathbf{x}_i'\boldsymbol{\beta}^*) + (1-y_i)\ln(1-\Phi(\mathbf{x}_i'\boldsymbol{\beta}^*))]$
    $-[\sum_i y_i \ln \Phi(\mathbf{x}_i'\boldsymbol{\beta}^{(s-1)}) + (1-y_i)\ln(1-\Phi(\mathbf{x}_i'\boldsymbol{\beta}^{(s-1)}))]$

Generated data

```
. * Generate data N = 100  Pr[y=1|x] = PHI(0.5 + 1.0*x) and x ~ N(0,1)
. set obs 100
Number of observations (_N) was 0, now 100.

. set seed 1234567

. generate x = rnormal(0,1)

. generate ystar = 0.5 + 1*x + rnormal(0,1)

. generate y = (ystar > 0)

. generate cons = 1   // Mata code below requires a regressor for the intercept

. summarize

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
           x |        100   -.1477064    1.003931  -2.583632   2.350792
       ystar |        100    .2901163     1.46373  -3.372719   3.316435
           y |        100         .59    .4943111          0          1
        cons |        100           1           0          1          1
```

Probit MLE

```
. * Estimate probit model by MLE
. probit y x, nolog

Probit regression                                       Number of obs =    100
                                                        LR chi2(1)    =  42.67
                                                        Prob > chi2   = 0.0000
Log likelihood = -46.350193                             Pseudo R2     = 0.3152

------------------------------------------------------------------------------
           y | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
           x |   1.137895   .2236915     5.09   0.000     .6994677    1.576322
       _cons |   .4810185   .1591173     3.02   0.003     .1691543    .7928827
------------------------------------------------------------------------------
```

Stata estimation using bayesmh with flat prior

- bayesmh y x, likelihood(probit) prior({y:_cons x}, flat) rseed(10101)

```
Bayesian probit regression                        MCMC iterations  =     12,500
Random-walk Metropolis-Hastings sampling          Burn-in          =      2,500
                                                  MCMC sample size =     10,000
                                                  Number of obs    =        100
                                                  Acceptance rate  =      .2081
                                                  Efficiency:  min =     .09261
                                                               avg =       .104
Log marginal-likelihood = -47.855029                           max =      .1154
```

y	Mean	Std. dev.	MCSE	Median	Equal-tailed [95% cred. interval]	
x	1.172479	.2315767	.006817	1.155511	.7693384	1.644086
_cons	.4912771	.1649868	.005421	.4913284	.1694699	.8135934

Code up instead

- Code up example in Mata with
  - uninformative (flat) prior
  - random walk MH with $\boldsymbol{\beta}^* = \boldsymbol{\beta}^{(s-1)} + \mathbf{v}$ where $\mathbf{v}$ is draw from $\mathcal{N}[\mathbf{0}, 0.25\mathbf{I}]$
    - $c = 0.25$ chosen after some trial and error
  - First $10,000$ MH draws are discarded (burn-in)
  - Next $10,000$ draws are kept.

Core Mata code

```
for (irep=1; irep<=20000; irep++) {
   bcandidate = bdraw + 0.25*rnormal(k,1,0,1)   // bdraw is previous value of b
   probitprob = normal(X*bcandidate)
   lpostcandidate = e'( y:*ln(probitprob) + (e-y):*ln(e-probitprob)   // e = J(n,1,1)
   laccprob = lpostcandidate - lpostdraw   // lpostdraw post. prob. from last round
   accept = 0
   if ( ln(runiform(1,1)) < laccprob ) {
      lpostdraw = lpostcandidate
      bdraw = bcandidate
     accept = 1
   }
   // Store the draws after burn-in of b
   if (irep>10000) {
   j = irep-10000
   b_all[.,j] = bdraw // These are the posterior draws
   }
}
```

