# Bayesian Synthesis Models #2

Jingchen (Monika) Hu

Vassar College

Data Confidentiality

# Outline

1. Case study: SynLBD
2. The sequential synthesis procedure
3. Bayesian synthesis models for different variable types
4. Case study revisit: SynLBD

# Outline

# The Longitudinal Business Database (LBD)

Kinney et al (2011) and appendix:

- The LBD is an annual economic census of establishments in the United States comprising more than 20 million records dating back to 1976.
- It supports an active research agenda on
  - business entry and exit
  - gross employment flows
  - employement volatility
  - industrial organization
  - and other topics that cannot be adequately addressed without establishment-level data

# The Longitudinal Business Database (LBD)

Kinney et al (2011) and appendix:

- The LBD is an annual economic census of establishments in the United States comprising more than 20 million records dating back to 1976.
- It supports an active research agenda on
  - business entry and exit
  - gross employment flows
  - employement volatility
  - industrial organization
  - and other topics that cannot be adequately addressed without establishment-level data
- Access to the LBD is regulated by Title 13 and Title 26 of the U.S. code
  - researchers desiring access must follow lenghty and potentially costly procedures to use the data
  - at one of the several Census Bureau Research Data Centers (RDCs)

# The Synthetic LBD (SynLBD) variables

Table 4.1. SynLBD variable description. Taken from Table 1 in Kinney et al (2011). Column name Not. stands for Notation.

| Name | Type | Description | Not. | Action |
|---|---|---|---|---|
| ID | Identifier | Unique Random Number of Establishment | | Created |
| County | Categorical | Geographic Location | $x_1$ | Not released |
| SIC | Categorical | Industry Code | $x_2$ | Unmodified |
| Firstyear | Categorical | First Year Establishment is Observed | $y_1$ | Synthesized |
| Lastyear | Categorical | Last Year Establishment is Observed | $y_2$ | Synthesized |
| Year | Categorical | Year dating of annual variables | | Created |
| Multiunit | Categorical | Multiunit Status | $y_3$ | Synthesized |
| Employment | Continuous | March 12 Employment (annual) | $y_4$ | Synthesized |
| Payroll | Continuous | Payroll (annual) | $y_5$ | Synthesized |

- There are additional variables in the confidential LBD, such as firm structure, were not used to generate the SynLBD.
- No geographic or firm-level information are included, though County and State were used in the synthesis.

# The Synthetic LBD (SynLBD) features

- The LBD is a universe file, therefore there are no sampling weights.
- SynLBD is based on a cleaned version of the confidential database
    - several data cleaning steps
- SynLBD comprises one record for each of 21 million establishments active in the Business Register any time between 1976 and 2001.

# Outline

# Sequential synthesis of SynLBD

Table 4.1. SynLBD variable description. Taken from Table 1 in Kinney et al (2011). Column name Not. stands for Notation.

| Name | Type | Description | Not. | Action |
|---|---|---|---|---|
| ID | Identifier | Unique Random Number of Establishment | | Created |
| County | Categorical | Geographic Location | $x_1$ | Not released |
| SIC | Categorical | Industry Code | $x_2$ | Unmodified |
| Firstyear | Categorical | First Year Establishment is Observed | $y_1$ | Synthesized |
| Lastyear | Categorical | Last Year Establishment is Observed | $y_2$ | Synthesized |
| Year | Categorical | Year dating of annual variables | | Created |
| Multiunit | Categorical | Multiunit Status | $y_3$ | Synthesized |
| Employment | Continuous | March 12 Employment (annual) | $y_4$ | Synthesized |
| Payroll | Continuous | Payroll (annual) | $y_5$ | Synthesized |

1. `Firstyear` $f(y_1 \mid x_1, x_2)$

# Sequential synthesis of SynLBD

Table 4.1. SynLBD variable description. Taken from Table 1 in Kinney et al (2011). Column name Not. stands for Notation.

| Name | Type | Description | Not. | Action |
|---|---|---|---|---|
| ID | Identifier | Unique Random Number of Establishment | | Created |
| County | Categorical | Geographic Location | $x_1$ | Not released |
| SIC | Categorical | Industry Code | $x_2$ | Unmodified |
| Firstyear | Categorical | First Year Establishment is Observed | $y_1$ | Synthesized |
| Lastyear | Categorical | Last Year Establishment is Observed | $y_2$ | Synthesized |
| Year | Categorical | Year dating of annual variables | | Created |
| Multiunit | Categorical | Multiunit Status | $y_3$ | Synthesized |
| Employment | Continuous | March 12 Employment (annual) | $y_4$ | Synthesized |
| Payroll | Continuous | Payroll (annual) | $y_5$ | Synthesized |

