20592 - Probability and Statistics

Final Project Computational Module

Group 8 - Report

January, 2023

1 Introduction

The data provided contain information about students' math scores from a sample of 100 schools in the United States. Therefore, the data are hierarchical, with a population of schools and a population of students within each school.

$$Y_{i,j} = \text{the score of student i in school j}$$
$$\mathbf{x}_{i,j} = \text{covariates of student i in school j}$$
$$m = \text{number of schools}$$
$$n_j = \text{number of students' scores in school j}$$

The goal is to estimate global and school-specific effects so that the performance of different schools can be analyzed and compared.

2 Hierarchical model without covariates

The hierarchical normal model is used to describe the difference of means across populations.

$$Y_{i,j}|\ \theta_j, \sigma^2 \sim N(y|\ \theta_j, \sigma^2), \quad i = 1, ..., n_j, j = 1, ..., m$$
$$\theta_j|\ \mu, \tau^2 \sim N(\theta|\ \mu, \tau^2)$$
$$(\mu, \tau^2, \sigma^2) \sim p_0$$

The fixed but unknown quantities of the system are the within-group sampling variability σ^2, the group-specific means $\{\theta_1, ..., \theta_m\}$, and the mean and variance of the population of group-specific means (μ, τ^2). The within-group sampling variability σ^2 is assumed to be the same across schools.

A Gibbs sampler is used to approximate the joint posterior distribution of these unknown parameters. The following factorization of the posterior can be used to simplify the computation of the full conditionals:

$$p(\theta_1, ..., \theta_m, \mu, \tau^2, \sigma^2 |\ \boldsymbol{y}_1, ..., \boldsymbol{y}_m)$$

$$\propto p(\mu, \tau^2, \sigma^2) \times p(\theta_1, ..., \theta_m |\ \mu, \tau^2, \sigma^2) \times p(\boldsymbol{y}_1, ..., \boldsymbol{y}_m |\ \theta_1, ..., \theta_m, \mu, \tau^2, \sigma^2)$$

$$= p(\mu)p(\tau^2)p(\sigma^2) \left\{ \prod_{j=1}^{m} p(\theta_j |\ \mu, \tau^2) \right\} \left\{ \prod_{j=1}^{m} \prod_{i=1}^{n_j} p(y_{i,j} |\ \theta_j, \sigma^2) \right\}$$

The second term stems from the two following assumptions:

- conditionally on $(\theta_1, ..., \theta_m, \mu, \tau^2, \sigma^2)$, $Y_{i,j}, ..., Y_{n_j,j}$ are independent.
- μ and τ^2 do not directly provide information about $\boldsymbol{Y}_j$, but only indirectly through θ_j

The following prior distributions were chosen for the parameters:

$$\frac{1}{\sigma^2} \sim Gamma\left(\frac{\nu_0}{2}, \nu_0 \frac{\sigma_0^2}{2}\right)$$

$$\frac{1}{\tau^2} \sim Gamma\left(\frac{\eta_0}{2}, \eta_0 \frac{\tau_0^2}{2}\right)$$

$$\mu \sim N(\mu_0, \gamma_0^2)$$

Notice that the following parametrization of the Gamma distribution is used

$$X \sim Gamma(\alpha, \beta),\ f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$$

where β is the rate parameter of the Gamma distribution, equal to 1/scale.
Thus,

$$1/X \sim Inv\text{-}Gamma(\alpha, \beta),\ f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} \left(\frac{1}{x}\right)^{\alpha+1} e^{-\frac{\beta}{x}}$$

where β is the scale parameter of the Inverse Gamma distribution.

2.1 Full conditionals of μ and τ^2

As a function of μ and τ^2, the factorization of the posterior distribution is proportional to:

$$p(\mu)p(\tau^2) \prod_{j=1}^{m} p(\theta_j |\ \mu, \tau^2)$$

Therefore, the full conditional of μ is such that:

$$p(\mu|\ \theta_1, ..., \theta_m, \tau^2, \sigma^2, \boldsymbol{y}_1, ..., \boldsymbol{y}_m) \propto p(\mu) \prod_{j=1}^{m} p(\theta_j|\ \mu, \tau^2)$$

$$\propto exp\left\{-\frac{1}{2\gamma_0^2}(\mu - \mu_0)^2\right\} exp\left\{-\frac{1}{2\tau^2}\sum_{j=1}^{m}(\theta_j - \mu)^2\right\}$$

$$= exp\left\{-\frac{1}{2\gamma_0^2}(\mu - \mu_0)^2 - \frac{1}{2\tau^2}\sum_{j=1}^{m}(\theta_j - \mu)^2\right\}$$

$$= exp\left\{-\frac{1}{2}\left(\frac{1}{\gamma_0^2}(\mu - \mu_0)^2 + \frac{1}{\tau^2}\sum_{j=1}^{m}(\theta_j - \mu)^2\right)\right\}$$

Ignoring the $-1/2$ and looking at the argument of the exponential

$$\frac{1}{\gamma_0^2}(\mu - \mu_0)^2 + \frac{1}{\tau^2}\sum_{j=1}^{m}(\theta_j - \mu)^2$$

$$= \frac{1}{\gamma_0^2}(\mu^2 + \mu_0^2 - 2\mu\mu_0) + \frac{1}{\tau^2}\left(\sum_{j=1}^{m}\theta_j^2 - 2\mu\sum_{j=1}^{m}\theta_j + m\mu^2\right)$$

$$= a\mu^2 - 2b\mu + c$$

where

$$a = \frac{1}{\gamma_0^2} + \frac{m}{\tau^2}$$

$$b = \frac{\mu_0}{\gamma_0^2} + \frac{\sum_{j=1}^{m}\theta_j}{\tau^2}$$

$$c = c(\mu_0, \tau^2, \gamma_0^2, \theta_1, ..., \theta_m)$$

Then

$$p(\mu|\ \theta_1, ..., \theta_m, \tau^2, \sigma^2, \boldsymbol{y}_1, ..., \boldsymbol{y}_m) \propto exp\left\{-\frac{1}{2}(a\mu^2 - 2b\mu)\right\}$$

$$= exp\left\{-\frac{1}{2}(a\mu^2 - 2b\mu) - \frac{1}{2}\left(\frac{b^2}{a}\right) + \frac{1}{2}\left(\frac{b^2}{a}\right)\right\}$$

$$= exp\left\{-\frac{1}{2}a\left(\mu^2 - \frac{2b\mu}{a} + \frac{b^2}{a^2}\right) + \frac{1}{2}\left(\frac{b^2}{a}\right)\right\}$$

$$\propto exp\left\{-\frac{1}{2}a\left(\mu - \frac{b}{a}\right)^2\right\}$$

which is the kernel of a normal distribution $N(\mu_n, \tau_n^2)$ with mean $\mu_n = \frac{b}{a}$ and variance $\tau_n^2 = \frac{1}{a}$.

