STAT714 Final Project

Statistical Inference in Missing Data by MCMC and Non-MCMC Multiple Imputation Algorithms:

Assessing the Effects of Between-Imputation Iterations

(Takahashi, 2017)

2020021218 이소연
2020020338 전지현
2020020316 황서진

Assumptions of Imputation Methods

1. Missing Data Mechanism

-MCAR : $f\left(m_i \middle| y_i, \phi\right) = f\left(m_i \middle| y_i^*, \phi\right)$ for all i.

-MAR : $f\left(m_i \middle| y_{i(0)}, y_{i(1)}, \phi\right) = f\left(m_i \middle| y_{i(0)}, y_{i(1)}^*, \phi\right)$ for all i

-NMAR : $f\left(m_i \middle| y_{i(0)}, y_{i(1)}, \phi\right) \neq f\left(m_i \middle| y_{i(0)}, y_{i(1)}^*, \phi\right)$ for some i.

2. Ignorability

If both of conditions satisfied:

(1) MAR condition

(2) the distinctness condition

Assumptions of Imputation Methods

3. Proper Imputation

- Confidence proper : Simplified version of proper imputation

  by van Buuren(2012: 39)

Incomplete Sample	Complete Sample	Population
Y_{obs}	$Y = (Y_{\text{obs}}, Y_{\text{mis}})$	
$\bar{Q}$	$\hat{Q}$	Q
$\bar{U}$	$U \doteq V(\hat{Q})$	
$B \doteq V(\bar{Q})$		

confidence proper if …:

$$
\begin{aligned}
E(\bar{Q}|Y) &= \hat{Q} \\
E(\bar{U}|Y) &= U \\
\left(1 + \frac{1}{m}\right) E(B|Y) &\geq V(\bar{Q})
\end{aligned}
$$

MI Algorithms

1. Data Augmentation (DA)

-It is hard to judge the convergence in MCMC because its convergence is stochastic

-R package : norm2 'mcmcNorm' (MCMC for incomplete multivariate normal data)

Step 1) Imputation step: generates imputed values from the predictive distribution of missing values, given the observed values and the parameter values at iteration $\boldsymbol{t}$

$$Y_{(1)}^{(t+1)} \sim p\left(Y_{(1)} \middle| Y_{(0)}, \theta^{(t)}\right)$$

Step 2) Posterior step: generates parameter values from the posterior distribution, given the observed values and the imputed values at iteration $\boldsymbol{t+1}$

$$\theta^{(t+1)} \sim p(\theta | Y_{(0)}, Y_{(1)}^{(t+1)})$$

MI Algorithms

2. Fully Conditional Specification (FCS)

-Allows for the creation of flexible multivariate models

-R package : MICE 'mice' (Multivariate Imputation By Chained Equations)

Step 1) Imputations based on a set of conditional distributions for each variable on the other variables, one at a time.

$$y_{j(1)}^{(t+1)} \sim p\left(y_{j(1)} \middle| y_{(0)}, y_{1(1)}^{(t+1)}, \dots, y_{j-1(1)}^{(t+1)}, y_{j+1(1)}^{(t)}, \dots, y_{K(1)}^{(t)}, x\right) \qquad \mathrm{j} = 1, \dots \mathrm{K}$$

Step 2) Draws $\theta_j^{(t+1)}$ given the observed values and the $(\boldsymbol{t}+\mathbf{1})$th imputations

$$\theta_j^{(t+1)} \sim p(\theta_j | y_{(0)}, y_{1(1)}^{(t+1)}, \dots, y_{K(1)}^{(t+1)}, x)$$

MI Algorithms

3. Expectation-Maximization with Bootstrapping (EMB)

-EM algorithm is applied to each of the M bootstrap resamples to refine M point estimates of parameter θ.

