Measurement Error Model

1. Introduction

We consider linear regression with independent variables observed with error. In other words, assume linear model as

$$y = \beta_0 + \beta_1 x + \epsilon$$

We cannot observe x, rather $w = x + v$ where $v \sim \mathrm{N}(0, \sigma_v^2)$, where error is simultaneously observed with independent variable. From now on, we call w as **proxy** because it is observed with error. We consider linear regression where two proxies are observed instead of observing independent variables. In other words, we want to model linear relationship between y and x where w and z are observed.

$$\begin{aligned} y &= \beta_0 + \beta_1 x + \epsilon \\ w &= x + v \\ z &= \gamma_0 + \gamma_1 x + \delta \end{aligned}$$

Note that two proxies do not have identical error structure. Proxy w is assumed to have structure where error is simply added to independent variable. Proxy z is assumed to have linear relationship between independent variable.

2. Model

We assume bayesian hierarchical model to simple linear regression with two proxies. The reasons why we choose bayesian inference are

1. It is easy to stack more assumptions to model because MCMC is chosen as baseline inference method
2. Bayesian inference is more robust when sample size is small

2.1. Model Structure

Full model specification is

$$\begin{aligned} y &= \beta_0 + \beta_1 x + \epsilon \\ w &= x + v \\ z &= \gamma_0 + \gamma_1 x + \delta \\ \epsilon &\sim \mathrm{N}(0, \sigma_\epsilon^2) \\ x &\sim \mathrm{N}(\mu_x, \sigma_x^2) \\ v &\sim \mathrm{N}(0, \sigma_v^2) \\ \delta &\sim \mathrm{N}(0, \sigma_\delta^2) \end{aligned}$$

Priors of model parameters are assumed as

$$\begin{aligned} (\beta_0, \beta_1)^T &\sim \mathrm{N}_2(\mathbf{0}, \sigma_\beta^2 \mathbf{I}) \\ (\gamma_0, \gamma_1)^T &\sim \mathrm{N}_2(\mathbf{0}, \sigma_\gamma^2 \mathbf{I}) \\ \sigma_\epsilon^2 &\sim \text{Inv-Gamma}(A_\epsilon, B_\epsilon) \\ \mu_x &\sim \mathrm{N}(0, \sigma_{\mu_x}^2) \\ \sigma_x^2 &\sim \text{Inv-Gamma}(A_x, B_x) \\ \sigma_v^2 &\sim \text{Inv-Gamma}(A_v, B_v) \\ \sigma_\delta^2 &\sim \text{Inv-Gamma}(A_\delta, B_\delta) \end{aligned}$$

2.2. Posterior distribution

We derive posterior distribution of model parameters through gibbs sampling. The full conditional distributions of each parameters are specified below.

$$
\begin{aligned}
\boldsymbol{\beta} \mid rest &\sim N\left(\left(\mathbf{X}^T\mathbf{X}\sigma_\epsilon^{-2} + \sigma_\beta^{-2}\mathbf{I}\right)^{-1}\mathbf{y}^T\mathbf{X}\sigma_\epsilon^2, \left(\mathbf{X}^T\mathbf{X}\sigma_\epsilon^{-2} + \sigma_\beta^{-2}\mathbf{I}\right)^{-1}\right)\\
\boldsymbol{\gamma} \mid rest &\sim N\left(\left(\mathbf{X}^T\mathbf{X}\sigma_\delta^{-2} + \sigma_\gamma^{-2}\mathbf{I}\right)^{-1}\mathbf{z}^T\mathbf{X}\sigma_\delta^2, \left(\mathbf{X}^T\mathbf{X}\sigma_\delta^{-2} + \sigma_\gamma^{-2}\mathbf{I}\right)^{-1}\right)\\
\sigma_\epsilon^2 \mid rest &\sim IG\left(A_\epsilon + n/2, B_\epsilon + \frac{1}{2}||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2\right)\\
\mu_x \mid rest &\sim N\left(\frac{\mathbf{1^T x}/\boldsymbol{\sigma_x^2}}{n\sigma_x^{-2} + \sigma_{\mu_x}^{-2}}, \frac{1}{n\sigma_x^{-2} + \sigma_{\mu_x}^{-2}}\right)\\
\sigma_x^2 \mid rest &\sim IG\left(A_x + \frac{n}{2}, B_x + \frac{1}{2}||\mathbf{x} - \mu_x\mathbf{1}||^2\right)\\
\sigma_v^2 \mid rest &\sim IG\left(A_v + \frac{n}{2}, B_v + \frac{1}{2}||\mathbf{w} - \mathbf{x}||^2\right)\\
\sigma_\delta^2 \mid rest &\sim IG\left(A_\delta + \frac{n}{2}, B_\delta + \frac{1}{2}||\mathbf{z} - \mathbf{X}\boldsymbol{\gamma}||^2\right)\\
x_i \mid rest &\sim N\left(\frac{\beta_1(y_i - \beta_0)/\sigma_\epsilon^2 + w_i/\sigma_v^2 + \mu_x/\sigma_x^2}{\beta_1^2/\sigma_\epsilon^2 + 1/\sigma_v^2 + 1/\sigma_x^2}, \frac{1}{\beta_1^2/\sigma_\epsilon^2 + 1/\sigma_v^2 + 1/\sigma_x^2}\right)
\end{aligned}
$$