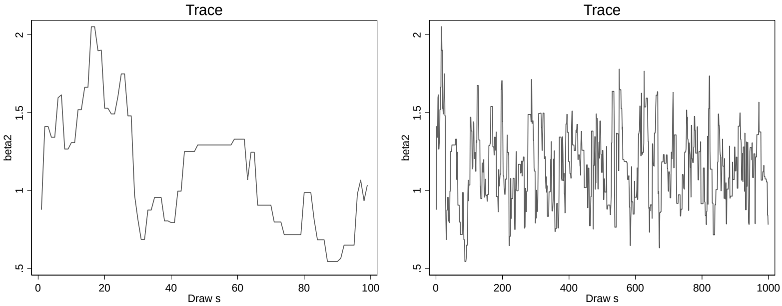
Results for slope parameter

- Posterior mean is 1.168 versus bayesmh 1.172 and MLE 1.137
- Posterior stand. dev. 0.226 versus bayesmh 0.232 and MLE 0.224
- A 95% percent Bayesian credible interval for β_2 is $(0.746, 1.631)$.

```
. * Analyze the posterior draws from probit MH algorithm
. summarize beta* accept

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
       beta1 |     10,000    .4868512    .1573034  -.0620936   1.173329
       beta2 |     10,000    1.167617    .2263273   .4205724   2.051684
      accept |     10,000       .4302    .4951287          0          1

. centile beta2, centile(2.5, 97.5)

                                                         Binom. interp.
    Variable |       Obs  Percentile    Centile        [95% conf. interval]
-------------+-------------------------------------------------------------
       beta2 |    10,000        2.5    .7456415        .7262759    .7578553
             |                 97.5    1.631214        1.621459     1.63932
```

Correlated draws

- The first 100 and first 1,000 draws (after burn-in) from the posterior density of β_2



- Flat sections are where the candidate draw was not accepted
  - The acceptance rate for 10,000 draws was 0.4302.

- Correlations of the 10,000 draws of β_2 die out reasonably quickly

```
. * Compute the first 12 autocorrelations of beta2
. corrgram beta2, lags(12)

                                          -1       0       1 -1       0       1
 LAG       AC       PAC      Q     Prob>Q  [Autocorrelation]  [Partial autocor]
-------------------------------------------------------------------------------
1        0.8282   0.8283   6861.7  0.0000          |------            |------
2        0.6848  -0.0044    11553  0.0000          |-----             |
3        0.5715   0.0167    14821  0.0000          |----              |
4        0.4747  -0.0070    17076  0.0000          |---               |
5        0.3917  -0.0083    18611  0.0000          |---               |
6        0.3247   0.0041    19666  0.0000          |--                |
7        0.2671  -0.0068    20380  0.0000          |--                |
8        0.2169  -0.0083    20850  0.0000          |-                 |
9        0.1762   0.0005    21161  0.0000          |-                 |
10       0.1444   0.0035    21370  0.0000          |-                 |
11       0.1174  -0.0023    21508  0.0000          |                  |
12       0.0922  -0.0109    21593  0.0000          |                  |
```

- This varies a lot with choice of c in $\boldsymbol{\beta}^* = \boldsymbol{\beta}^{(s-1)} + \mathcal{N}[\mathbf{0}, c\mathbf{I}]$

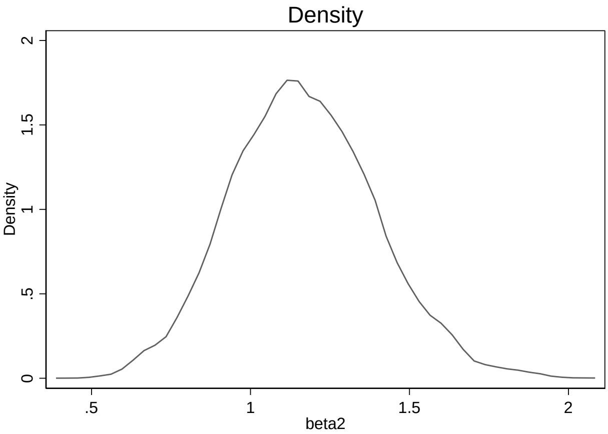
Efficiency statistic

- This is $1/\{1+2\sum_{k=1}^{\max}\widehat{\rho}_k^2\}$ where $\widehat{\rho}_j = Cor[\widehat{\beta},\widehat{\beta}_{-s}]$.
- Efficiency statistic is 0.1829
  - so 10,000 correlated draws equivalent to 1,827 independent draws.

```
. * Compute the efficiency of the MH algorithm for beta2
. qui ac beta2, lags(100) gen(ac)

. qui generate ac_sq = ac^2

. qui summarize ac_sq

. di "Efficiency = " 1/(1+2*r(sum))
Efficiency = .18277753
```

Posterior density

- Kernel density estimate of the 10,000 draws of β_2
  - centered around approx. 1.2 with standard deviation of 0.2 or so.