$$\left\{\mu|\ \theta_1, ..., \theta_m, \tau^2\right\} \sim N\left(\frac{\frac{\mu_0}{\gamma_0^2} + \frac{m}{\tau^2}\bar{\theta}}{\frac{1}{\gamma_0^2} + \frac{m}{\tau^2}}, \left(\frac{1}{\gamma_0^2} + \frac{m}{\tau^2}\right)^{-1}\right)$$

Similarly, the full conditional of τ^2 is such that:

$$p(\tau^2|\ \theta_1, ..., \theta_m, \mu, \sigma^2, \boldsymbol{y}_1, ..., \boldsymbol{y}_m) \propto p(\tau^2) \prod_{j=1}^{m} p(\theta_j|\ \mu, \tau^2)$$

where

$$\tau^2 \sim Inv\text{-}Gamma\left(\frac{\eta_0}{2}, \eta_0 \frac{\tau_0^2}{2}\right)$$

$$p(\tau^2|\ \theta_1, ..., \theta_m, \mu, \sigma^2, \boldsymbol{y}_1, ..., \boldsymbol{y}_m) \propto p(\tau^2)$$

$$\propto (\tau^2)^{-\frac{\eta_0}{2}-1} exp\left\{-\frac{\eta_0\tau_0^2}{2\tau^2}\right\} (\tau^2)^{-\frac{m}{2}} exp\left\{-\frac{1}{2\tau^2}\sum_{j=1}^{m}(\theta_j - \mu)^2\right\}$$

$$= (\tau^2)^{-\left(\frac{\eta_0}{2}+\frac{m}{2}\right)-1} exp\left\{-\frac{1}{\tau^2}\left(\frac{\eta_0\tau_0^2}{2} + \frac{\sum_{j=1}^{m}(\theta_j - \mu)^2}{2}\right)\right\}$$

which is the kernel of an inverse gamma distribution with

$$\alpha = \frac{\eta_0}{2} + \frac{m}{2}$$

$$\beta = \frac{\eta_0\tau_0^2}{2} + \frac{\sum_{j=1}^{m}(\theta_j - \mu)^2}{2}$$

Therefore

$$\left\{1/\tau^2|\ \theta_1, ..., \theta_m, \mu\right\} \sim Gamma\left(\frac{\eta_0 + m}{2}, \frac{\eta_0\tau_0^2 + \sum_{j=1}^{m}(\theta_j - \mu)^2}{2}\right)$$

where β is the rate parameter of the Gamma distribution.

2.2 Full conditional of θ_j

Given the factorization of the posterior distribution, the full conditional of θ_j should be proportional to

$$p(\theta_j|\ \mu, \tau^2, \sigma^2, \boldsymbol{y}_1, ..., \boldsymbol{y}_m) \propto p(\theta_j|\ \mu, \tau^2) \prod_{i=1}^{n_j} p(y_{i,j}|\ \theta_j, \sigma^2)$$

Therefore, conditional on $\{\mu, \tau^2, \sigma^2, \boldsymbol{y}_j\}$, each θ_j must be independent of any other θ_k, $k \neq j$ and of the data coming from groups other than j. In other words, the θ's do not provide information about each other directly but only through μ, τ^2 and σ^2.

$$p(\theta_j|\ \mu, \tau^2) \prod_{i=1}^{n_j} p(y_{i,j}|\ \theta_j, \sigma^2)$$

$$\propto exp\left\{-\frac{1}{2\tau^2}\left(\theta_j - \mu\right)^2\right\} exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n_j}\left(y_{i,j} - \theta_j\right)^2\right\}$$

$$= exp\left\{-\frac{1}{2}\left(\frac{1}{\tau^2}\left(\theta_j - \mu\right)^2 + \frac{1}{\sigma^2}\sum_{i=1}^{n_j}\left(y_{i,j} - \theta_j\right)^2\right)\right\}$$

Ignoring the $-1/2$ and looking at the argument of the exponential

$$\frac{1}{\tau^2}(\theta_j - \mu)^2 + \frac{1}{\sigma^2}\sum_{i=1}^{n_j}(y_{i,j} - \theta_j)^2$$

$$= \frac{1}{\tau^2}(\theta_j^2 + \mu^2 - 2\theta_j\mu) + \frac{1}{\sigma^2}\left(\sum_{i=1}^{n_j} y_{i,j}^2 - 2\theta_j\sum_{i=1}^{n_j} y_{i,j} + n_j\theta_j^2\right)$$

$$= a\theta_j^2 - 2b\theta_j + c$$

where

$$a = \frac{1}{\tau^2} + \frac{n_j}{\sigma^2}$$

$$b = \frac{\mu}{\tau^2} + \frac{\sum_{i=1}^{n_j} y_{i,j}}{\sigma^2}$$

$$c = c(\mu, \tau^2, \sigma^2, y_{1,j}, ..., y_{n_j,j})$$

Then

$$p(\theta_j|\ \mu, \tau^2, \sigma^2, \boldsymbol{y}_1, ..., \boldsymbol{y}_m) \propto exp\left\{-\frac{1}{2}(a\theta_j^2 - 2b\theta_j)\right\}$$

$$= exp\left\{-\frac{1}{2}(a\theta_j^2 - 2b\theta_j) - \frac{1}{2}\left(\frac{b^2}{a}\right) + \frac{1}{2}\left(\frac{b^2}{a}\right)\right\}$$

$$= exp\left\{-\frac{1}{2}a\left(\theta_j^2 - \frac{2b\theta_j}{a} + \frac{b^2}{a^2}\right) + \frac{1}{2}\left(\frac{b^2}{a}\right)\right\}$$

$$\propto exp\left\{-\frac{1}{2}a\left(\theta_j - \frac{b}{a}\right)^2\right\}$$

which is the kernel of a normal distribution with mean $\frac{b}{a}$ and variance $\frac{1}{a}$.

$$\left\{\theta_j|\ y_{1,j},...,y_{n_j,j},\tau^2,\sigma^2\right\} \sim N\left(\frac{\frac{\mu}{\tau^2}+\frac{n_j\bar{y_j}}{\sigma^2}}{\frac{1}{\tau^2}+\frac{n_j}{\sigma^2}}, \left(\frac{1}{\tau^2}+\frac{n_j}{\sigma^2}\right)^{-1}\right)$$

