Bayesian Analysis of Linear Models

with Endogeneity using Instrumental Variables

Student: Dongyuan Lin

Supervisors: A/Prof Clara Grazian, Dr Linh Nghiem

Table of Contents

Motivation

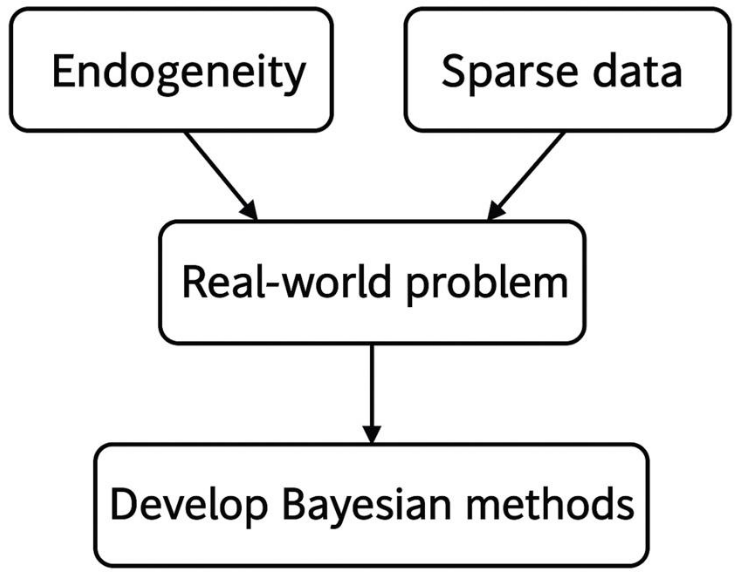
- Many real-world regression problems suffer from **endogeneity**:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

  - OLS assumes $\mathbb{E}[\mathbf{X}^\top \boldsymbol{\varepsilon}] = 0$, but this is often violated because of:
    - **Omitted variables**: hidden factors affect both **X** and **Y**
    - **Measurement error**: noisy or misreported **X**
  - When ignore them, OLS estimates of **β** are biased and inconsistent

- **Sparse data** makes variable selection challenging
- These two problems often appear together in practice

- **Goal:** Develop Bayesian methods to address both simultaneously



Introduction

Bayesian Inference

- **Core idea:**

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}$$

  - **Prior**: What we believe before seeing the data
  - **Likelihood**: What the data tells us
  - **Posterior**: What we believe after seeing the data

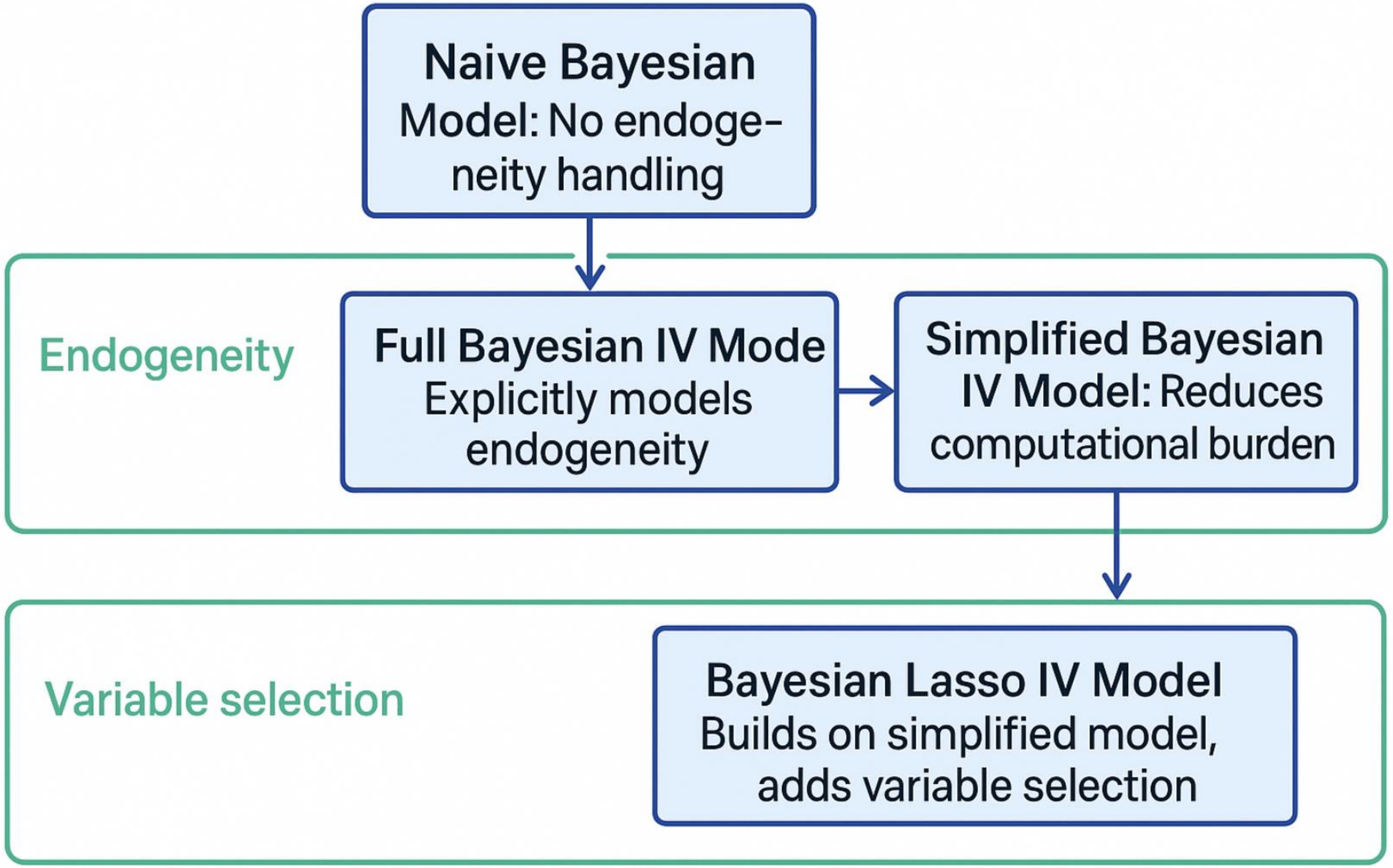
- **Why use Bayesian in IV models?**
- Handles uncertainty (posterior, not just point estimates)
- Naturally incorporates prior knowledge
- Provides credible intervals (not asymptotic)