1. `Firstyear` $f(y_1 \mid x_1, x_2)$
2. `Lastyear` $f(y_2 \mid y_1, x_1, x_2)$

# Sequential synthesis of SynLBD

Table 4.1. SynLBD variable description. Taken from Table 1 in Kinney et al (2011). Column name Not. stands for Notation.

| Name | Type | Description | Not. | Action |
|---|---|---|---|---|
| ID | Identifier | Unique Random Number of Establishment | | Created |
| County | Categorical | Geographic Location | $x_1$ | Not released |
| SIC | Categorical | Industry Code | $x_2$ | Unmodified |
| Firstyear | Categorical | First Year Establishment is Observed | $y_1$ | Synthesized |
| Lastyear | Categorical | Last Year Establishment is Observed | $y_2$ | Synthesized |
| Year | Categorical | Year dating of annual variables | | Created |
| Multiunit | Categorical | Multiunit Status | $y_3$ | Synthesized |
| Employment | Continuous | March 12 Employment (annual) | $y_4$ | Synthesized |
| Payroll | Continuous | Payroll (annual) | $y_5$ | Synthesized |

1. `Firstyear` $f(y_1 \mid x_1, x_2)$
2. `Lastyear` $f(y_2 \mid y_1, x_1, x_2)$
3. `Multiunit` $f(y_3 \mid y_2, y_1, x_1, x_2)$

# Sequential synthesis of SynLBD

Table 4.1. SynLBD variable description. Taken from Table 1 in Kinney et al (2011). Column name Not. stands for Notation.

| Name | Type | Description | Not. | Action |
|---|---|---|---|---|
| ID | Identifier | Unique Random Number of Establishment | | Created |
| County | Categorical | Geographic Location | $x_1$ | Not released |
| SIC | Categorical | Industry Code | $x_2$ | Unmodified |
| Firstyear | Categorical | First Year Establishment is Observed | $y_1$ | Synthesized |
| Lastyear | Categorical | Last Year Establishment is Observed | $y_2$ | Synthesized |
| Year | Categorical | Year dating of annual variables | | Created |
| Multiunit | Categorical | Multiunit Status | $y_3$ | Synthesized |
| Employment | Continuous | March 12 Employment (annual) | $y_4$ | Synthesized |
| Payroll | Continuous | Payroll (annual) | $y_5$ | Synthesized |

1. `Firstyear` $f(y_1 \mid x_1, x_2)$
2. `Lastyear` $f(y_2 \mid y_1, x_1, x_2)$
3. `Multiunit` $f(y_3 \mid y_2, y_1, x_1, x_2)$
4. `Employment` $f(y_4^{(t)} \mid y_4^{(t-1)}, y_3, y_2, y_1, x_1, x_2),\ t \in [1976, 2001]$

# Sequential synthesis of SynLBD

Table 4.1. SynLBD variable description. Taken from Table 1 in Kinney et al (2011). Column name Not. stands for Notation.

| Name | Type | Description | Not. | Action |
|---|---|---|---|---|
| ID | Identifier | Unique Random Number of Establishment | | Created |
| County | Categorical | Geographic Location | $x_1$ | Not released |
| SIC | Categorical | Industry Code | $x_2$ | Unmodified |
| Firstyear | Categorical | First Year Establishment is Observed | $y_1$ | Synthesized |
| Lastyear | Categorical | Last Year Establishment is Observed | $y_2$ | Synthesized |
| Year | Categorical | Year dating of annual variables | | Created |
| Multiunit | Categorical | Multiunit Status | $y_3$ | Synthesized |
| Employment | Continuous | March 12 Employment (annual) | $y_4$ | Synthesized |
| Payroll | Continuous | Payroll (annual) | $y_5$ | Synthesized |

1. `Firstyear` $f(y_1 \mid x_1, x_2)$
2. `Lastyear` $f(y_2 \mid y_1, x_1, x_2)$
3. `Multiunit` $f(y_3 \mid y_2, y_1, x_1, x_2)$
4. `Employment` $f(y_4^{(t)} \mid y_4^{(t-1)}, y_3, y_2, y_1, x_1, x_2)$, $t \in [1976, 2001]$
5. `Payroll` $f(y_5^{(t)} \mid y_4^{(t)}, y_5^{(t-1)}, y_3, y_2, y_1, x_1, x_2)$, $t \in [1976, 2001]$