2.3 Full conditional of σ^2

Given $\{\boldsymbol{y}_1,...\boldsymbol{y}_m,\theta_1,...,\theta_m\}$, σ^2 is conditionally independent of $\{\mu,\tau^2\}$. The full conditional distribution of σ^2 should be proportional to

$$p(\sigma^2|\ \boldsymbol{y}_1,...\boldsymbol{y}_m,\theta_1,...,\theta_m) \propto p(\sigma^2)\prod_{j=1}^{m}\prod_{i=1}^{n_j}p(y_{i,j}|\ \theta_j,\sigma^2)$$

where

$$\sigma^2 \sim Inv\text{-}Gamma\left(\frac{\nu_0}{2},\nu_0\frac{\sigma_0^2}{2}\right)$$

$$p(\sigma^2)\prod_{j=1}^{m}\prod_{i=1}^{n_j}p(y_{i,j}|\ \theta_j,\sigma^2)$$

$$\propto (\sigma^2)^{-\frac{\nu_0}{2}-1}exp\left\{-\frac{\nu_0\sigma_0^2}{2\sigma^2}\right\}(\sigma^2)^{-\frac{1}{2}\sum_{j=1}^{m}n_j}exp\left\{-\frac{1}{2\sigma^2}\sum_{j=1}^{m}\sum_{i=1}^{n_j}(y_{i,j}-\theta_j)^2\right\}$$

$$= (\sigma^2)^{-\frac{\nu_0}{2}-\frac{1}{2}\sum_{j=1}^{m}n_j-1}exp\left\{-\frac{1}{\sigma^2}\left(\frac{\nu_0\sigma_0^2}{2}+\frac{1}{2}\sum_{j=1}^{m}\sum_{i=1}^{n_j}(y_{i,j}-\theta_j)^2\right)\right\}$$

which is the kernel of an inverse gamma distribution with

$$\alpha = \frac{1}{2}\left(\nu_0+\sum_{j=1}^{m}n_j\right)$$

$$\beta = \frac{1}{2}\left(\nu_0\sigma_0^2+\sum_{j=1}^{m}\sum_{i=1}^{n_j}(y_{i,j}-\theta_j)^2\right)$$

Therefore

$$\left\{1/\sigma^2|\ \boldsymbol{\theta},\boldsymbol{y}_1,...\boldsymbol{y}_m\right\}$$

$$\sim Gamma\left(\frac{1}{2}\left(\nu_0+\sum_{j=1}^{m}n_j\right),\frac{1}{2}\left(\nu_0\sigma_0^2+\sum_{j=1}^{m}\sum_{i=1}^{n_j}(y_{i,j}-\theta_j)^2\right)\right)$$

where β is the rate parameter of the Gamma distribution.

2.4 Gibbs Sampler

Given the current state of the parameters $\left\{\theta_1^{(t)}, \theta_2^{(t)}, ..., \theta_m^{(t)}, \mu^{(t)}, \tau^{2(t)}, \sigma^{2(t)}\right\}$, at each iteration the new state is generated by:

- sampling $\mu^{(t+1)} \sim p(\mu|\ \theta_1^{(t)}, \theta_2^{(t)}, ..., \theta_m^{(t)}, \tau^{2(t)})$
- sampling $\tau^{2(t+1)} \sim p(\tau^2|\ \theta_1^{(t)}, \theta_2^{(t)}, ..., \theta_m^{(t)}, \mu^{(t+1)})$
- sampling $\sigma^{2(t+1)} \sim p(\sigma^2|\ \theta_1^{(t)}, \theta_2^{(t)}, ..., \theta_m^{(t)}, \boldsymbol{y}_1, ..., \boldsymbol{y}_m)$
- sampling $\theta_j^{(t+1)} \sim p(\theta_j|\ \mu^{(t+1)}, \tau^{2(t+1)}, \sigma^{2(t+1)}, \boldsymbol{y}_j)$, for each $j \in \{1, ..., m\}$

where each parameter gets updated conditional on the most current value of the other parameters. The prior parameters for which a value should be chosen are (ν_0, σ_0^2) for $p(\sigma^2)$, (η_0, τ_0^2) for $p(\tau^2)$, and (μ_0, γ_0^2) for $p(\mu)$.

The math exam was designed to produce a nationwide mean of 50 and a nationwide variance of 100, which includes both within-school and between-school variability. Thus, we set $\mu_0 = 50$ and $\gamma_0^2 = 25$, so that the prior probability that μ is in the interval $(40, 60)$ is about 95%. Assuming that the within school variance is at most 100, we take $\sigma_0^2 = 100$. As this is likely an overestimate, we take $\nu_0 = 1$ so to only weakly concentrate the prior distribution around σ_0^2. Similarly, the variance between schools should not be greater than 100, so we use $\tau_0^2 = 100$ and $\eta_0 = 1$.

A Gibbs Sampler is run for 5,000 iterations after 5,000 iterations of burn-in. The initial values of the parameters are sampled at random from their prior distributions. Before making inference using the resulting MCMC samples, trace plots, box plots and autocorrelation plots are used to see if there is any evidence that the chains have not converged. Each boxplot represents 1000 MCMC samples.

Figure 1 depicts the diagnostic checks for the Gibbs Sampler. Because the sample distribution in each boxplot appears to be the same, stationary has most likely been achieved. The autocorrelation plots suggest the Gibbs sampler moves quickly around the parameter space, never staying in one place for too long.