-R package : Amelia 'amelia' (Uses a bootstrap(default=100)+EM algorithm)

Step 1) Expectation step: calculates the Q-function by averaging the complete-data log-likelihood over the predictive distribution of missing data.

$$Q\left(\theta\middle|\theta^{(t)}\right) = \int l(\theta|y) f(Y_{(0)}|Y_{(1)}, \theta^{(t)}) dY_{(0)}$$

Step 2) Maximization step: finds parameter values at iteration t + 1 by maximizing the Q-function.

$$\theta^{(t+1)} = \underset{\theta}{\operatorname{argmax}}\, Q\left(\theta\middle|\theta^{(t)}\right)$$

MI Algorithms

Between-imputation Iterations

the number of times the imputation process is iterated between saving one complete dataset to memory and the next (e.g. between m = 1 and m = 2)

	Joint Modeling	Conditional Modeling	Between-imputation Iterations required
MCMC	DA	FCS	O
Non-MCMC	EMB		X

1. Data Augmentation (DA)
2. Fully Conditional Specification (FCS)
3. Expectation-Maximization with Bootstrapping (EMB)

실험 설계

Hyperparameters from Meta-Analysis

-Literatures that compared imputation methods

[Table] Summary of the 20 Studies on Multiple Imputation

Authors	MI Algorithms	Sample Size	Number of Variables	Number of Imputations	Number of Iterations	Missing Rate
Barnard and Rubin (1999)	DA	10, 20, 30	2	3, 5, 10	Unknown	10%, 20%, 30%
Hardt, Herke, and Leonhart (2012)	**DA, EMB, FCS**	50, 100, 200	3, 13, 23, 43, 83	20	**Unknown**	20%, 50%
⋮						
McNeish (2017)	DA, FCS	20, 50, 100, 250	4	5, 25, 100	Unknown	10%, 20%, 30%, 50%

- N = 1000 (Sample size)
- p = 10 (Number of variables)
- M = 5(Number of Imputations)
- T: the number of EMB iteration, doubling for DA and FCS

실험 설계

Monte Carlo Simulated Data (Theoretical)

MC 시뮬레이션 Run 횟수는 500
관측치 수 = 1000

설명변수 X 의 개수 p-1는 1, …, 9으로 변화하며, multivariate normal을 따름
→ $X \sim N_{p-1}(0, 1)$

X 의 공분산(상관계수)행렬은 $Unif(-1, 1)$에서 9^2개의 난수를 거듭 제곱하여 생성
→ r = matrix(runif(9^2,-1,1), ncol=9); Cor<-cov2cor(r%*%t(r))

```
> Cor
             [,1]         [,2]       [,3]        [,4]        [,5]         [,6]        [,7]         [,8]        [,9]
 [1,]  1.000000000  0.001192676  0.4845351 -0.28138331  0.12096665 -0.004177945 -0.16580310 -0.201378893 -0.43339747
 [2,]  0.001192676  1.000000000  0.1711035 -0.51210209 -0.22414282 -0.397915620 -0.72671978 -0.006932479 -0.05650666
 [3,]  0.484535091  0.171103534  1.0000000 -0.38341631  0.06247200 -0.126390475 -0.20427691 -0.171238401 -0.77807966
 [4,] -0.281383307 -0.512102093 -0.3834163  1.00000000 -0.06436935  0.491598014  0.61026175  0.599824987  0.31183735
 [5,]  0.120966646 -0.224142821  0.0624720 -0.06436935  1.00000000 -0.606624257  0.33116141  0.078619688 -0.22061605
 [6,] -0.004177945 -0.397915620 -0.1263905  0.49159801 -0.60662426  1.000000000  0.09940189  0.391326164  0.37946220
 [7,] -0.165803099 -0.726719775 -0.2042769  0.61026175  0.33116141  0.099401894  1.00000000 -0.045070208  0.15252656
 [8,] -0.201378893 -0.006932479 -0.1712384  0.59982499  0.07861969  0.391326164 -0.04507021  1.000000000  0.30701544
 [9,] -0.433397470 -0.056506657 -0.7780797  0.31183735 -0.22061605  0.379462197  0.15252656  0.307015435  1.00000000
```