# Why sequential synthesis?

- The fully conditional specification (FCS) approach
  - multiple imputation for missing data
  - Raghunathan et al. (2001), Drechsler (2011), van Buuren and Oudshoorn (2011)
- Several variables to be synthesized
  - sequentially synthesize one variable at a time
  - given what has been synthesized already

# Why sequential synthesis works?

- Challenging to develop a joint density for all variables.
- May be less challenging to work with univariate conditional density for each variable.

# Why sequential synthesis works?

- Challenging to develop a joint density for all variables.
- May be less challenging to work with univariate conditional density for each variable.

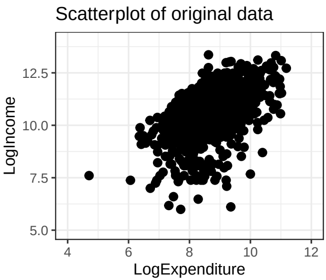
- For illustration purpose, ignoring $t$ for $y_4$ and $y_5$, the joint density can be expressed as product of a sequence of conditional density:

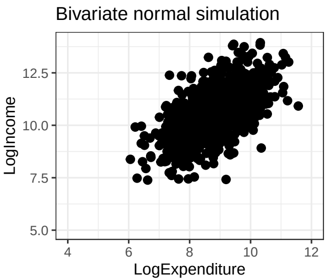
$$\begin{aligned} f(y_5, y_4, y_3, y_2, y_1 \mid x_1, x_2) &= f(y_5 \mid y_4, y_3, y_2, y_1, x_1, x_2) \\ & \quad f(y_4 \mid y_3, y_2, y_1, x_1, x_2) \\ & \quad f(y_3 \mid y_2, y_1, x_1, x_2) \\ & \quad f(y_2 \mid y_1, x_1, x_2) \\ & \quad f(y_1 \mid x_1, x_2) \end{aligned}$$

- The product of conditionals (right) may be a good approximation to the joint (left).

# Joint modeling vs FCS

- Joint modeling
  - ideal if empirical data follows a standard multivariate distribution, though seldomly true
  - example: `log(Income)` and `log(Expenditure)` in the CE sample
    1. simple linear regression
    2. bivariate normal model



# Joint modeling vs FCS cont'd

- FCS
  - flexible to account for bounds, interactions or constraints between different variables
  - challenging to guarantee a good approximation to the joint density

# Joint modeling vs FCS cont'd

- FCS
  - flexible to account for bounds, interactions or constraints between different variables
  - challenging to guarantee a good approximation to the joint density
  - example: `log(Income)` and `log(Expenditure)` in the CE sample
  - how to use sequential synthesis for synthesizing both variables?

# Outline

# Continuous variables

- Use normal linear regression model, as for the CE sample.

# Binary variables

- Binary outcome examples: labor participation (0 or 1), loan default (yes or no).
- Idea from normal linear regression: model the outcome variable as a function of explanatory variable(s).

# Binary variables

- Binary outcome examples: labor participation (0 or 1), loan default (yes or no).
- Idea from normal linear regression: model the outcome variable as a function of explanatory variable(s).

- Outcome variable $Y_i \in \{0, 1\}$ as a binomial random variable with trial 1:

$$Y_i \overset{ind}{\sim} \mathrm{Binomial}(1, p_i) \tag{1}$$

  - $p_i$ is the success probability of observation $i$ taking $Y_i = 1$
  - 1 indicating 1 trial of this binomial experiment

# Logistic model

$$Y_i \overset{ind}{\sim} \text{Binomial}(1, p_i)$$

- $p_i \in (0, 1)$ and is continuous.
- The odds, $\frac{p_i}{1-p_i}$ is then a positive, continuous quantity.
- The log odds, $\log\left(\frac{p_i}{1-p_i}\right)$ is then a continuous quantity on the real line, i.e. $\log\left(\frac{p_i}{1-p_i}\right) \in (-\infty, \infty)$.