Markov Chain Monte Carlo (MCMC)

- **MCMC**: simulates a Markov chain to sample from the posterior
  - Used when the posterior is hard to compute directly
- **Gibbs sampling**: a simple and efficient MCMC method
  - Works when each parameter has a standard full conditional (e.g., Normal)

Introduction

Methodological Roadmap



Naive Bayesian Model

- **Model:**

$$\mathbf{Y}_{n\times 1} = \mathbf{X}_{n\times p}\boldsymbol{\beta}_{p\times 1} + \boldsymbol{\varepsilon}_{n\times 1} \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}_n(\mathbf{0},\ \sigma_\varepsilon^2\mathbf{I}_n)$$

  - Assumes exogeneity: $\mathbb{E}[\mathbf{X}^\top\boldsymbol{\varepsilon}] = 0$
  - **Likelihood:** $p(\mathbf{Y} \mid \mathbf{X}, \boldsymbol{\beta}, \sigma^2)$
  - **Priors:**

$$p(\boldsymbol{\beta}) = \mathcal{N}_p(\boldsymbol{\mu}_\beta, \sigma_\beta^2\mathbf{I}_p)$$

$$p(\sigma^2) = \text{Inverse-Gamma}(a, b)$$

$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}$

  - **Posterior:** prior family with data-updated parameters; inference via Gibbs sampling

- **Why it fails when endogeneity exists?**
  - This model ignores how **X** is generated
  - If **X** is endogenous, estimates of **β** are biased
  - No way to account for omitted variables and measurement error

Full Bayesian IV Model

- To address the endogeneity in **X**, we use instrumental variables **Z:**
  - **Z** is correlated with **X**
  - **Z** is uncorrelated with both **U** and **ε**
  - These assumptions ensure that **Z** provides valid variation for identifying **β**
- **Model:**

$$\mathbf{X}_{n\times p} = \mathbf{Z}_{n\times q}\boldsymbol{\Gamma}_{q\times p} + \mathbf{U}_{n\times p}$$

$$\mathbf{Y}_{n\times 1} = \mathbf{X}_{n\times p}\boldsymbol{\beta}_{p\times 1} + \boldsymbol{\varepsilon}_{n\times 1}$$

with

$$\begin{pmatrix} \boldsymbol{U} \\ \boldsymbol{\varepsilon} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_u & \boldsymbol{\Sigma}_{u\varepsilon} \\ \boldsymbol{\Sigma}_{u\varepsilon}^\top & \sigma_\varepsilon^2 \end{pmatrix} \right)$$

  - This joint model structure allows endogeneity, i.e., $\boldsymbol{\Sigma}_{u\varepsilon} \neq \mathbf{0}$.
  - **Likelihood:** $p(\mathbf{Y}, \mathbf{X} \mid \mathbf{Z}, \boldsymbol{\beta}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma})$
  - **Priors:**

$$p(\boldsymbol{\beta}) = \mathcal{N}_p(\boldsymbol{\mu}_\beta, \sigma_\beta^2 \mathbf{I}_p)$$

$$p(\boldsymbol{\Gamma}) = \mathcal{MN}_{q\times p}(\boldsymbol{\mu}_\Gamma, \mathbf{I}_q, \sigma_\Gamma^2 \mathbf{I}_p)$$

$$p(\boldsymbol{\Sigma}) = \text{Inverse-Wishart}_{p+1}(\nu_0, \boldsymbol{\Psi}_0)$$

where

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_u & \boldsymbol{\Sigma}_{u\varepsilon} \\ \boldsymbol{\Sigma}_{u\varepsilon}^\top & \sigma_\varepsilon^2 \end{bmatrix}$$

$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}$

  - **Posterior:**

$$p(\boldsymbol{\beta} \mid \mathbf{Y}, \mathbf{X}, \mathbf{Z}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma}) = \mathcal{N}_p(\cdot, \cdot)$$

$$p(\text{vec}(\boldsymbol{\Gamma}) \mid \mathbf{Y}, \mathbf{X}, \mathbf{Z}, \boldsymbol{\beta}, \boldsymbol{\Sigma}) = \mathcal{N}_{pq}(\cdot, \cdot)$$

$$p(\boldsymbol{\Sigma} \mid \mathbf{Y}, \mathbf{X}, \mathbf{Z}, \boldsymbol{\beta}, \boldsymbol{\Gamma}) = \text{Inverse-Wishart}_{p+1}(\cdot, \cdot)$$