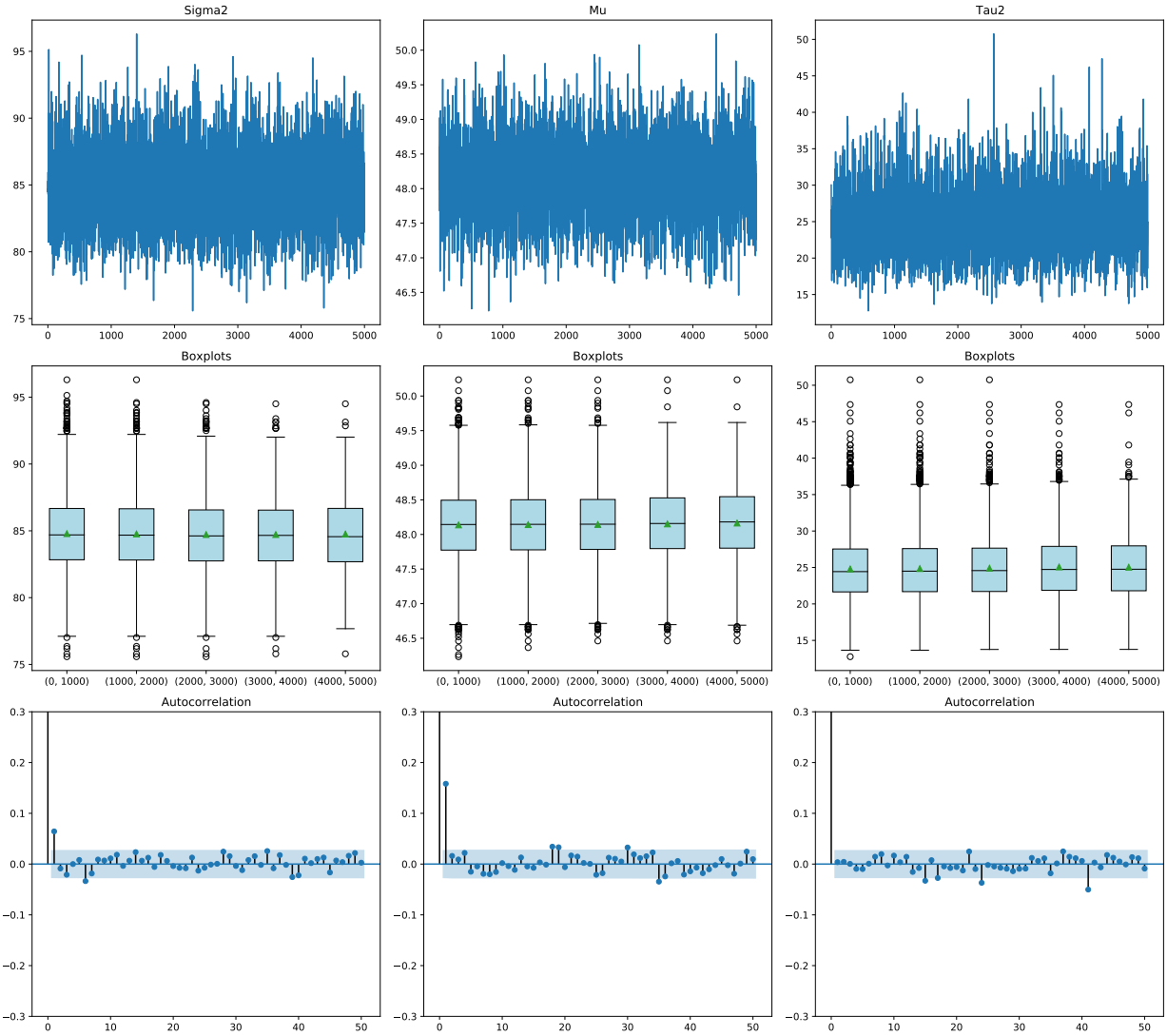

Figure 1: Trace Plots, Box Plots, and Autocorrelation Plots of σ^2, μ, τ^2

As $\sigma_0^2 = 100$ and $\tau_0^2 = 100$ are probably overestimates, another Gibbs Sampler is run with lower within-school and between-school variances. In particular, we have set $\sigma_0^2 = \tau_0^2 = 60$, while maintaining $\nu_0 = 1$ and $\eta_0 = 1$. Moreover, we have considered a 95% prior probability that μ is in the interval $(30, 70)$ by keeping $\mu_0 = 50$ and setting $\gamma_0^2 = 100$. Figure 2 shows the trace plots, box plots, and autocorrelation plots of the three parameter chains.