# Logistic model

$$Y_i \overset{ind}{\sim} \text{Binomial}(1, p_i)$$

- $p_i \in (0, 1)$ and is continuous.
- The odds, $\frac{p_i}{1-p_i}$ is then a positive, continuous quantity.
- The log odds, $\log\left(\frac{p_i}{1-p_i}\right)$ is then a continuous quantity on the real line, i.e. $\log\left(\frac{p_i}{1-p_i}\right) \in (-\infty, \infty)$.
- We can then model $\log\left(\frac{p_i}{1-p_i}\right)$ as a linear function of potential explanatory variable(s).

# Logistic model cont'd

- Assume one explanatory variable, $X_i$.
- A linear function of $X_i$ for the logit of $p_i$ can be expressed as

$$\mathrm{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_i \tag{2}$$

  - two parameters, $\beta_0$ and $\beta_1$
  - Bayesian inference: prior for $\beta_0$ and $\beta_1$, MCMC estimation, posterior draws

# Logistic model cont'd

- Assume one explanatory variable, $X_i$.
- A linear function of $X_i$ for the logit of $p_i$ can be expressed as

$$\mathrm{logit}(p_i) = \log\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \beta_1 X_i \tag{2}$$

  - two parameters, $\beta_0$ and $\beta_1$
  - Bayesian inference: prior for $\beta_0$ and $\beta_1$, MCMC estimation, posterior draws
  - for Bayesian synthesis: use these posterior parameter draws for simulating synthetic data from the posterior predictive distribution

# Sample JAGS script for a logistic model

```
modelString <-"
model {
## sampling
for (i in 1:N){
   y[i] ~ dbern(p[i])
   logit(p[i]) <- beta0 + beta1*x[i]
}

## priors
beta0 ~ dnorm(mu0, g0)
beta1 ~ dnorm(mu1, g1)
}
"
```

- For illustration purpose, we use `beta0 ~ dnorm(mu0, g0)` and `beta1 ~ dnorm(mu1, g1)` as placeholders for the two prior distributions.
- In practice, directly specifying priors for $\beta_0$ and $\beta_1$ could be challenging, and we can use a conditional means prior.

# Synthesis from a logistic model

- To simulate a posterior predictive draw of $\tilde{Y}_i$ given explanatory variable $X_i$ and parameter draws of $\{\beta_0, \beta_1\}$:

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_i, \quad (3)$$

$$\tilde{Y}_i \overset{ind}{\sim} \text{Binomial}(1, p_i). \quad (4)$$

# Synthesis from a logistic model

- To simulate a posterior predictive draw of $\tilde{Y}_i$ given explanatory variable $X_i$ and parameter draws of $\{\beta_0, \beta_1\}$:

$$\begin{aligned} \mathrm{logit}(p_i) &= \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_i, && (3) \\ \tilde{Y}_i &\overset{ind}{\sim} \mathrm{Binomial}(1, p_i). && (4) \end{aligned}$$

- Note that to get $p_i$ from $\beta_0$, $\beta_1$ and $X_i$, we need the following algebra transformation:

$$\begin{aligned} \log\left(\frac{p_i}{1-p_i}\right) &= \beta_0 + \beta_1 X_i \\ \frac{p_i}{1-p_i} &= \exp(\beta_0 + \beta_1 X_i) \\ p_i &= \frac{\exp(\beta_0 + \beta_1 X_i)}{1 + \exp(\beta_0 + \beta_1 X_i)}. && (5) \end{aligned}$$

# Synthesis from a logistic model cont'd

- To simulate synthetic values for all $n$ observations:

$$
\begin{aligned}
\text{simulate } p_1 = \beta_0 + \beta_1 X_1 \quad &\rightarrow \quad \text{sample } \tilde{Y}_1 \sim \mathrm{Binomial}(1, p_1) \\
\text{simulate } p_2 = \beta_0 + \beta_1 X_2 \quad &\rightarrow \quad \text{sample } \tilde{Y}_2 \sim \mathrm{Binomial}(1, p_2) \\
&\vdots \\
\text{simulate } p_n = \beta_0 + \beta_1 X_n \quad &\rightarrow \quad \text{sample } \tilde{Y}_n \sim \mathrm{Binomial}(1, p_n),
\end{aligned}
$$

we will have simulated one synthetic vector $(\tilde{Y}_i)_{i=1,\cdots,n}$.