3. Simulation study

3.1. Data generation

$$
\begin{aligned}
y_i &= -1 + 2x_i + \epsilon_i\\
w_i &= x_i + v_i\\
z_i &= -1 + 3x_i + d_i\\
x_i &\sim N(1/2, 1)\\
v_i &\sim N(0, 1)\\
d_i &\sim N(0, 1)\\
\epsilon_i &\sim N(0, 1)
\end{aligned}
$$

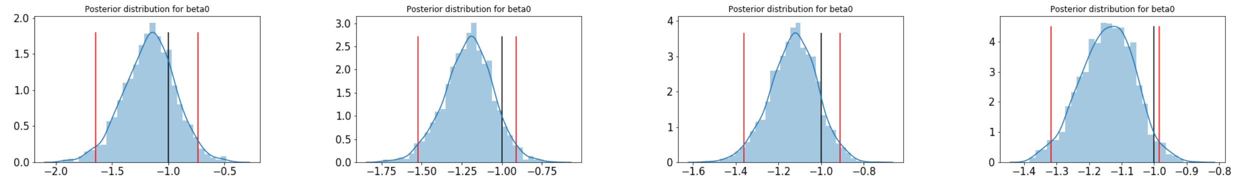
3.2. Set up

1. Generate n data with parameters defined in 3.1.
2. Fit posterior distribution for each parameter with 10000 iterations Gibbs sampler
3. Repeat 1 and 2 with $n = 120, 300, 500, 1000$

3.3. Result

1. 95% credible interval is specified in red vertical line
2. True value of parameters are specified in black vertical line

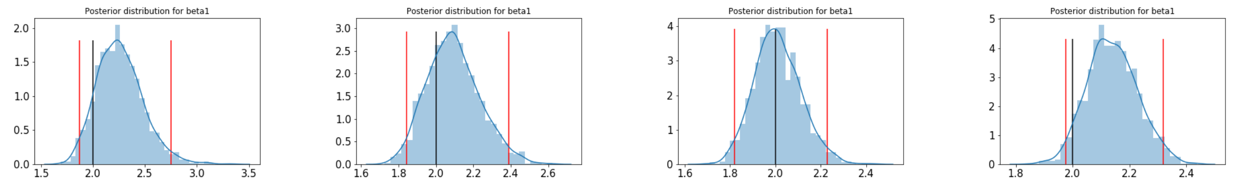
Results with β_0



(a) **(b)** **(c)** **(d)**

Figure 1: **(a)** $n = 120$ **(b)** $n = 300$ **(c)** $n = 500$ **(d)** $n = 1000$

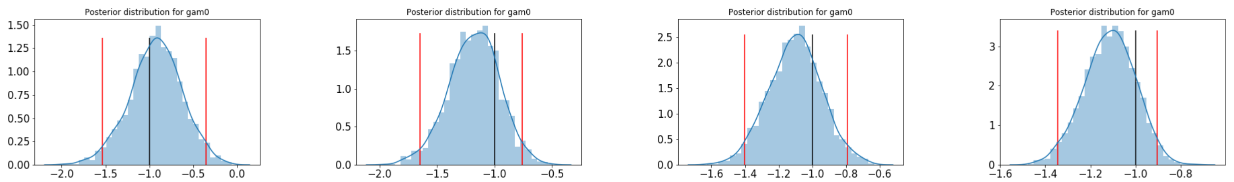
Results with β_1



(a) **(b)** **(c)** **(d)**

Figure 2: **(a)** $n = 120$ **(b)** $n = 300$ **(c)** $n = 500$ **(d)** $n = 1000$

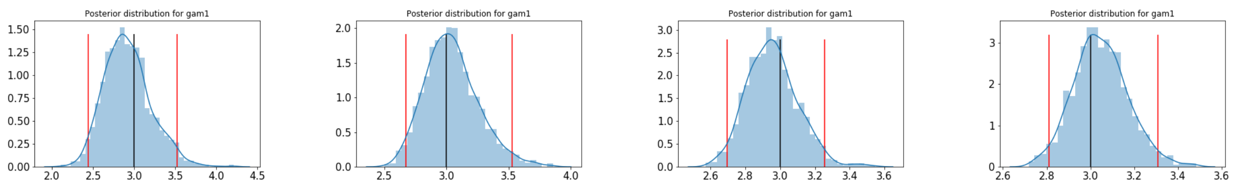
Results with γ_0



(a) **(b)** **(c)** **(d)**

Figure 3: **(a)** $n = 120$ **(b)** $n = 300$ **(c)** $n = 500$ **(d)** $n = 1000$