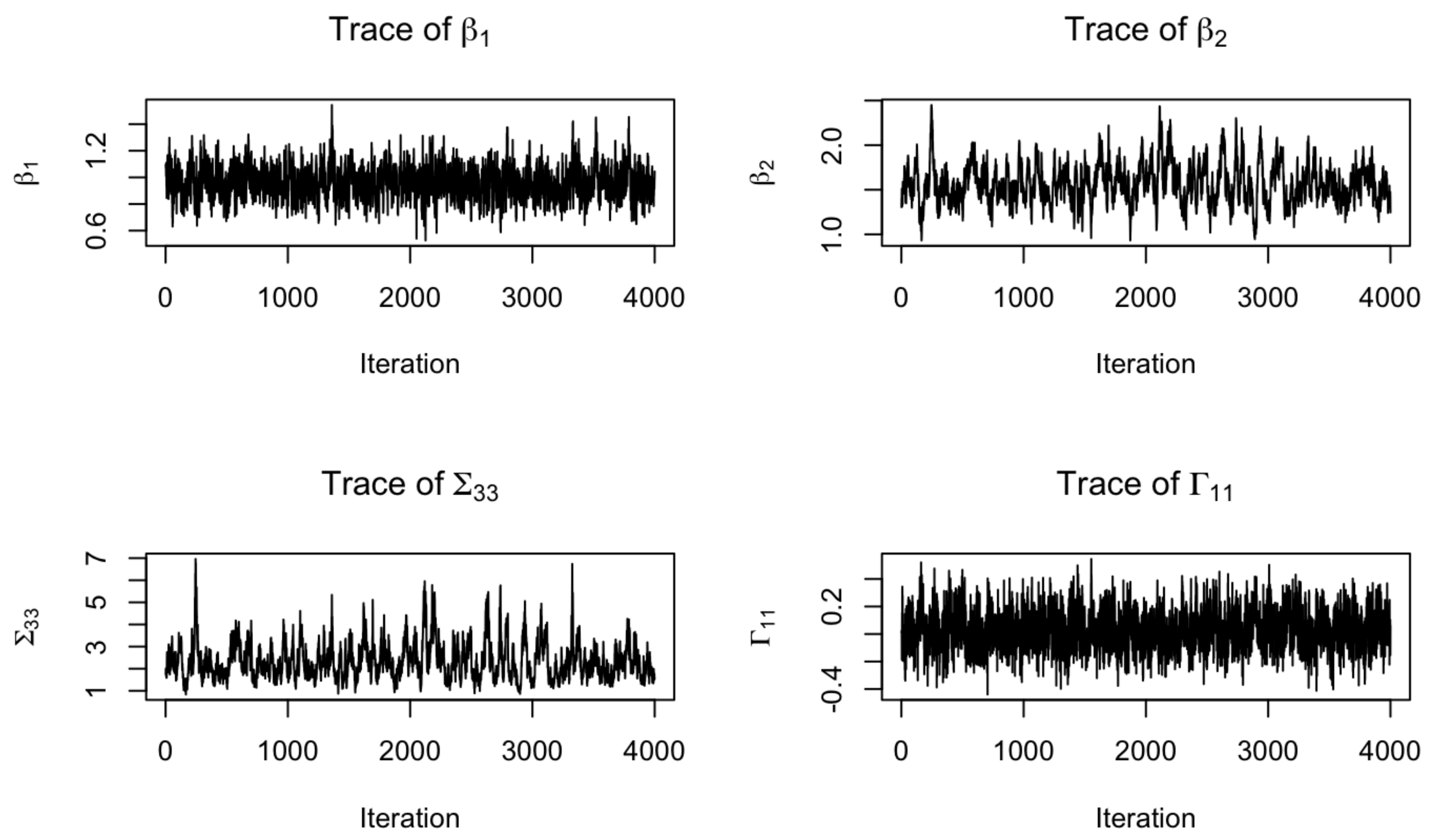
inference via Gibbs sampling

Full Bayesian IV Model

Limitations of Full Bayesian IV in Practice

- **Simulation setup:** n=100, p=2, q=3; Gibbs sampling: 5,000 iterations, 1,000 burn-in
- **Simulation results:**

Parameter	True Value	Posterior Mean
β	$\begin{bmatrix} 1.00 \\ 1.00 \end{bmatrix}$	$\begin{bmatrix} 0.9629 \\ 1.5687 \end{bmatrix}$
$\mathbf{\Gamma}$	$\begin{bmatrix} 1.0 & 0.5 \\ 0.3 & -0.2 \\ 0.0 & 0.8 \end{bmatrix}$	$\begin{bmatrix} 0.7204 & 0.0258 \\ 0.3905 & -0.0517 \\ -0.3170 & 0.2023 \end{bmatrix}$
$\mathbf{\Sigma}$	$\begin{bmatrix} 1.0 & 0.5 & -0.4 \\ 0.5 & 1.0 & -0.4 \\ -0.4 & -0.4 & 1.0 \end{bmatrix}$	$\begin{bmatrix} 1.0979 & 0.9194 & -0.8516 \\ 0.9194 & 1.8031 & -1.5564 \\ -0.8516 & -1.5564 & 2.3433 \end{bmatrix}$



- **Discussion:**
  - Posterior means **deviate** from truth, especially for **Γ, Σ**
  - Trace plots indicate **poor convergence** for some parameters