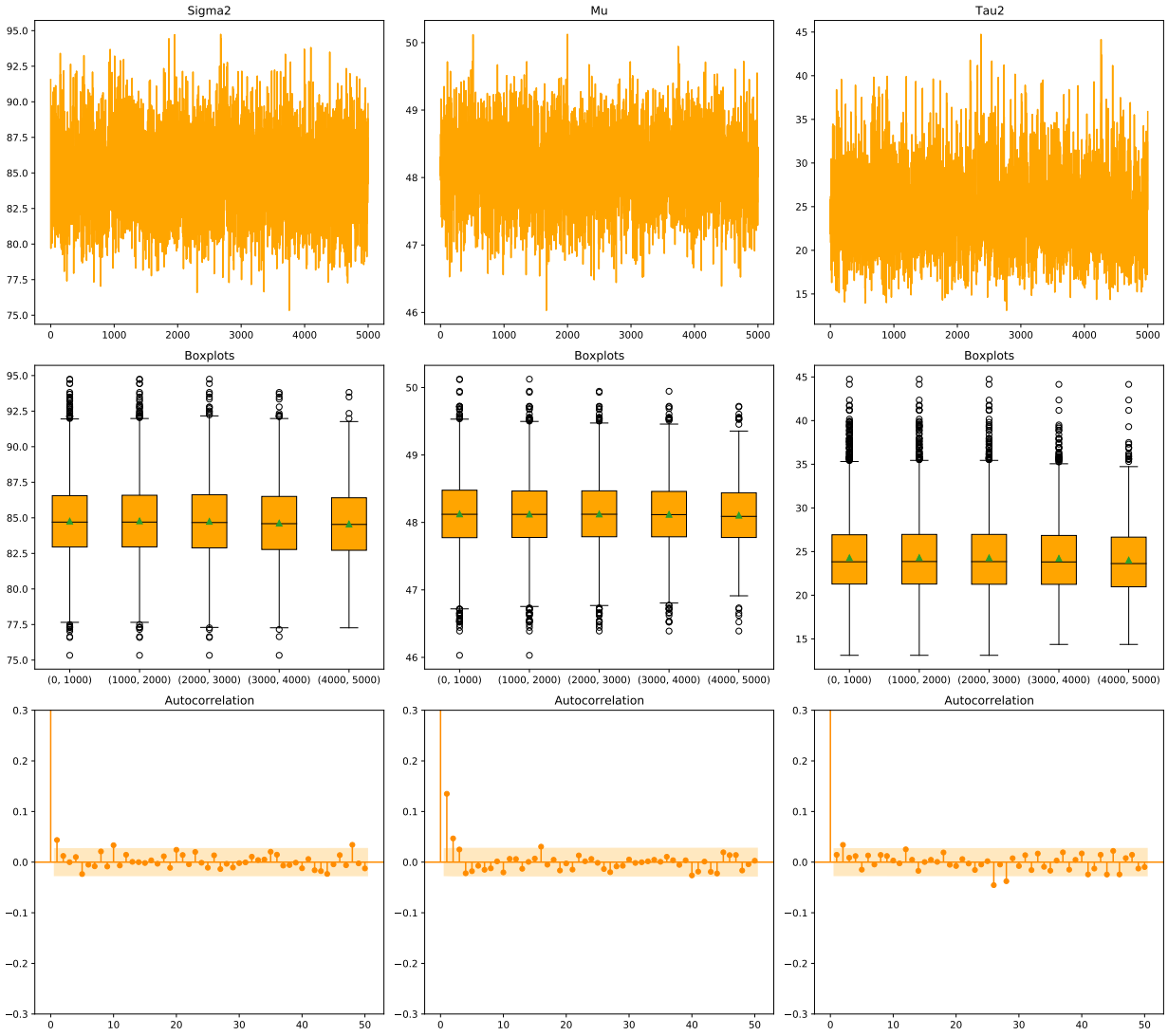

Figure 2: Trace Plots, Box Plots, and Autocorrelation Plots of σ^2, μ, τ^2

According to the trace and box plots, the three chains appear to have converged in the given number of iterations, despite the smaller prior variability. The autocorrelation plots show that, also in this case, the Gibbs Sampler is quick in exploring the parameter space.

Using the same prior parameter values as the first Gibbs Sampler, two different starting points for the iterations are considered. First, we have set the following initial values

for the parameters: $\sigma^2 = \sigma_0^2$, $\mu = \mu_0$, and $\tau^2 = \tau_0^2$. The diagnostic checks of the corresponding Gibbs Sampler are depicted in Figure 3.

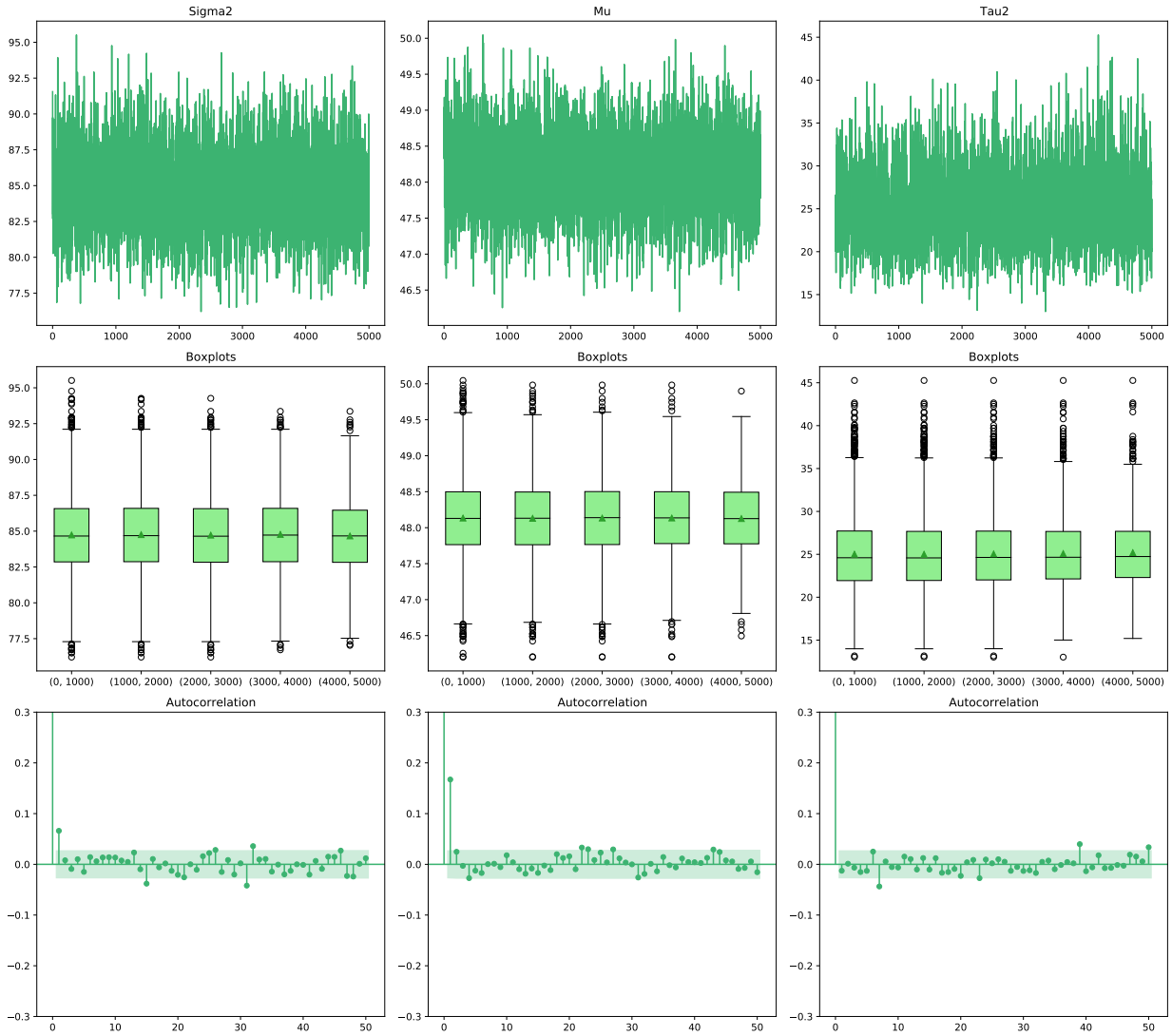

Figure 3: Trace Plots, Box Plots, and Autocorrelation Plots of σ^2, μ, τ^2

Second, we have considered completely non-informative starting values by setting $\sigma^2 = 1$, $\mu = 0$ and $\tau^2 = 1$. Figure 4 shows the corresponding trace plots, boxplots and autocorrelation plots of the parameter chains.