실험 설계

Monte Carlo Simulated Data (Theoretical)

$$\mathbf{y_i} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1 x_{1i}} + \cdots + \boldsymbol{\beta_{p-1} x_{p-1i}} + \boldsymbol{\varepsilon_i} \quad where \;\; \beta_j \; includes \; \beta_0,$$

$$\beta_j \sim Unif(-2,2), \qquad j = 0, 1, \cdots 10, \qquad \varepsilon_i \sim N(0, \sigma)\,, \qquad \sigma \sim Unif(0.5, 2.0)$$

p = 10 일 때

```
> head(data.frame(y[,9],x))
     y...9. x1         x2          x3          x4         x5          x6          x7         x8         x9        x10
1 -1.4115333  1 -0.1017198  0.12778270  0.44936004 -0.6612612 -0.41693025 -0.01227498 -0.2568702 -0.7493783 -0.6128758
2  0.0263533  1 -0.6035207 -1.02797327 -0.09449877  0.3961243  0.17830756  0.39451484  0.2995638  0.1544598 -0.3923068
3 -4.1944419  1 -2.2970526  0.79491288 -1.13751627  0.1328191  0.45196918 -0.33301380 -0.3743568  1.3601124  1.2041334
4 -2.7578483  1 -0.3620129 -1.03691812 -0.39652645  0.3520109 -0.37513022  1.64731656  0.4177494  0.6462230  1.1177978
5  8.2246782  1 -1.4253012 -0.31026385  0.40620533  1.6409277 -1.17923998  0.80536091  0.9227506  0.5867630 -0.1005743
6 -0.7916779  1 -0.5255071 -0.01947445 -2.24576316  1.3217166  0.07705503  0.18488363  0.6650921  0.8003130  1.7023881
```

```
> beta
 [1] -0.8496899  1.1532205 -0.3640923  1.5320696  1.7618691 -1.8177740  0.1124220  1.5696762  0.2057401 -0.1735411
```

실험 설계

Monte Carlo Simulated Data (Theoretical)

결측 발생 메커니즘: MAR

y_i 는 완전히 관측된 변수이며, x_j 는 다음과 같이 y_i 와 난수 $\boldsymbol{u_{ij}}$에 의존하여 결측치를 가진다.

$$u_{ij} \sim Unif(0,1)$$

$$\boldsymbol{x_{ji}} \begin{cases} \boldsymbol{missing} & if \;\; y_i < median(y_i) \;\; and \; u_{ij} < 0.5 \quad \boldsymbol{or} \;\; y_i > median(y_i) and \; u_{ij} > 0.9 \\ \boldsymbol{not \;\; missing} & otherwise \end{cases}$$

위 메커니즘에 따라 만들어진 데이터는 각 변수 별 약 30%의 결측률을 가진다.

```
Fraction Missing for original variables:
-----------------------------------------

         Fraction Missing
y...9.            0.000
x1                0.292
x2                0.288
x3                0.289
x4                0.319
x5                0.309
x6                0.327
x7                0.294
x8                0.303
x9                0.296
```

실험 설계

Evaluation criteria

- Unbiasedness

$$Bias(\hat{\theta}) = E\left(\hat{\theta}\right) - \theta$$

- Efficiency

$$RMSE(\hat{\theta}) = \sqrt{E(\hat{\theta} - \theta)^2}$$

- Confidence Validity

$$SE(CR) = \sqrt{\frac{CR(1 - CR)}{s}}$$

$$where\ CR = proportion\ of\ simulated\ smaples\ for\ which\ CI\ includes\ \beta_1 (Coverage\ rate)$$