- **Full Bayesian** showed poor mixing and slow convergence
- This led us to try a **simplified** version, which worked better empirically

Simplified Bayesian IV Model

- **Model:**
  - Same **model specification** and **priors** as full Bayesian
  - **Posterior:**
    - Only the conditional posterior of **Γ** is modified:

$$p(\mathbf{\Gamma} \mid \mathbf{X}, \mathbf{Z}, \mathbf{\Sigma}_u) \propto p(\mathbf{X} \mid \mathbf{Z}, \mathbf{\Gamma}, \mathbf{\Sigma}_u) \cdot p(\mathbf{\Gamma}) \quad \Rightarrow \quad \mathbf{\Gamma} \mid \mathbf{X}, \mathbf{Z}, \mathbf{\Sigma}_u \sim \mathcal{MN}_{q\times p}(\cdot, \cdot, \cdot)$$

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}$$

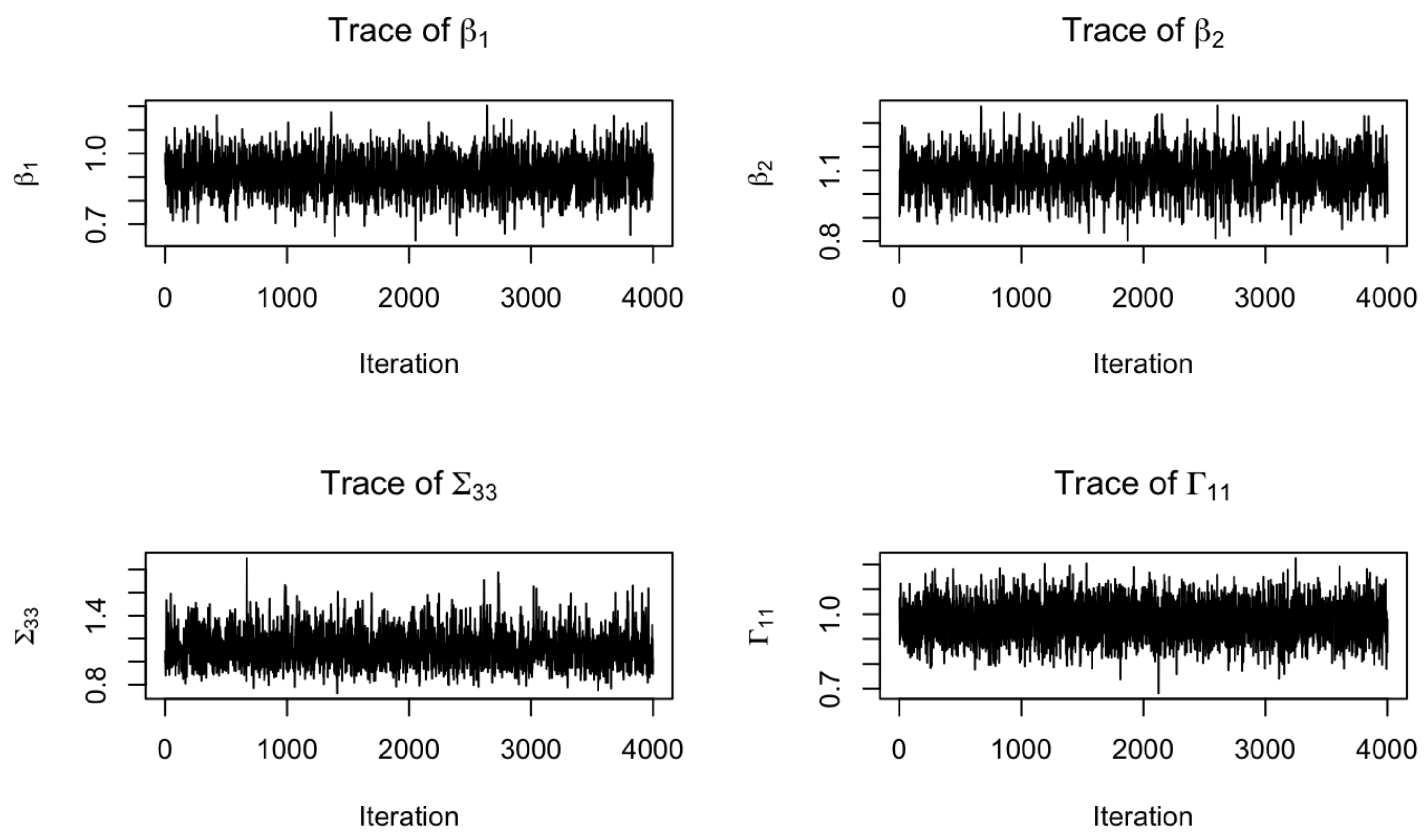
    - **Γ** is modified to no longer depend on **β** and $\mathbf{\Sigma}_{u\varepsilon}$

Simplified Bayesian IV Model

Simplified Bayesian IV in Practice

- **Simulation setup:** n=100, p=2, q=3, same as full Bayesian approach
- **Simulation results:**