# Synthesis from a logistic model cont'd

- Suppose the output from the `run.jags` function is saved as `posterior`.

```
post <- as.mcmc(posterior)
synthesize <- function(X, index, n){
  log_p <- post[index, "beta0"] +  X * post[index, "beta1"]
  p <- exp(log_p) / (1 + exp(log_p))
  synthetic_Y <- rbinom(length(p), size = n, prob = p)
  data.frame(X, synthetic_Y)
}
```

- The input `X` is a vector of the un-synthesized variable, i.e. the explanatory variable. `index` indicates which set of posterior draws to be used.
- If multiple synthetic datasets are needed, for example $m = 20$, we can then repeat this process $m$ times using $m$ independent MCMC iterations.

# Categorical variables

- A categorical outcome variable, $Y_i$, which takes value from 1 to $C$.
- We could model it as a multinomial random variable with trial 1:

$$Y_i \overset{ind}{\sim} \mathrm{Multinomial}(1, p_{i1}, \cdots, p_{iC}) \tag{6}$$

  - $p_{ic}$ is the success probability of observation $i$ taking $Y_i = c$
  - 1 indicating 1 trial of this multionmial experiment
  - $\sum_{i=1}^{C} p_{ic} = 1$.

# Categorical variables

- A categorical outcome variable, $Y_i$, which takes value from 1 to $C$.
- We could model it as a multinomial random variable with trial 1:

$$Y_i \stackrel{ind}{\sim} \mathrm{Multinomial}(1, p_{i1}, \cdots, p_{iC}) \tag{6}$$

  - $p_{ic}$ is the success probability of observation $i$ taking $Y_i = c$
  - 1 indicating 1 trial of this multionmial experiment
  - $\sum_{i=1}^{C} p_{ic} = 1$.

- Two types of models for categorical outcomes:
  1. Multinomial logistic model (with explanatory variable(s))
  2. Dirichlet-multinomial model (without explanatory variable(s))

# Multinomial logistic model

- A generalization of binary logistic regression to the multi-categorical outcome case.
- For illustration purpose, assume one explanatory variable, $X_i$.

# Multinomial logistic model

- A generalization of binary logistic regression to the multi-categorical outcome case.
- For illustration purpose, assume one explanatory variable, $X_i$.

- Similar to logistic model, we can define the log odds ratio for category $c$ relative to category 1 as

$$\log\left(\frac{p_{ic}}{p_{i1}}\right) = \beta_{0c} + \beta_{1c} X_i. \tag{7}$$

  - two parameters, $\beta_{0c}$ and $\beta_{1c}$ for each category $c$
  - Bayesian inference: prior for $\beta_{0c}$ and $\beta_{1c}$, MCMC estimation, posterior draws
  - for Bayesian synthesis: use these posterior parameter draws for simulating synthetic data from the posterior predictive distribution

# Multinomial logistic model cont'd

- After algebra transformation,

$$p_{ic} = \frac{\exp(\beta_{0c} + \beta_{1c} X_i)}{\sum_{c'=1}^{C} \exp(\beta_{0c'} + \beta_{1c'} X_i)}, \tag{8}$$

with the constraint that $\exp(\beta_{01} + \beta_{11} X_i) = 1$.

# Sample JAGS script for a multinomial logistic model

```
modelString <-"
model {
## sampling
for (i in 1:N){
   y[i] ~ dmulti(p[i,1:C],1)
   for (c in 1:C){
     p[i,c] <- q[i,c]/sum(q[i,])
     log(q[i,c]) <- beta0[c] + beta1[c]*x[i]
   }
}

## priors
beta0[1] <- 0
beta1[1] <- 0
for (c in 2:C){
  beta0[c] ~ dnorm(0, 0.00001)
  beta1[c] ~ dnorm(0, 0.00001)
}
}
"
```

# Sample JAGS script for a multinomial logistic model cont'd

- `q[i,c]` to represent $\exp(\beta_{0c} + \beta_{1c} X_i)$.
- `beta0[c] ~ dnorm(0, 0.00001)` and `beta1[c] ~ dnorm(0, 0.00001)` as default weakly informative priors for the two parameters.
- `beta0[1] <- 0` and `beta1[1] <- 0` are needed to satisfy the $\exp(\beta_{01} + \beta_{11} X_i) = 1$ constraint.

# Dirichlet-multinomial model

For a multinomial sampling model, there exists a conjugate prior, the Dirichlet distribution.