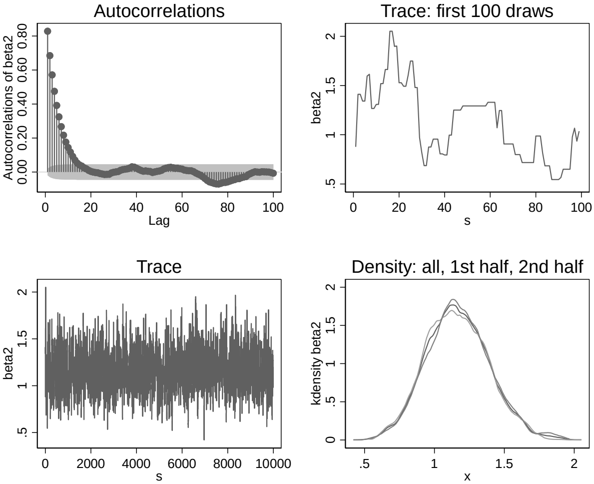
Various diagnostics

- Stata code

```
. * Plot various diagnostics for the posterior draws of b2
. qui ac beta2, title("Autocorrelations") lags(100)        ///
>     note(" ", ring(0) pos(3)) saving(graph1.gph, replace)

. qui line beta2 s if s < 100, title("Trace: first 100 draws") ///
>     saving(graph2.gph, replace)

. qui line beta2 s, title("Trace") saving(graph3.gph, replace)

. qui graph twoway (kdensity beta2) (kdensity beta2 if s<=5000)          ///
>     (kdensity beta2 if s>5000), title("Density: all, 1st half, 2nd half") ///
>     legend(off) note(" ", ring(0) pos(3)) saving(graph4.gph, replace)

. graph combine graph1.gph graph2.gph graph3.gph graph4.gph, ///
>     iscale(0.7) ysize(5) xsize(6) rows(2)
```

Various diagnostics

- Graph



3. Data Augmentation: Summary

- Latent variable models (probit, Tobit, ...) observe $y_1, ..., y_N$ based on latent variables $y_1^*, ..., y_N^*$.
- Bayesian data augmentation introduces $y_1^*, ..., y_N^*$ as additional parameters
  - then posterior is $p(y_1^*, ..., y_N^*, \boldsymbol{\theta})$
  - which includes $p(\boldsymbol{\theta})$ the desired posterior for $\boldsymbol{\theta}$.
- Use Gibbs sampler
  - alternating draws between $p(\boldsymbol{\theta}|y_1^*, ..., y_N^*)$ and $p(y_1^*, ..., y_N^*|\boldsymbol{\theta})$.
- Draws of $\boldsymbol{\theta}|y_1^*, ..., y_N^*$ can use known results for linear regression
  - since regular regression once $y_1^*, ..., y_N^*$ are known
- Draws from $p(y_1^*, ..., y_N^*|\boldsymbol{\theta})$ are called **data augmentation**
  - since we augment observed $y_1, ..., y_N$ with unobserved $y_1^*, ..., y_N^*$.