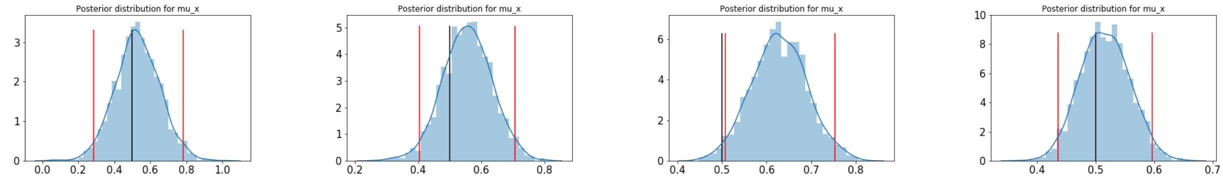
Results with γ_1



(a) **(b)** **(c)** **(d)**

Figure 4: **(a)** $n = 120$ **(b)** $n = 300$ **(c)** $n = 500$ **(d)** $n = 1000$

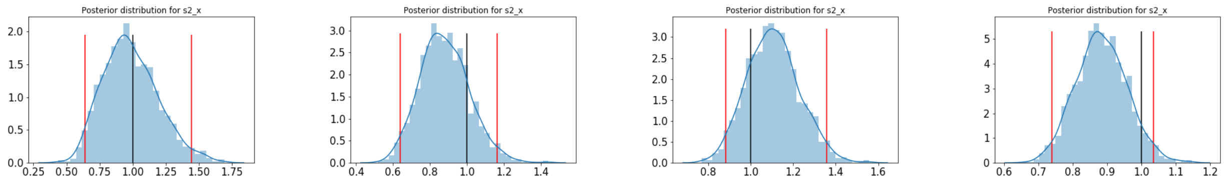
Results with μ_x



(a) **(b)** **(c)** **(d)**

Figure 5: **(a)** $n = 120$ **(b)** $n = 300$ **(c)** $n = 500$ **(d)** $n = 1000$

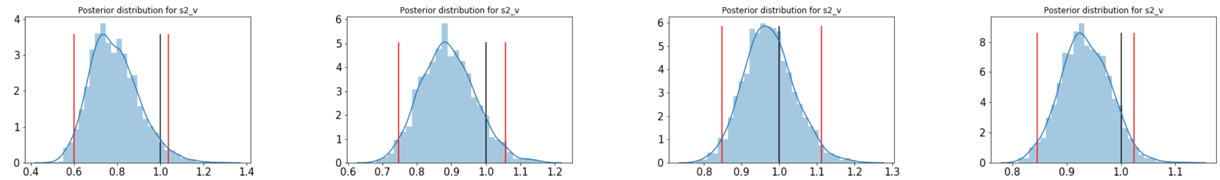
Results with σ_x^2



(a) **(b)** **(c)** **(d)**

Figure 6: **(a)** $n = 120$ **(b)** $n = 300$ **(c)** $n = 500$ **(d)** $n = 1000$

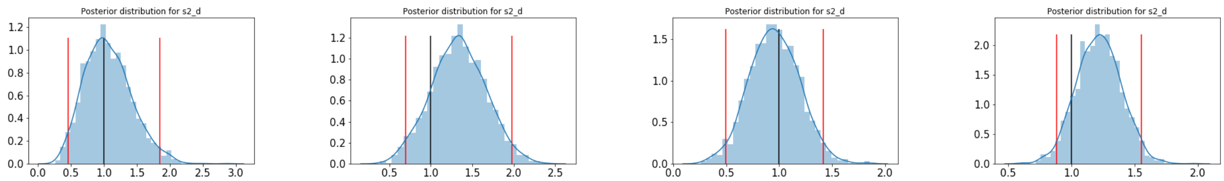
Results with σ_v^2



(a) **(b)** **(c)** **(d)**

Figure 7: **(a)** $n = 120$ **(b)** $n = 300$ **(c)** $n = 500$ **(d)** $n = 1000$

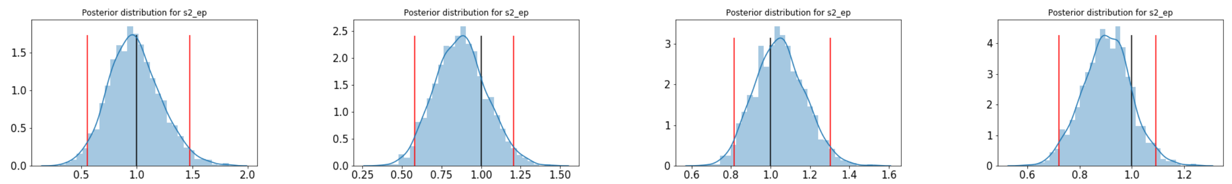
Results with σ_d^2



(a) **(b)** **(c)** **(d)**

Figure 8: **(a)** $n = 120$ **(b)** $n = 300$ **(c)** $n = 500$ **(d)** $n = 1000$

Results with σ_ϵ^2



(a) (b) (c) (d)

Figure 9: **(a)** $n = 120$ **(b)** $n = 300$ **(c)** $n = 500$ **(d)** $n = 1000$

3.4. Interpretation

1. Even if when sample size is samll, 95% credibel interval includes true parameters.
2. Most of credible intervals from posterior distributions include true parameters.
3. As expected, as sample size becomes bigger, credible intervals are getting narrower to true parameters.

4. Future work

1. Nonparametric measurement error model.
2. Variational inference for measurement error model.