Parameter	True Value	Posterior Mean
$\boldsymbol{\beta}$	$\begin{bmatrix} 1.00 \\ 1.00 \end{bmatrix}$	$\begin{bmatrix} 0.9127 \\ 1.0821 \end{bmatrix}$
$\boldsymbol{\Gamma}$	$\begin{bmatrix} 1.0 & 0.5 \\ 0.3 & -0.2 \\ 0.0 & 0.8 \end{bmatrix}$	$\begin{bmatrix} 0.9746 & 0.4807 \\ 0.3328 & -0.1631 \\ 0.0111 & 0.8081 \end{bmatrix}$
$\boldsymbol{\Sigma}$	$\begin{bmatrix} 1.0 & 0.5 & -0.4 \\ 0.5 & 1.0 & -0.4 \\ -0.4 & -0.4 & 1.0 \end{bmatrix}$	$\begin{bmatrix} 0.9318 & 0.6202 & -0.3208 \\ 0.6202 & 1.2536 & -0.5828 \\ -0.3208 & -0.5828 & 1.1137 \end{bmatrix}$



- **Discussion:**
  - Estimates are **close** to true values
  - MCMC chains **mix better** than full Bayesian approach

Simulation Study

- **Design:**
  - Compare all three models
  - Settings: n = 100 and 500; p = 2, 3, 5, q = 3; Gibbs sampling: 5,000 iterations, 1,000 burn-in
  - True parameters：

$$\boldsymbol{\beta}_{\text{true}} = \mathbf{1}_p$$

$$\boldsymbol{\Gamma}_{\text{true}} \sim \text{Uniform}(0,1) \text{ element-wise}$$

$$\boldsymbol{\Sigma}_u\text{: AR(1) with } (\Sigma_u)_{ij} = 0.5^{|i-j|},\ \sigma_\varepsilon^2 = 1$$

$$\boldsymbol{\Sigma}_{u\varepsilon} = (-0.4, \ldots, -0.4)^T$$

  - 100 replications per setting

- **Evaluation metrics:**
  - Bias: average posterior error from true value
  - MSE: mean squared error across replications
  - Coverage: % of 95% credible intervals containing true value

Simulation Results

- We show results for p=2,q=3 as a representative example; other settings yield similar patterns

- Key takeaways:
  - **Simplified Bayesian** achieves the lowest bias and MSE for most β's
  - Coverage is consistently close to the nominal 95%
  - Improvement is most pronounced when sample size increases

- Considering all metrics together, the **Simplified Bayesian** approach provides the best overall performance

Table 2.4.1: Simulation results for $p = 2$, $q = 3$

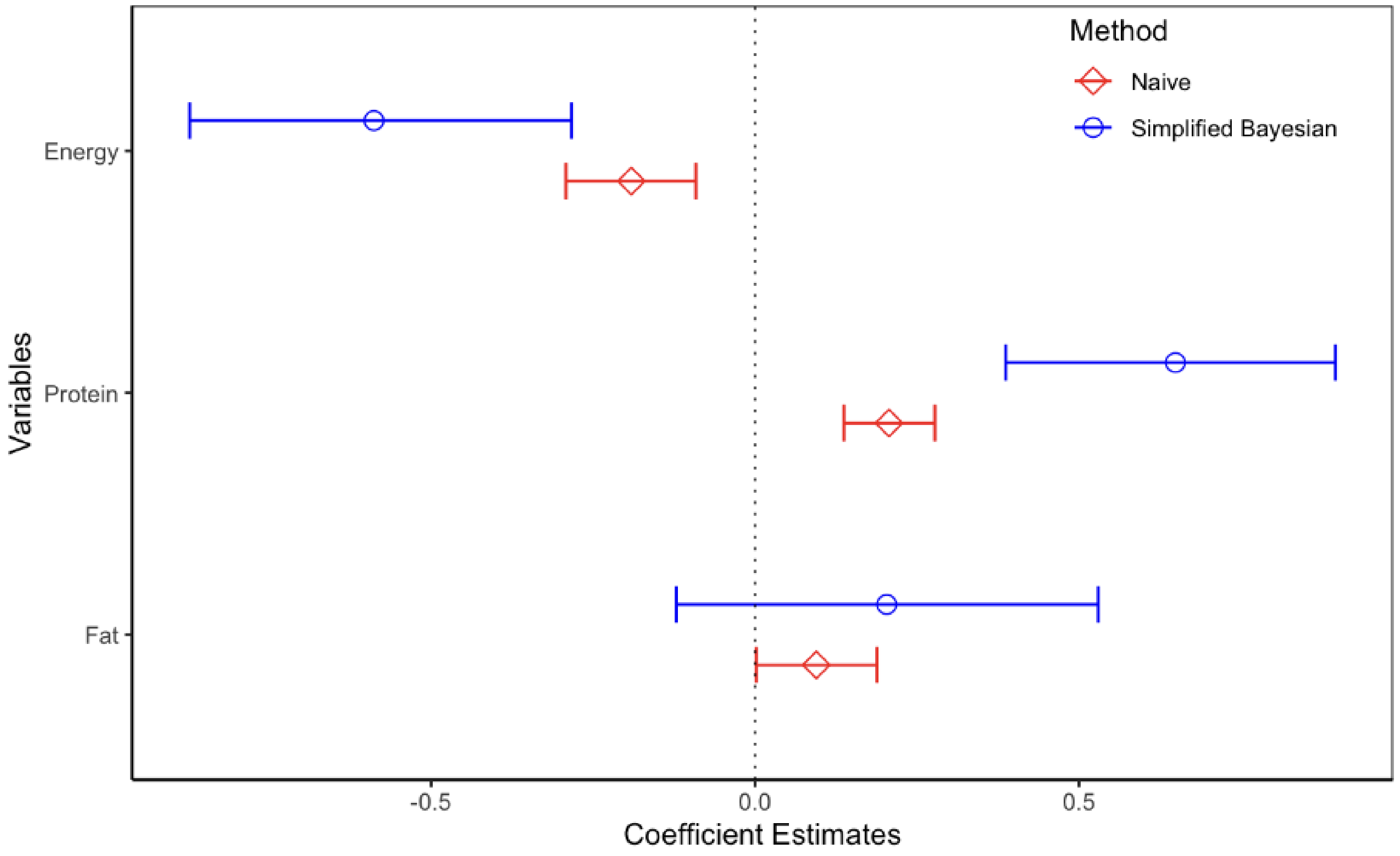
n	Metric	Method	β_1	β_2
100	Bias	Naive	-0.1503	-0.1313
		Full Bayesian	0.1270	0.1972
		Simplified Bayesian	-0.0525	-0.0119
	MSE	Naive	0.0373	0.0334
		Full Bayesian	0.1299	0.1396
		Simplified Bayesian	0.0568	0.0443
	Coverage	Naive	0.67	0.65
		Full Bayesian	0.79	0.76
		Simplified Bayesian	0.94	0.96
500	Bias	Naive	-0.1470	-0.1243
		Full Bayesian	0.5448	0.6446
		Simplified Bayesian	-0.0247	0.0097
	MSE	Naive	0.0266	0.0221
		Full Bayesian	0.3996	0.5719
		Simplified Bayesian	0.0174	0.0184
	Coverage	Naive	0.17	0.27
		Full Bayesian	0.25	0.13
		Simplified Bayesian	0.95	0.93