Probit Model

- Likelihood: Probit model with latent variable formulation
  - $y_i^* = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i,\ \varepsilon_i \sim \mathcal{N}[0,1]$.
  - $y_i = \begin{cases} 1 & y_i^* > 0 \\ 0 & y_i^* \leq 0 \end{cases}$
- Since $y_i^*|\mathbf{x}_i, \boldsymbol{\beta} \sim N[\mathbf{x}_i'\boldsymbol{\beta}, 1]$ we have
  - $p(\mathbf{y}^*|\boldsymbol{\beta}, \mathbf{X}) = -\frac{N}{2}\ln(2\pi) - \frac{1}{2}\sum_{i=1}^{n}(y_i^* - \mathbf{x}_i'\boldsymbol{\beta})^2$.
- Prior: uniform prior (all values equally likely)
  - $\pi(\boldsymbol{\beta}) = 1$
- We have "parameters" $\boldsymbol{\beta}, \mathbf{y}^*$ and data $(\mathbf{y}, \mathbf{X})$
  - posterior is $p(\boldsymbol{\beta}, \mathbf{y}^*|\mathbf{y}, \mathbf{X})$.
- We will make alternating draws from
  - $p(\boldsymbol{\beta}|\mathbf{y}^*, \mathbf{y}, \mathbf{X})$ and $p(\mathbf{y}^*|\boldsymbol{\beta}, \mathbf{y}, \mathbf{X})$
  - data want the posterior density $p(\boldsymbol{\beta}, \mathbf{y}^*|\mathbf{y}, \mathbf{X})$

Probit Model (continued)

- Make alternating draws from $p(\boldsymbol{\beta}|\mathbf{y}^*, \mathbf{y}, \mathbf{X})$ and $p(\mathbf{y}^*|\boldsymbol{\beta}, \mathbf{y}, \mathbf{X})$
- $p(\boldsymbol{\beta}|\mathbf{y}^*, \mathbf{y}, \mathbf{X}) = p(\boldsymbol{\beta}|\mathbf{y}^*, \mathbf{X})$ since knowledge of $\mathbf{y}^* \Rightarrow \mathbf{y}$ is known
  - the normal distribution is symmetric in the mean
  - since $\mathbf{y}^*|\mathbf{X}\boldsymbol{\beta} \sim N[\mathbf{X}\boldsymbol{\beta}, \mathbf{I}]$ we have $\mathbf{X}\boldsymbol{\beta}|\mathbf{y}^* \sim N[\mathbf{y}^*, \mathbf{I}]$
    - ★ density is $(2\pi)^{-N/2} \exp\{(\mathbf{y}^* - \mathbf{X}\boldsymbol{\beta})(\mathbf{y}^* - \mathbf{X}\boldsymbol{\beta})/2\}$
  - so $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta}|\mathbf{y}^* \sim N[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}^*, \ (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{I}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}]$
  - or $\boldsymbol{\beta}|\mathbf{y}^* \sim N[\widetilde{\boldsymbol{\beta}}, \ (\mathbf{X}'\mathbf{X})^{-1}]$ where $\widetilde{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$.
- $p(\mathbf{y}^*|\boldsymbol{\beta}, \mathbf{y}, \mathbf{X}) = p(y_1^*, ..., y_N^*|\boldsymbol{\beta}, \mathbf{y}, \mathbf{X})$ is truncated normal so
  -
    - ★ If $y_i = 1$ draw from $\mathcal{N}[\mathbf{x}_i'\boldsymbol{\beta}, 1]$ left truncated at 0
    - ★ If $y_i = 0$ draw from $\mathcal{N}[\mathbf{x}_i'\boldsymbol{\beta}, 1]$ right truncated at 0
- Posterior step: draw $\boldsymbol{\beta}^{(s)}$ from $p(\boldsymbol{\beta}|y_1^{*(s-1)}, ..., y_N^{*(s-1)}, \mathbf{X})$
  Augmentation step: draw $y_1^{*(s)}, ..., y_N^{*(s)}$ from $p(y_1^*, ..., y_N^*|\boldsymbol{\beta}^{(s)}, \mathbf{y}, \mathbf{X})$.

Core Mata code

```
for (irep=1; irep<=s; irep++) {
   // Posterior-step: draw from beta | y* ~N[bols*, (X'X)^-1]
  bols = Xtxinvchol*cross(X,ystar)
  b1 = bols
  bdraw = b1 + Xtxinvchol*rnormal(k,1,0,1) //invnormal(uniform(k,1))
  // Imputation step: make one draw of vector ystar
  // where for ith observation ystar_i | y,b is truncated normal
  for (i=1; i<=n; i++) {
    mu = X[i,.]*bdraw
     if (y[i,1]==0) {
       uright = normal(-mu)*uniform(1,1)
       ystar[i,1] = mu + invnormal(uright)
     }
    else {
      uleft = normal(-mu) + (1-normal(-mu))*uniform(1,1)
     ystar[i,1] = mu + invnormal(uleft)
    }
```

Results for slope parameter

- Posterior mean is 1.174 versus bayesmh 1.172 and MLE 1.137
- Posterior stand. dev. 0.226 versus bayesmh 0.232 and MLE 0.224
- A 95% percent Bayesian credible interval for β_2 is $(0.7516, 1.645)$.