실험 결과

Abbreviations	Missing Data Methods
CD	Complete data without missing values
LD	Listwise deletion (Complete case)
EMB	MI by AMELIA II
DA1	MI by NORM2 with no iterations
DA2	MI by NORM2 with 2*EM iterations
FCS1	MI by MICE with no iterations
FCS2	MI by MICE with 2*EM iterations
S-SI	Stochastic SI by `norm.nob` in MICE

실험 결과

Bias (Theoretical Data)

*red: absolute value over 0.01

β_1	Number of Independent Variables (p-1)								
	1	2	3	4	5	6	7	8	9
Complete	-0.0041	-0.0043	-0.0040	-0.0049	-0.0055	-0.0060	-0.0057	-0.0088	-0.0071
LD	-0.0059	-0.0073	-0.0043	-0.0080	**-0.0110**	-0.0097	**-0.0103**	**-0.0146**	**0.0119**
SSI	-0.0024	-0.0075	-0.0040	-0.0044	-0.0049	0.0027	-0.0032	**-0.0325**	**-0.0859**
EMB	-0.0046	-0.0038	-0.0039	-0.0062	-0.0071	-0.0082	-0.0048	**-0.0486**	**-0.0480**
DA1	-0.0024	-0.0038	-0.0132	**-0.0178**	**-0.0173**	**-0.0207**	**-0.0191**	**-0.0110**	-0.0087
DA2	**-0.0004**	-0.0006	**-0.0006**	**-0.0008**	**0.0001**	**-0.0005**	-0.0003	0.0088	**0.0131**
FCS1	-0.0016	-0.001	**-0.0102**	-0.0095	**-0.0106**	**-0.0115**	**-0.0118**	-0.0053	-0.0067
FCS2	-0.0033	**0.0002**	-0.0014	-0.0019	-0.0024	-0.0020	**-0.0001**	**-0.0050**	**0.0399**

실험 결과

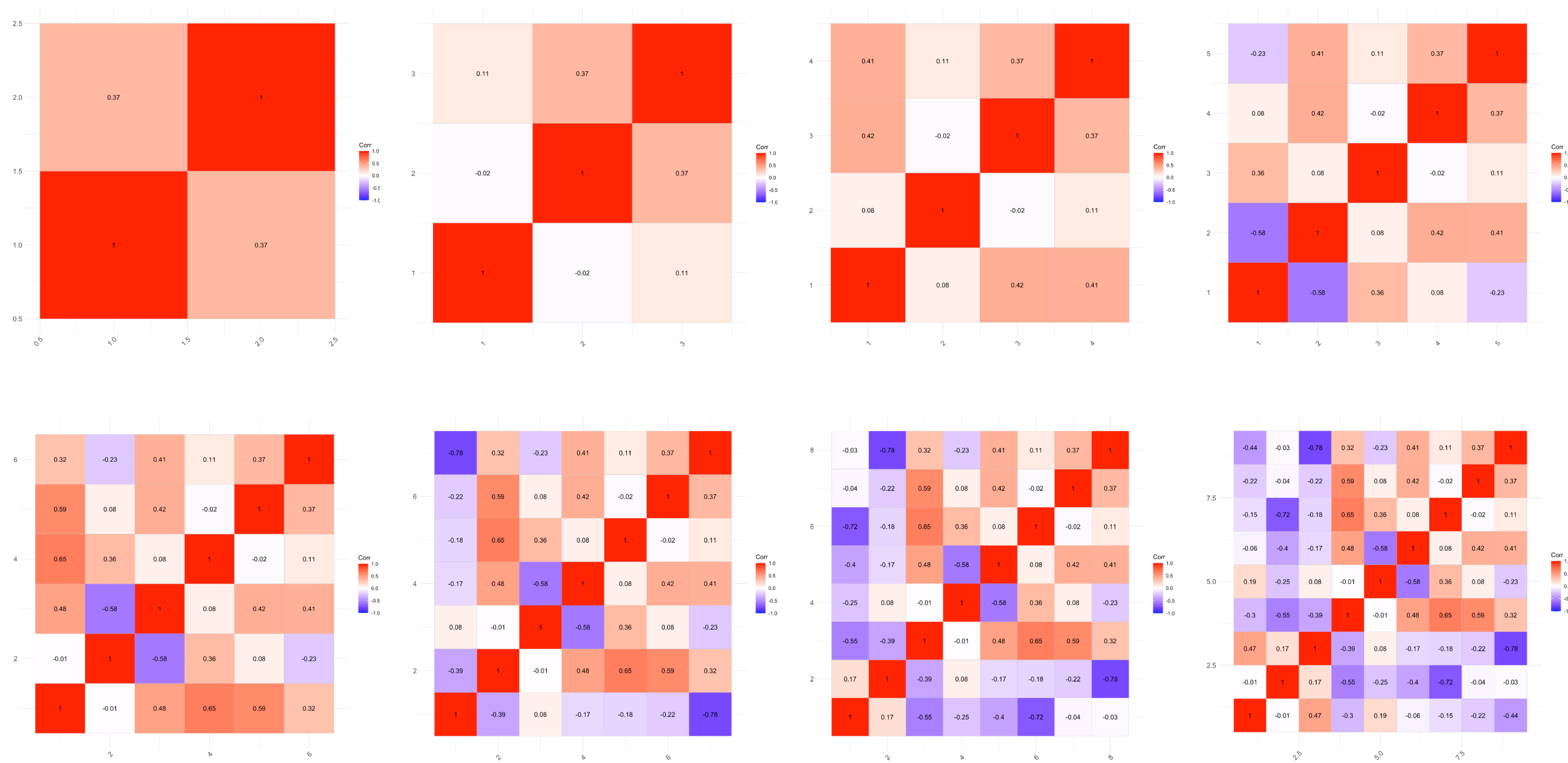
RMSE (Theoretical Data)