NHANES data analysis

- **Study background:**
  - Dataset: NHANES 2009–2010
  - Objective: Estimate effect of long-term nutrient intake (energy, protein, fat) on BMI
  - Challenge: 24 - hour recalls are noisy proxies → measurement error → endogeneity
- **Model setup:**
  - Outcome model: $y_i = \mathbf{W}_i^\top \boldsymbol{\beta} + \tilde{\varepsilon}_i$
  - Measurement model: $\mathbf{D}_{ij} = \mathbf{W}_i + \mathbf{V}_{ij}, \quad j = 1, 2$
  - IV setup: $\mathbf{X} = \mathbf{D}_{i1}, \quad \mathbf{Z} = \mathbf{D}_{i2}$
- **Key takeaways:**
  - **Simplified Bayesian** gives stronger effects
  - **Naive** model underestimates effect size and uncertainty

Table 2.5.1: Posterior estimates for BMI model under Naive and Simplified Bayesian

Method	Variable	Posterior Mean	95% CI Lower	95% CI Upper
Naive	Energy	−0.190	−0.290	−0.091
	Protein	0.206	0.134	0.277
	Fat	0.094	0.004	0.188
Simplified Bayesian	Energy	−0.611	−0.914	−0.303
	Protein	0.672	0.397	0.942
	Fat	0.212	−0.109	0.544



Bayesian Lasso IV Model

- To incorporate **variable selection** when some components of **β** may be exactly zero
- **Model:**
  - Based on **simplified Bayesian** model, BUT change the prior distribution of **β**
  - Same **likelihood** and **priors** for **Γ** and **Σ** as in the simplified Bayesian model
  - **Prior:**

$$\begin{aligned}
\boldsymbol{\beta} \mid \tau_1^2, \ldots, \tau_p^2 &\sim \mathcal{N}_p(\mathbf{0}, \mathbf{D}_\tau), \quad \mathbf{D}_\tau = \operatorname{diag}(\tau_1^2, \ldots, \tau_p^2) \\
\tau_j^2 &\sim \text{Exponential}(\lambda^2/2), \quad j = 1, \ldots, p \\
\lambda^2 &\sim \text{Gamma}(r, \delta)
\end{aligned}$$

  - **Posterior:**

$$\begin{aligned}
\boldsymbol{\beta} \mid \cdot &\sim \mathcal{N}_p(\cdot, \cdot) \\
1/\tau_j^2 \mid \cdot &\sim \text{Inverse-Gaussian}(\cdot, \cdot) \\
\lambda^2 \mid \cdot &\sim \text{Gamma}(\cdot, \cdot)
\end{aligned}$$