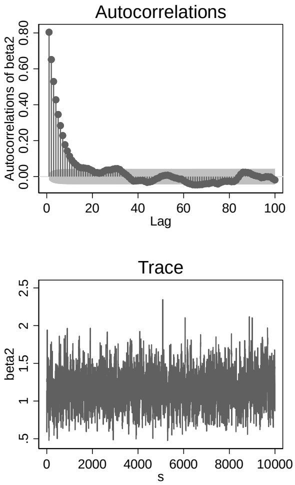
```
. * Analyze the posterior draws from probit Gibbs sampler algorithm
. summarize beta*

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
       beta1 |     10,000    .4842779    .1582509  -.2082249   1.158187
       beta2 |     10,000    1.173542    .2259341   .4758451   2.346783

. centile beta2, centile(2.5, 97.5)

                                                          Binom. interp.
    Variable |       Obs  Percentile    Centile        [95% conf. interval]
-------------+-------------------------------------------------------------
       beta2 |    10,000        2.5    .7512463        .7392763    .7636364
             |                 97.5    1.646951        1.629367    1.660316
```

Efficiency statistic

- This is $1/\{1+2\sum_{k=1}^{max}\widehat{\rho}_k^2\}$ where $\widehat{\rho}_j = Cor[\widehat{\beta}, \widehat{\beta}_{-s}]$.
- Efficiency statistic is 0.205

```
. * Compute the efficiency of the Gibbs sampler algorithm
. generate s= _n

. tsset s

Time variable: s, 1 to 10000
        Delta: 1 unit

. qui ac beta2, lags(100) gen(ac)

. qui generate ac_sq = ac^2

. qui summarize ac_sq

. di "Efficiency = " 1/(1+2*r(sum))
Efficiency = .20486695
```

Diagnostics plot



More complicated example: Multinomial probit

- Likelihood: Multinomial probit model (MLE has high-dimensional integral)
  - $U_{ij}^* = \mathbf{x}_{ij}'\boldsymbol{\beta} + \varepsilon_{ij}$, $\boldsymbol{\varepsilon}_i \sim \mathcal{N}[\mathbf{0}, \Sigma_\varepsilon]$
  - $y_{ij} = 1$ if $U_{ij}^* > U_{ik}^*$ all $k \neq j$
- Prior for $\boldsymbol{\beta}$ and Σ_ε^{-1} may be normal and Wishart
- Data augmentation
  - Latent utilities $\mathbf{U}_i = (U_{i1}, ..., U_{im})$ are introduced as auxiliary variables
  - Let $\mathbf{U} = (\mathbf{U}_1, ..., \mathbf{U}_N)$ and $\mathbf{y} = (y_1, ..., y_N)$
- Gibbs sampler for joint posterior $p(\boldsymbol{\beta}, \mathbf{U}, \Sigma_\varepsilon | \mathbf{y}, \mathbf{X})$ cycles between
  - 1. Conditional posterior for $\boldsymbol{\beta} | \mathbf{U}, \Sigma_\varepsilon, \mathbf{y}, \mathbf{X}$
  - 2. Conditional posterior for $\Sigma_\varepsilon | \boldsymbol{\beta}, \mathbf{U}, \mathbf{y}, \mathbf{X}$, and
  - 3. Conditional posterior for $\mathbf{U}_i | \boldsymbol{\beta}, \Sigma_\varepsilon, \mathbf{y}, \mathbf{X}$.
- Albert and Chib (1993) provide a quite general treatment.
- McCulloch and Rossi (1994) provide a substantive MNP application.