*red: absolute value over 0.2

β_1	Number of Independent Variables (p-1)								
	1	2	3	4	5	6	7	8	9
Complete	0.0606	0.0608	0.0714	0.0719	0.0723	0.0754	0.0758	0.1573	**0.3088**
LD	0.0774	0.1189	0.1910	**0.2543**	**0.3000**	**0.3377**	**0.3710**	**0.4850**	**0.7989**
SSI	0.0742	0.0815	0.1081	0.1115	0.1194	0.1353	0.1239	**0.4124**	**0.5199**
EMB	0.0596	0.0670	0.0911	0.1001	0.1071	0.1126	0.1131	**0.5220**	**0.4874**
DA1	0.0884	0.1020	0.1630	0.17970	0.1853	**0.2002**	**0.2019**	**0.2301**	**0.2262**
DA2	**0.0349**	**0.0377**	**0.0475**	**0.0517**	**0.0547**	**0.0559**	**0.0564**	**0.2015**	**0.4773**
FCS1	0.0442	0.0627	0.1174	0.1331	0.1403	0.1616	0.1632	**0.2061**	**0.2076**
FCS2	0.0431	0.0479	0.0638	0.0683	0.0743	0.0782	0.0787	**0.2948**	**0.6016**

실험 결과

Correlation (Theoretical Data)



실험 결과

Coverage (Theoretical Data)