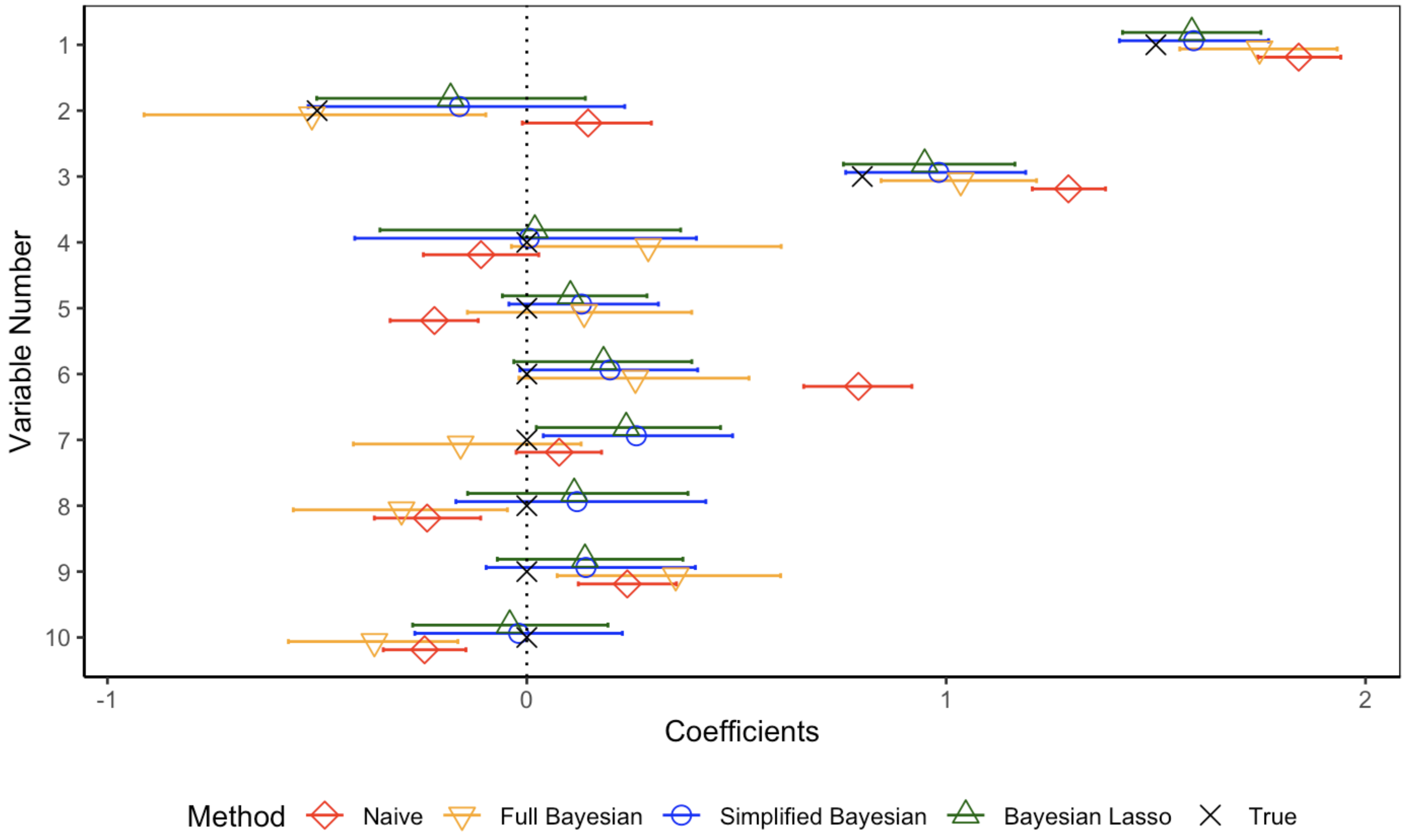
Simulation Study

Single Replication (Visual Insight):

- **Design:**
  - Settings: p = q = 10 ; n = 500
  - True parameters：

$$\boldsymbol{\beta}_{\text{true}} = (1.5,\ -0.5,\ 0.8,\ 0, \ldots, 0)^T$$
$$\mathbf{Z}_{ij} \sim \mathcal{N}(0,1) \quad \text{i.i.d.}$$
$$\mathbf{X} = \mathbf{Z}\boldsymbol{\Gamma} + \mathbf{U}$$
$$(\boldsymbol{\varepsilon}, \mathbf{U}) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$$
$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}_{\text{true}} + \boldsymbol{\varepsilon}$$

  - Sparse setting: only β_1, β_2, β_3 are nonzero



- **Key takeaways:**
  - **Bayesian Lasso** produces tight intervals for irrelevant (zero) coefficients
  - Less shrinkage for true signals compared to **Naive** and **Full Bayesian**

Simulation Study

Aggregated Performance (100 Simulations):

- **Design:**
  - Same setting and true parameters as in **Single Replication**
- **Evaluation metrics:**
  - Bias, MSE, Coverage
  - TP (True Positives), FP (False Positives), FPR (False Positive Rate), FNR (False Negative Rate)
  - Precision: TP divided by the total number of selected coefficients

- **Key takeaways:**
  - Bayesian Lasso gives best overall performance
  - Accurately selects true signals with low false positives
  - Balances estimation accuracy and sparsity well

Table 3.5.1: Overall performance metrics across 100 simulations

Method	MSE	Coverage	FPR	FNR	Precision	TP
Naive	0.9978	0.1740	0.8171	0.0333	0.3412	2.97
Full Bayesian	0.9379	0.3900	0.6071	0.0467	0.4248	2.95
Simplified Bayesian	0.5189	0.5890	0.4129	0.0067	0.5542	2.99
Bayesian Lasso	0.4942	0.5990	0.4029	0.0067	0.5603	2.99

Table 3.5.2: Average bias by coefficient across 100 simulations

Method	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
Naive	-0.027	0.020	-0.027	0.058	-0.048	-0.003	-0.018	0.049	-0.050	0.024
Full Bayesian	0.086	-0.030	0.043	0.051	-0.055	-0.010	-0.029	0.045	-0.041	0.033
Simplified Bayesian	-0.010	-0.010	-0.030	0.019	-0.038	-0.013	0.019	0.002	-0.023	0.013
Bayesian Lasso	-0.018	-0.003	-0.031	0.020	-0.038	-0.010	0.019	0.002	-0.024	0.016