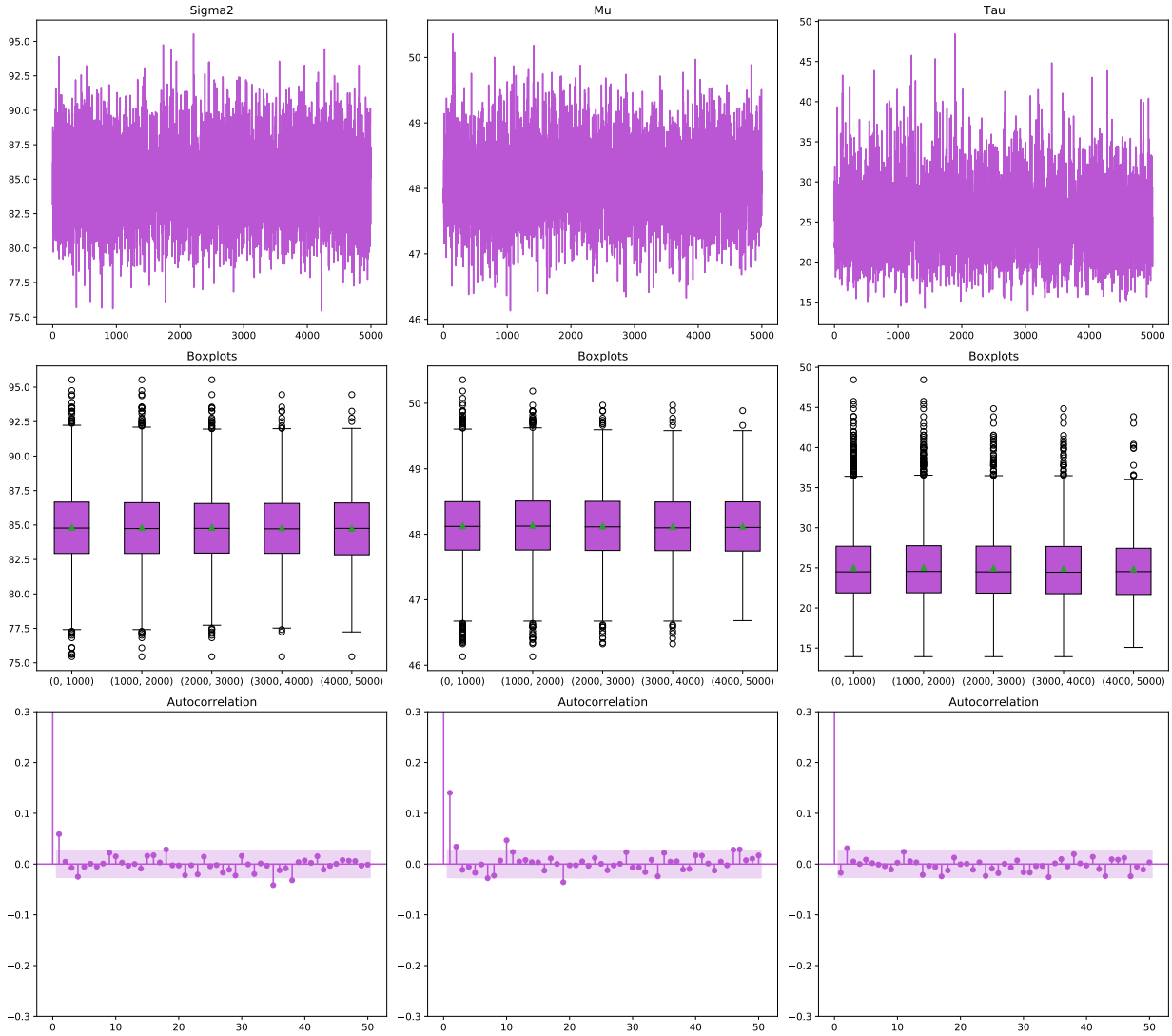

Figure 4: Trace Plots, Box Plots, and Autocorrelation Plots of σ^2, μ, τ^2

Regardless of the starting point for the iterations or the prior parameter values, the trace plots and boxplots show that, for each parameter, the four chains generated by the four Gibbs Samplers appear to converge and bounce around the same average value. Moreover, Figure 5 shows that the estimated posterior densities for σ^2, μ, τ^2 are almost the same for all the four Gibbs Samplers.

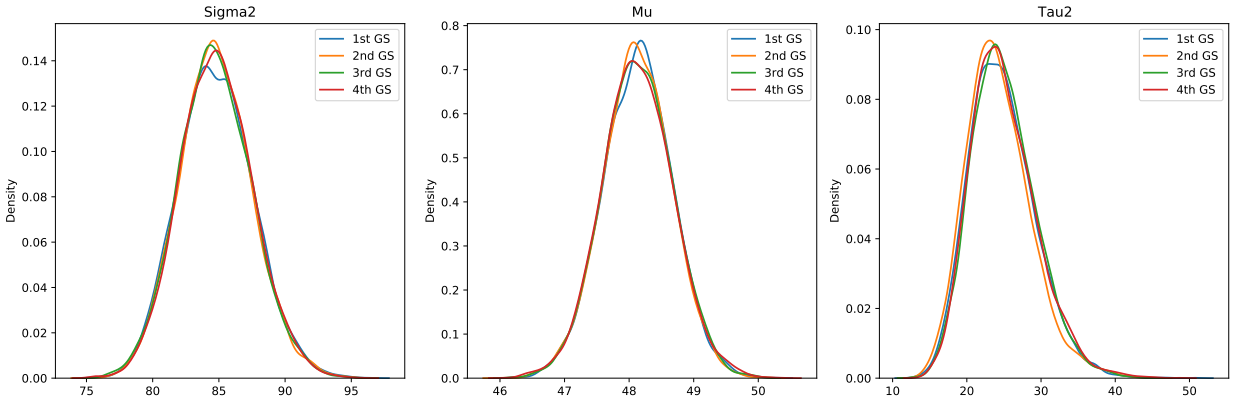

Figure 5: Estimated Posterior Densities of σ^2, μ, τ^2

For the analysis, we have considered the first Gibbs Sampler run with $\sigma_0^2 = 100$, $\tau_0^2 = 100$, $\nu_0 = 1$, $\eta_0 = 1$, $\mu_0 = 50$, and $\gamma_0^2 = 25$.

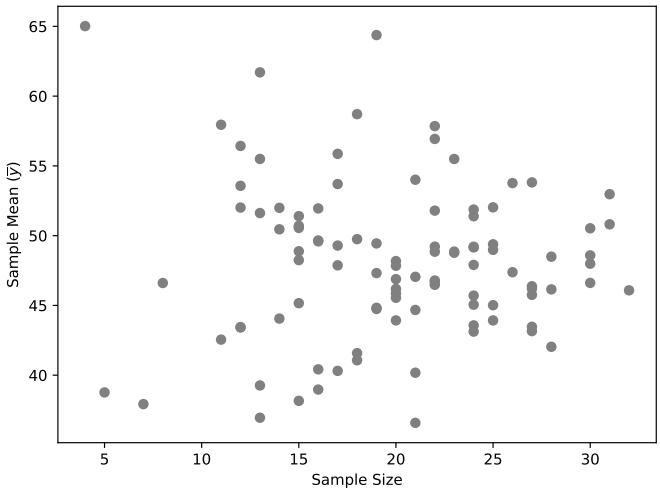

Figure 6: Relationship between sample mean and sample size

The relationship between sample mean and sample size is depicted in Figure 6. Schools with fewer observations have more extreme sample averages (very high or very low), which is common in hierarchical datasets. Hierarchical models address this issue by allowing information to be shared between groups. As a matter of fact, the expected value of θ_j is a weighted average of the sample mean $\bar{y}_j$ and μ, and it is pushed towards μ by a factor dependent on the sample size. In particular, Figure 7 shows how the schools

with the smallest sample size get shrunk the most as they borrow more information from the rest of the population.