*red: absolute value under 0.9

β_1	Number of Independent Variables (p-1)								
	1	2	3	4	5	6	7	8	9
Complete	0.946	0.942	0.942	0.940	0.942	0.932	0.940	0.940	0.938
LD	0.924	**0.824**	**0.708**	**0.600**	**0.524**	**0.526**	**0.532**	**0.818**	0.906
SSI	**0.882**	**0.82**	**0.766**	**0.752**	**0.730**	**0.670**	**0.754**	**0.430**	**0.352**
EMB	0.950	0.934	0.914	0.912	0.912	0.918	0.908	**0.884**	**0.874**
DA1	0.986	0.974	0.928	0.906	**0.874**	**0.842**	**0.862**	**0.704**	**0.738**
DA2	**0.962**	**0.974**	**0.960**	**0.966**	**0.950**	**0.954**	**0.954**	**0.952**	**0.962**
FCS1	0.944	**0.876**	**0.670**	**0.628**	**0.632**	**0.60**	**0.604**	**0.522**	**0.506**
FCS2	0.934	0.934	0.944	0.936	0.910	0.898	0.930	**0.874**	**0.862**

실험 결과

Lengths of the 95% CI (Theoretical Data)

*red: length too small or large than complete data

β_1	Number of Independent Variables (p-1)								
	1	2	3	4	5	6	7	8	9
Complete	0.2186	0.2188	0.2518	0.2568	0.2575	0.2653	0.2669	0.5601	1.0797
LD	0.2593	0.2992	0.3843	0.4323	0.4741	**0.5321**	**0.5830**	**1.2608**	**2.6525**
SSI	0.2186	0.2194	0.2539	0.2606	0.2622	0.2697	0.2728	0.4416	**0.4907**
EMB	0.2575	0.2752	0.3422	0.3677	0.3872	0.4044	0.4104	**1.6975**	**2.7171**
DA1	**0.5417**	**0.5301**	**0.6125**	**0.6349**	**0.6287**	**0.6315**	**0.6366**	0.6266	**0.6635**
DA2	**0.2488**	**0.2963**	**0.2475**	**0.2653**	**0.2795**	**0.2935**	**0.2963**	**1.0208**	**2.4805**
FCS1	0.2796	0.2089	0.2653	0.2863	0.3087	0.3213	0.3257	**0.3480**	**0.3541**
FCS2	**0.2493**	**0.2938**	**0.2432**	**0.2555**	**0.2703**	**0.2814**	**0.2820**	**0.9266**	**1.8429**

실험 결과

Computation Time in Sec (Theoretical Data)

-intel xeon gold 6230s(2.1GHz CPU) with 384GB RAM

β_1	Number of Independent Variables (p-1)								
	1	2	3	4	5	6	7	8	9
EMB	0.0402	**0.0419**	**0.0450**	**0.0503**	**0.0594**	**0.0756**	**0.1039**	**0.1550**	**0.1758**
DA1	0.0113	0.0115	0.0121	0.0129	0.0139	0.0151	0.0166	0.0181	0.0197
DA2	**0.0227**	0.0452	0.0774	0.1227	0.1922	0.2912	0.4316	0.6342	0.8843
FCS1	0.0175	0.0322	0.0492	0.0700	0.0882	0.1110	0.1346	0.1611	0.1900
FCS2	0.1144	0.4392	**1.0894**	**2.2967**	**4.0312**	**6.5446**	**10.0143**	**13.724**	**18.2476**

결론

실제 실험 결과 ≈ 논문 결과

-Bias, RMSE, Coverage: DA2의 성능이 가장 우수

-CI length: DA2, FCS2의 성능이 가장 우수

-Computation time: EMB가 가장 우수

→ DA와 FCS는 between-imputation iteration이 없을 경우 성능 저하 심각

→ 효과 면에서는 DA2와 FCS2가 EMB보다 뛰어났으나, 계산 시간 면에서 between-imputation iteration이 없는 EMB가 두드러지게 우수했으며 효과 차이는 용인할 수 있는 수준