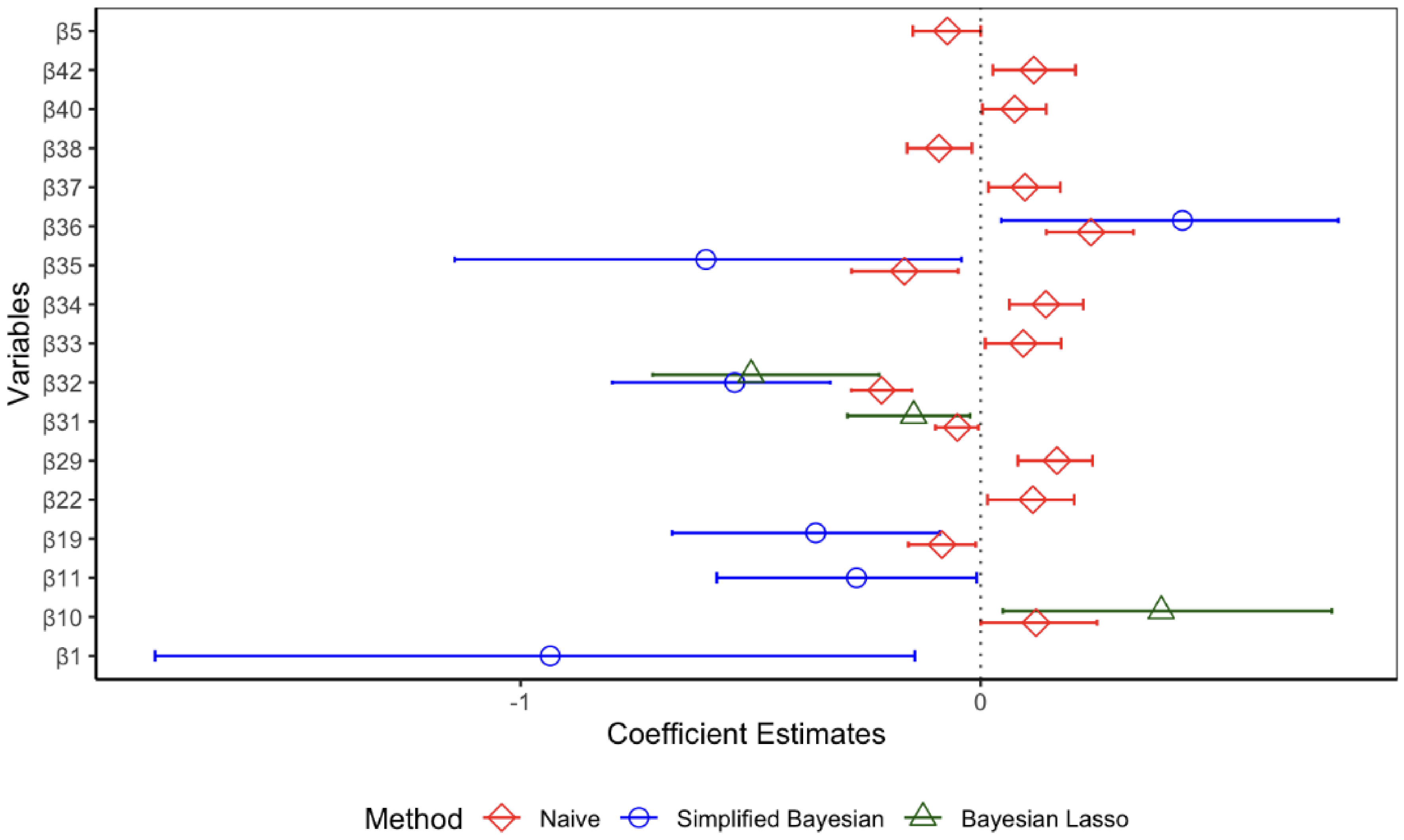
NHANES data analysis

- **Data set:**
  - NHANES dietary data, 42 nutrient intake variables
  - **X**: Day 1 intakes (endogenous)
    Z: Day 2 intakes (instruments)
  - All variables standardized; no missing values

- **Key takeaways:**
  - Naive: 15 variables (risk of over-selection)
  - Simplified Bayesian: 6 variables (more conservative)
  - Bayesian Lasso select 3 robust variables (most selective):
    β_{10} – Cholesterol (mg)
    β_{31} – Added vitamin B12 (mcg)
    β_{32} – Vitamin C (mg)

Number and identity of variables with credible intervals excluding zero

Method	Significant Variables
Naive	β_5, β_{10}, β_{19}, β_{22}, β_{29}, β_{31}, β_{32}, β_{33}, β_{34}, β_{35}, β_{36}, β_{37}, β_{38}, β_{40}, β_{42}
Simplified Bayesian	β_1, β_{11}, β_{19}, β_{32}, β_{35}, β_{36}
Bayesian Lasso	β_{10}, β_{31}, β_{32}



Summary

Modeling Framework

- Developed Bayesian IV models to correct for endogeneity in regression
- Introduced a **Simplified Bayesian model** for computational efficiency
- Extended to **Bayesian Lasso IV** for **sparse settings**

Simulation Results

- **Simplified Bayesian**: good bias–variance balance, reliable coverage
- **Bayesian Lasso**: accurate selection under sparsity

Real Data Insights (In NHANES data)

- Simplified Bayesian detects stronger effects of energy and protein
- Bayesian Lasso selects a smaller, more robust subset of nutrients

Takeaway

- **Simplified Bayesian** is a practical and reliable baseline
- **Bayesian Lasso** is preferred when **sparsity** is expected
- Correcting for endogeneity is **crucial** in dietary regression

Thank you！