4. Multiple Imputation: Missingness Mechanisms

- Define data $\mathbf{W} = \mathbf{W}_{obs} \cup \mathbf{W}_{miss}$ and selection matrix $\mathbf{S}$

$\mathbf{W}$	$N \times p$	Complete data
$\mathbf{W}_{obs}$		Observed data
$\mathbf{W}_{miss}$		Missing data
$\mathbf{S}$	$N \times p$	Selection matrix of 1's and 0's

- Missing data mechanisms

MCAR	$\Pr(\mathbf{S}\vert\mathbf{W}_{obs},\mathbf{W}_{miss}) = \Pr(\mathbf{S})$	Missing completely at ran
MAR	$\Pr(\mathbf{S}\vert\mathbf{W}_{obs},\mathbf{W}_{miss}) = \Pr(\mathbf{S}\vert\mathbf{W_{obs}})$	Missing at random
MNAR	$\Pr(\mathbf{S}\vert\mathbf{W}_{obs},\mathbf{W}_{miss}) \neq \Pr(\mathbf{S}\vert\mathbf{W_{obs}})$	Missing not at random

Missingness Mechanisms

- MCAR (Missing completely at random) $\Pr(\mathbf{S}|\mathbf{W}_{obs}, \mathbf{W}_{miss}) = \Pr(\mathbf{S})$
  - can validly use only nonmissing data
  - but imputation can improve efficiency
- MAR (Missing at random) $\Pr(\mathbf{S}|\mathbf{W}_{obs}, \mathbf{W}_{miss}) = \Pr(\mathbf{S}|\mathbf{W_{obs}})$
  - regression case where only missing data is exogenous regressors
  - then can use e.g. case deletion
  - but imputation can improve efficiency
- MNAR (Missing not at random) $\Pr(\mathbf{S}|\mathbf{W}_{obs}, \mathbf{W}_{miss}) \neq \Pr(\mathbf{S}|\mathbf{W_{obs}})$
  - regression case where data on endogenous regressors or the dependent variable is missing
  - then problems - selection on unobservables
  - much stronger stochastic assumptions will be needed e.g. tobit model.

Imputation Methods

- Simple methods such as replace missing values with nonmissing mean values can lead to inconsistent estimation and invalid statistical inference.
- Single imputation uses a model to obtain imputation $\mathbf{W}_{miss\_imp}$ of $\mathbf{W}_{miss}$ and form complete data $\mathbf{W}_{imp} = (\mathbf{W}_{obs}, \mathbf{W}_{miss\_imp})$.
  - standard inference on consequent $\widehat{\theta}$ obtained using $\mathbf{W}_{imp}$ is invalid as it ignores the additional randomness in imputing $\mathbf{W}_{miss\_imp}$.

Multiple Imputation

- Multiple imputation accounts for this by imputing $\mathbf{W}_{miss\_imp}$ m independent times.
  - impute $\mathbf{W}_{miss\_imp,r}$, $r = 1, ...m$, leading to $\mathbf{W}_{imp,r}$, $r = 1, ..., m$.
  - obtain $\widehat{\boldsymbol{\theta}}_r$, $r = 1, .., m$, with usual variance estimates $\widehat{\mathbf{V}}_r$, $r = 1, .., m$
- $\widehat{\boldsymbol{\theta}}$ is the average of the m $\widehat{\boldsymbol{\theta}}_r$s

$$\widehat{\boldsymbol{\theta}} = \frac{1}{m} \sum\nolimits_{r=1}^{m} \widehat{\boldsymbol{\theta}}_r.$$

- $\widehat{\mathrm{Var}}(\widehat{\boldsymbol{\theta}})$ is the average of the m $\widehat{\mathbf{V}}_r$ plus the variability in the m $\widehat{\boldsymbol{\theta}}_r$s around $\widehat{\boldsymbol{\theta}}$

$$\widehat{\mathrm{Var}}(\widehat{\boldsymbol{\theta}}) = \frac{1}{m} \sum\nolimits_{r=1}^{m} \widehat{\mathbf{V}}_r + \frac{1 + (1/m)}{m-1} \sum\nolimits_{r=1}^{m} (\widehat{\boldsymbol{\theta}}_r - \widehat{\boldsymbol{\theta}})(\widehat{\boldsymbol{\theta}}_r - \widehat{\boldsymbol{\theta}})'.$$