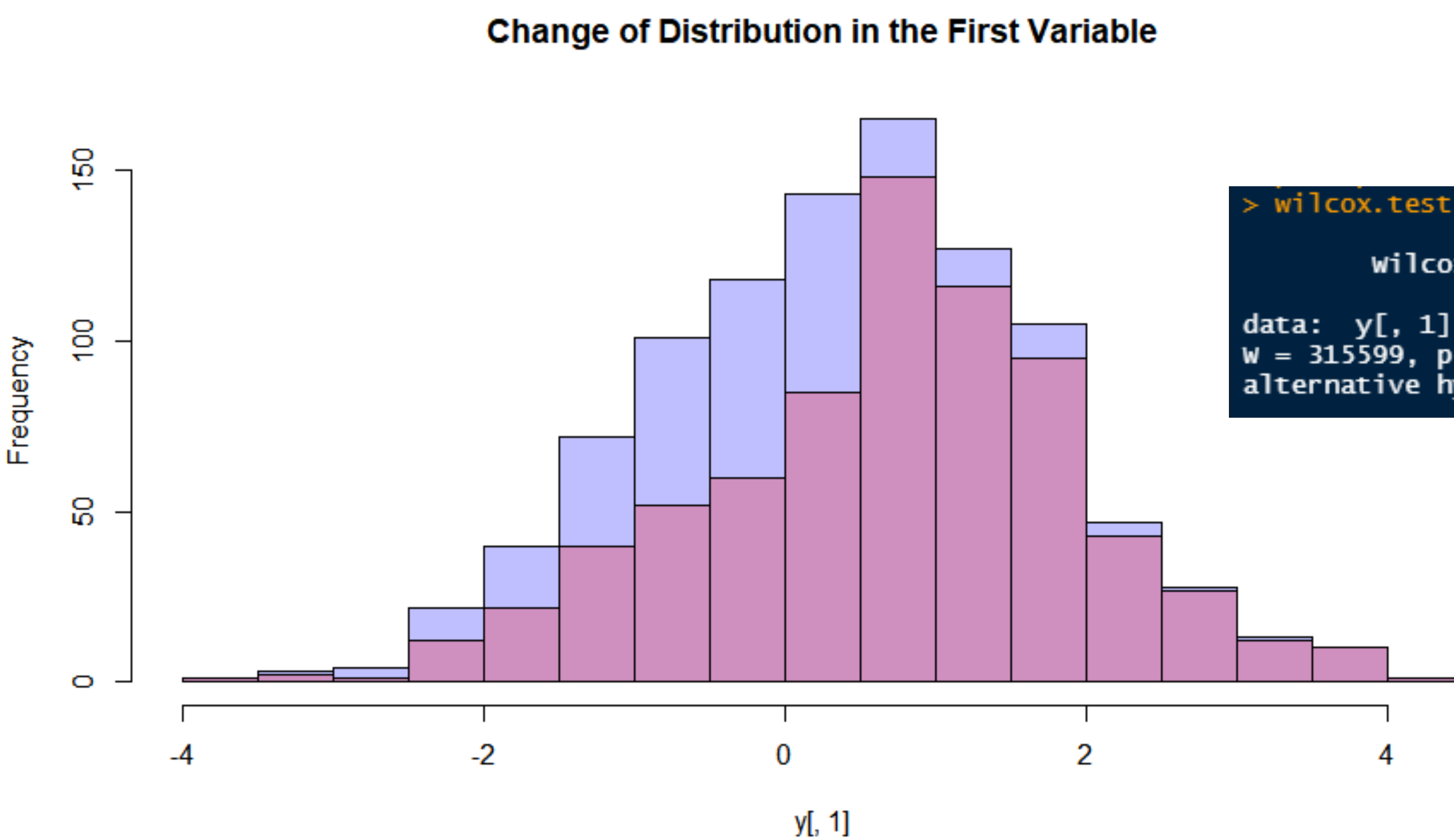
→ EMB: confidence proper without between-imputation iteration

→ Single imputation(stochastic regression): coverage rate가 낮고 CI length 과소추정

감사합니다 !