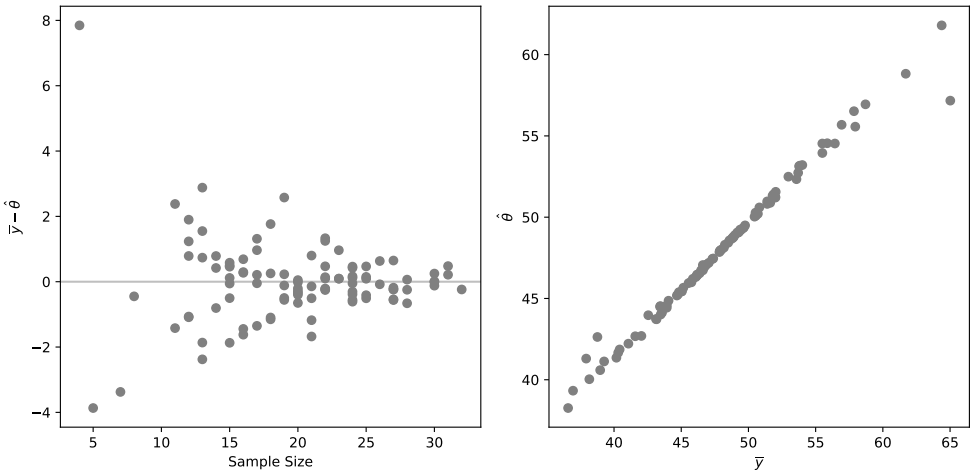

Figure 7: Shrinkage as a function of the sample size

The posterior expected value of the within-school variance σ^2 is 84.789, with a 95% quantile-based posterior confidence interval equal to $(79.551, 90.539)$. In comparison, the posterior expected value of the between-school variability τ^2 is much lower at 24.801, with a 95% quantile-based posterior confidence interval equal to $(17.389, 34.414)$. The mean of school-specific means μ has a posterior point estimate equal to 48.138, with a 95% quantile-based posterior confidence interval equal to $(47.070, 49.180)$.

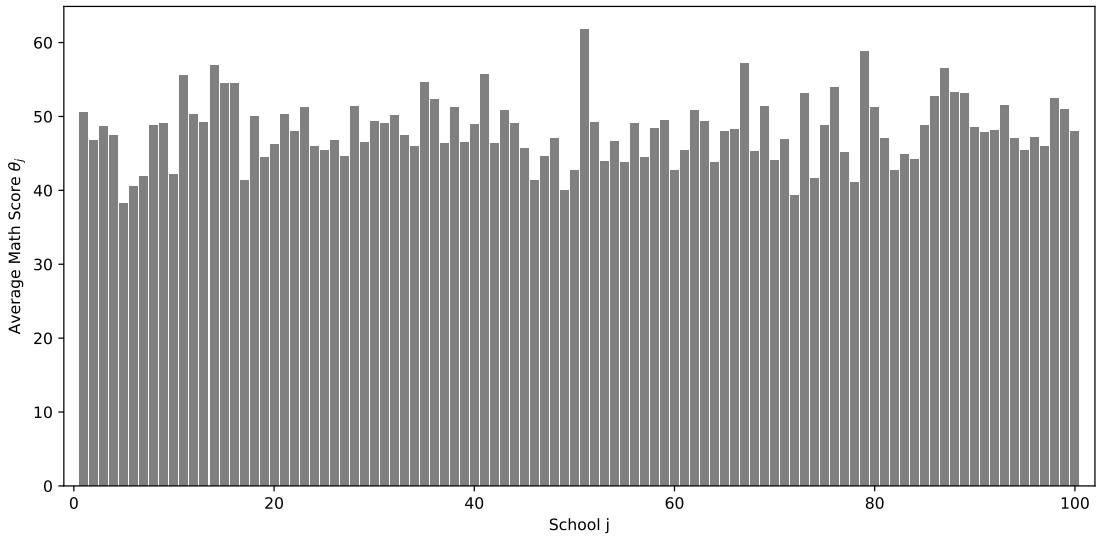

Figure 8: Estimated school-specific posterior averages

Let's assume every student in each school took the math exam. To rank the schools according to their performances, we can use the school-specific posterior expectations.

Figure 8 shows the posterior estimate of the school-specific means (θs). School 5 performs the worst, with an estimated average math score of 38.260, while school 51 performs the best, with an estimated average math score of 61.780.

3 Hierarchical model with covariates

As in the previous section, a hierarchical normal model is used

$$Y_{i,j}|\ \boldsymbol{\beta}_j, \sigma^2 \sim N(y|\ \boldsymbol{\beta}_j^{\mathrm{T}}\mathbf{x}_{i,j}, \sigma^2), \quad i = 1, ..., n_j, j = 1, ..., m$$
$$\boldsymbol{\beta}_j|\ \boldsymbol{\theta}, \Sigma \sim N(\boldsymbol{\beta}|\ \boldsymbol{\theta}, \boldsymbol{\Sigma})$$
$$(\boldsymbol{\theta}, \Sigma, \sigma^2) \sim p_0$$

The fixed but unknown parameters of the system are the within-school sampling variability σ^2, the school-specific regression parameters $\boldsymbol{\beta}_j$, and the mean and variance of the population of school-specific regression parameters $(\boldsymbol{\theta}, \boldsymbol{\Sigma})$. The within-school sampling variability σ^2 is again assumed to be the same across schools. A Gibbs sampler is used to approximate the joint posterior distribution $p(\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_m, \boldsymbol{\theta}, \Sigma, \sigma^2|\boldsymbol{X}_1, ..., \boldsymbol{X}_m, \boldsymbol{y}_1, ..., \boldsymbol{y}_m)$ of these unknown parameters, given the prior distributions for $\boldsymbol{\theta}, \Sigma$, σ^2, and the observed $\boldsymbol{Y}_1 = \boldsymbol{y}_1, ..., \boldsymbol{Y}_m = \boldsymbol{y}_m$. The following semi-conjugate prior distributions for $\boldsymbol{\theta}, \Sigma$, and σ^2 were used:

$$\boldsymbol{\theta} \sim Multivariate\ Normal(\boldsymbol{\mu}_0, \Lambda_0)$$

$$\Sigma \sim Inverse\text{-}Wishart(\eta_0, \boldsymbol{S}_0^{-1})$$

$$\sigma^2 \sim Inverse\text{-}Gamma\left(\frac{\nu_0}{2}, \nu_0\frac{\sigma_0^2}{2}\right)$$