Regression-based Imputation

- Example where a single regressor x, a count, has some missing values.
- Data $\mathbf{W} = (\mathbf{x}, \mathbf{Z})$
  - with observed and missing $\mathbf{x} = (\mathbf{x}_o, \mathbf{x}_m)$ and $\mathbf{Z} = (\mathbf{Z}_o, \mathbf{Z}_m)$.
- Imputation
  - Assume $x_i \sim \text{Poisson}(\mathbf{z}_i' \boldsymbol{\beta})$
  - Estimate $\widehat{\boldsymbol{\beta}}$ and robust $\widehat{\mathrm{V}}(\widehat{\boldsymbol{\beta}})$ from Poisson regression of $\mathbf{x}_o$ on $\mathbf{Z}_o$ (uses observed data)
  - Draw $\boldsymbol{\beta}^*$ from $N[\widehat{\boldsymbol{\beta}}, \widehat{\mathrm{V}}(\widehat{\boldsymbol{\beta}})]$ and for each missing observation x_{mi} make draw from $\text{Poisson}(\mathbf{z}_{mi}' \boldsymbol{\beta}^*)$.
- For multiple imputation make more draws.
- Since x is a regressor $\mathbf{z}$ should include y (as x and y are then related).
- The MAR assumption is made (fine if y is completely observed).

Imputation by Data Augmentation

- Partition $\mathbf{w} = (\mathbf{x}, \mathbf{z})$
  - where $\mathbf{z}$ is completely observed and $\mathbf{x}$ is partially observed.
- Assume $\mathbf{x}_i \sim N[\Pi \mathbf{z}_i, \Sigma]$ where Π and Σ are unknown.
  - may need to first transform transform variables such as right-skewed.
- Goal is impute $\mathbf{X}_m$ given $\mathbf{X}_o$ and $\mathbf{Z}$ where partition $\mathbf{X} = (\mathbf{X}_o, \mathbf{X}_m)$.
- Use data augmentation where $\mathbf{X}_m$ are additional parameters
  - posterior density $p(\Pi, \Sigma, \mathbf{X}_m | \mathbf{X}_o, \mathbf{Z})$
  - use Gibbs sampler
    - ★ P step (posterior step) $p(\Pi, \Sigma | \mathbf{X}_m, \mathbf{X}_o, \mathbf{Z})$
    - ★ I step (imputation step) $p(\mathbf{X}_m | \Pi, \Sigma, \mathbf{X}_o, \mathbf{Z})$
  - uniform prior for Π and Wishart prior for Σ.
  - use the draws $\mathbf{X}_m^*$ and discard the draws of Π, Σ.

Example Stata code - variable x2 has some missing values

```
* Declare dataset type (long) and summarize missingness
mi set mlong
mi misstable summarize
mi misstable patterns
* Register imputed variables and perform ten imputations
mi register imputed x2
mi register regular y x3
mi impute mvn x2 = y x3, add(10) rseed(10101) burnin(100) burnbetween(100)
* Multiple imputation creates the following data and variables
summarize
* Describe the imputed data
mi describe
* Estimate the model with imputed data
mi estimate, dots: regress y x2 x3
```

5. Some References

- Chapter 13 "Bayesian Methods" in A. Colin Cameron and Pravin K. Trivedi, Microeconometrics: Methods and Applications, Cambridge University Press.
- Chapter 30 "Bayesian Methods: MCMC Algorithms" in A. Colin Cameron and Pravin K. Trivedi, Microeconometrics using Stata, Second edition, forthcoming.
- Bayesian books by econometricians that feature MCMC are
  - Geweke, J. (2003), Contemporary Bayesian Econometrics and Statistics, Wiley.
  - Koop, G., Poirier, D.J., and J.L. Tobias (2007), Bayesian Econometric Methods, Cambridge University Press.
  - Koop, G. (2003), Bayesian Econometrics, Wiley.
  - Lancaster, T. (2004), Introduction to Modern Bayesian Econometrics, Wiley.
- Most useful (for me) book by statisticians
  - Gelman, A., J.B. Carlin, H.S. Stern, D.B. Dunson, A. Vehtari and D.B. Rubin (2013), Bayesian Data Analysis, Third Edition, Chapman & Hall/CRC.