부록

Check MAR Assumption of the Simulated Data




```
> wilcox.test(y[,1], y[-which(is.na(X_mis[,2])),1])

        Wilcoxon rank sum test with continuity correction

data:  y[, 1] and y[-which(is.na(X_mis[, 2])), 1]
W = 315599, p-value = 2.846e-06
alternative hypothesis: true location shift is not equal to 0
```

부록

Monte Carlo Simulation

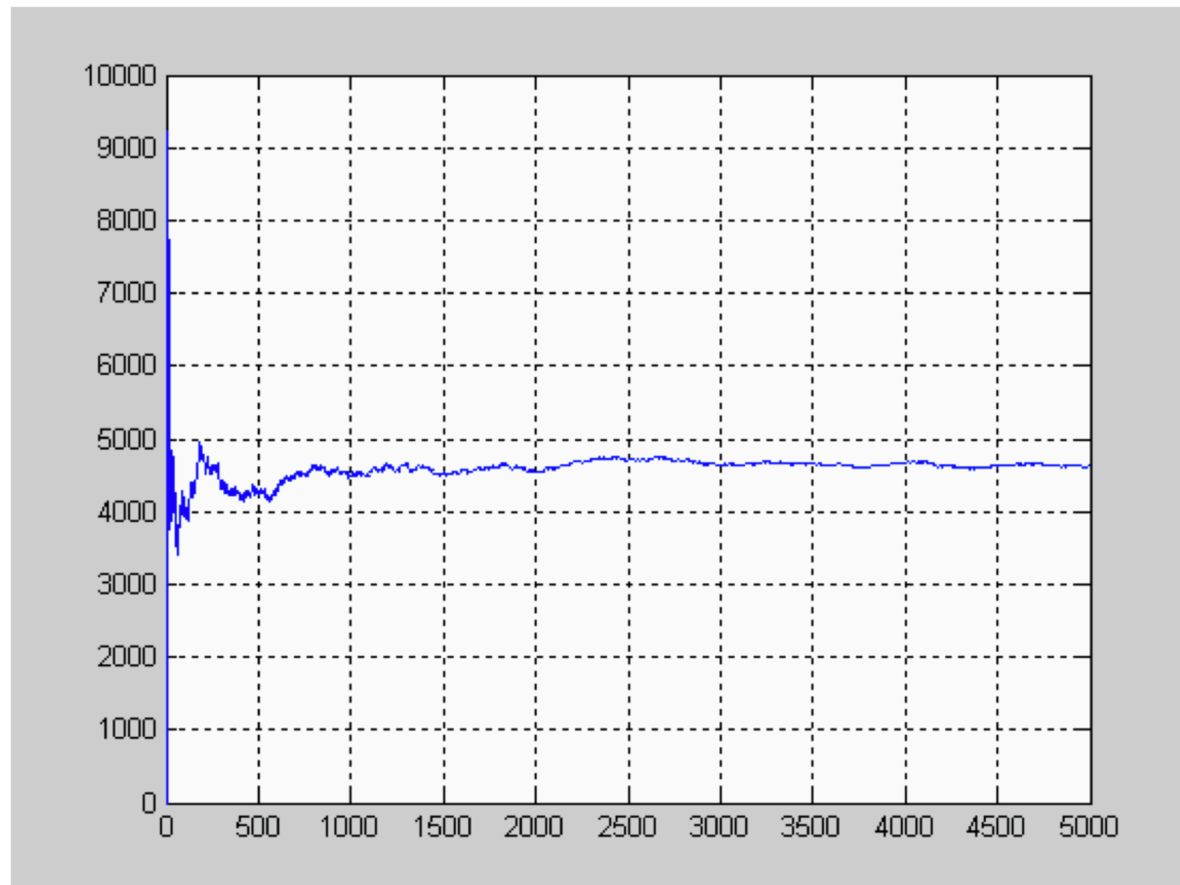


Figure 3. Number of iterations required vs. iteration number.

This shows how the number of iterations needed stabilizes within about 500-1000 iterations, and within about 100 iterations it is accurate to 20%.