- Sampling for $\{Y_i\}$:

$$Y_i \overset{ind}{\sim} \mathrm{Multinomial}(1, p_{i1}, \cdots, p_{iC}),$$

which produces the likelihood function of $(p_{i1}, \cdots, p_{iC})_{i=1,\cdots,n}$ as

$$L\left((p_{i1}, \cdots, p_{iC})_{i=1,\cdots,n}\right) = \prod_{i=1}^{n} \left( \frac{n!}{\prod_{c=1}^{C} Y_{ic}!} \prod_{c=1}^{C} p_{ic}^{Y_{ic}} \right) \propto \prod_{i=1}^{n} \prod_{c=1}^{C} p_{ic}^{Y_{ic}}, \quad (9)$$

where $Y_i = (Y_{i1}, \cdots, Y_{iC})$ with $(C-1)$ 0's and one 1.

# Dirichlet-multinomial model cont'd

- Conjugate prior for $(p_{i1}, \cdots, p_{iC})$ is a Dirichlet distribution:

$$(p_{i1}, \cdots, p_{iC}) \sim \mathrm{Dirichlet}(\alpha_1, \cdots, \alpha_C),$$

which produces the individual prior density of $(p_{i1}, \cdots, p_{iC})$ as

$$\pi(p_{i1}, \cdots, p_{iC}) = \frac{\Gamma(\sum_{c=1}^{C} \alpha_c)}{\prod_{c=1}^{C} \Gamma(\alpha_c)} \prod_{c=1}^{C} p_{ic}^{\alpha_c - 1} \propto \prod_{c=1}^{C} p_{ic}^{\alpha_c - 1},$$

and the joint prior density of $(p_{i1}, \cdots, p_{iC})_{i=1,\cdots,n}$ as

$$\pi((p_{i1}, \cdots, p_{iC})_{i=1,\cdots,n}) \propto \prod_{i=1}^{n} \prod_{c=1}^{C} p_{ic}^{\alpha_c - 1}. \tag{10}$$

# Dirichlet-multinomial model cont'd

- Take the likelihood and prior together and we can obtain the joint posterior density for $(p_{i1}, \cdots, p_{iC})_{i=1,\cdots,n} \mid (Y_i)_{i=1,\cdots,n}$ as

$$\pi((p_{i1}, \cdots, p_{iC})_{i=1,\cdots,n} \mid (Y_i)_{i=1,\cdots,n}) \propto \prod_{i=1}^{n} \prod_{c=1}^{C} p_{ic}^{\alpha_c + Y_{ic} - 1}, \tag{11}$$

which gives us a Dirichlet posterior for each $(p_{i1}, \cdots, p_{iC}) \mid Y_i$:

$$(p_{i1}, \cdots, p_{iC}) \mid Y_i \sim \text{Dirichlet}(\alpha_1 + Y_{i1}, \cdots, \alpha_C + Y_{iC}). \tag{12}$$

# Sample JAGS script for a Dirichlet-multinomial model

```
modelString <-"
model {
## sampling
for (i in 1:N){
   y[i,] ~ dmulti(p[i,1:C],1)
   p[i,] ~ ddirch(alpha[])
}

## priors
for (c in 1:C){
  alpha[c] <- 1
}
}
"
```

- `y[i,]` is a vector of $(C-1)$ 0's and one 1 at the observed category $c$ for observation $i$.
- The prior choice above is a uniform distribution, such that each category $c$ is equally likely.

# Outline

# Discussion questions

1. What are the synthesized and un-synthesized variables in the SynLBD?
2. What approaches did they take when they have more than 1 variables deemed sensitive and to be synthesized?
3. What Bayesian synthesis model(s) did they use? Details of the synthesis models are in Kinney et al (2011) appendix.
4. How many synthetic datasets did they generate?
5. How did they evaluate the utility of the synthetic datasets? Can you think of any other utility measures?
6. How did they evaluate the disclosure risks? Can you think of any other disclosure risks measures?

# References


- Kinney, S. K., Reiter, J. P., Reznek, A. P., Miranda, J., Jarmin, R. S., and Abowd, J. M. (2011). Towards Unrestricted Public Use Business Microdata: The Synthetic Longitudinal Business Database. International Statistical Review, Vol. 79, No. 3, pp. 362-384.
- Raghunathan, T. E., Lepkowski, J. M., van Hoewyk, J., and Solenberger, P. (2001). A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models, Survey Methodology, Vol 27, No. 1, 85-95.
- Drechsler, J. (2011). Synthetic Datasets for Statistical Disclosure Control: Theory and Implementation, Lecture Notes in Statistics, Springer.
- van Buuren, S. and Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R, Journal of Statistical Software, Vol 45, Issue 3.