To simplify the computation of the full conditionals, we consider the following factorization of the target posterior distribution:

$$p(\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_m, \boldsymbol{\theta}, \Sigma, \sigma^2|\boldsymbol{X}_1, ..., \boldsymbol{X}_m, \boldsymbol{y}_1, ..., \boldsymbol{y}_m)$$

$$\propto p(\boldsymbol{\theta}, \Sigma, \sigma^2) \times p(\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_m|\ \boldsymbol{\theta}, \Sigma, \sigma^2) \times p(\boldsymbol{y}_1, ..., \boldsymbol{y}_m|\ \boldsymbol{X}_1\boldsymbol{\beta}_1, ..., \boldsymbol{X}_m\boldsymbol{\beta}_m, \boldsymbol{\theta}, \Sigma, \sigma^2)$$
$$= p(\boldsymbol{\theta})p(\Sigma)p(\sigma^2)\left\{\prod_{j=1}^{m} p(\boldsymbol{\beta}_j|\ \boldsymbol{\theta}, \Sigma)\right\}\left\{\prod_{j=1}^{m} p(\boldsymbol{y}_j|\ \mathbf{X}_j\boldsymbol{\beta}_j, \sigma^2)\right\}$$

3.1 Full conditionals of $\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_m$

Conditional on $\boldsymbol{\theta}, \Sigma, \sigma^2$, the regression coefficients $\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_m$ are independent. Considering a single $\boldsymbol{\beta}_j$, the model is like a one-group regression problem where the prior mean and variance of $\boldsymbol{\beta}_j$ are $\boldsymbol{\theta}$ and Σ. From the factorization of the posterior distribution,

the full conditional of $\boldsymbol{\beta}_j$ should be such that

$$p(\boldsymbol{\beta}_j|\ \boldsymbol{\theta}, \Sigma, \sigma^2, \boldsymbol{X}_j, \boldsymbol{y}_j) \propto p(\boldsymbol{\beta}_j|\ \boldsymbol{\theta}, \Sigma) \prod_{j=1}^{m} p(\boldsymbol{y}_j|\ \mathbf{X}_j\boldsymbol{\beta}_j, \sigma^2)$$

$$\propto exp\left\{-\frac{1}{2}(\boldsymbol{\beta}_j - \boldsymbol{\theta})^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{\beta}_j - \boldsymbol{\theta})\right\} exp\left\{-\frac{1}{2}(\boldsymbol{y}_j - \boldsymbol{X}_j\boldsymbol{\beta}_j)^{\mathrm{T}}(\sigma^2)^{-1}(\boldsymbol{y}_j - \boldsymbol{X}_j\boldsymbol{\beta}_j)\right\}$$

$$= exp\left\{-\frac{1}{2}(\boldsymbol{\beta}_j - \boldsymbol{\theta})^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{\beta}_j - \boldsymbol{\theta}) - \frac{1}{2}(\boldsymbol{y}_j - \boldsymbol{X}_j\boldsymbol{\beta}_j)^{\mathrm{T}}(\sigma^2)^{-1}(\boldsymbol{y}_j - \boldsymbol{X}_j\boldsymbol{\beta}_j)\right\}$$

$$= exp\left\{-\frac{1}{2}(\boldsymbol{\beta}_j^{\mathrm{T}}\Sigma^{-1}\boldsymbol{\beta}_j - 2\boldsymbol{\beta}_j^{\mathrm{T}}\Sigma^{-1}\boldsymbol{\theta} + \boldsymbol{\theta}^{\mathrm{T}}\Sigma^{-1}\boldsymbol{\theta}) - \frac{1}{2}\left(\frac{\boldsymbol{y}_j^{\mathrm{T}}\boldsymbol{y}_j}{\sigma^2} - 2\frac{\boldsymbol{\beta}_j^{\mathrm{T}}\boldsymbol{X}_j^{\mathrm{T}}\boldsymbol{y}_j}{\sigma^2} + \frac{\boldsymbol{\beta}_j^{\mathrm{T}}\boldsymbol{X}_j^{\mathrm{T}}\boldsymbol{X}_j\boldsymbol{\beta}_j)}{\sigma^2}\right)\right\}$$

$$\propto exp\left\{-\frac{1}{2}(\boldsymbol{\beta}_j^{\mathrm{T}}\Sigma^{-1}\boldsymbol{\beta}_j - 2\boldsymbol{\beta}_j^{\mathrm{T}}\Sigma^{-1}\boldsymbol{\theta}) - \frac{1}{2}\left(-2\frac{\boldsymbol{\beta}_j^{\mathrm{T}}\boldsymbol{X}_j^{\mathrm{T}}\boldsymbol{y}_j}{\sigma^2} + \frac{\boldsymbol{\beta}_j^{\mathrm{T}}\boldsymbol{X}_j^{\mathrm{T}}\boldsymbol{X}_j\boldsymbol{\beta}_j)}{\sigma^2}\right)\right\}$$

$$= exp\left\{\boldsymbol{\beta}_j^{\mathrm{T}}\left(\Sigma^{-1}\boldsymbol{\theta} + \frac{\boldsymbol{X}_j^{\mathrm{T}}\boldsymbol{y}_j}{\sigma^2}\right) - \frac{1}{2}\boldsymbol{\beta}_j^{\mathrm{T}}\left(\Sigma^{-1} + \frac{\boldsymbol{X}_j^{\mathrm{T}}\boldsymbol{X}_j}{\sigma^2}\right)\boldsymbol{\beta}_j\right\}$$

Thus, $\{\boldsymbol{\beta}_j|\ \boldsymbol{\theta}, \Sigma, \sigma^2, \boldsymbol{X}_j, \boldsymbol{y}_j\}$ has a multivariate normal density with

$$Var[\boldsymbol{\beta}_j|\ \boldsymbol{\theta}, \Sigma, \sigma^2, \boldsymbol{X}_j, \boldsymbol{y}_j] = \left(\Sigma^{-1} + \frac{\boldsymbol{X}_j^{\mathrm{T}}\boldsymbol{X}_j}{\sigma^2}\right)^{-1}$$

$$E[\boldsymbol{\beta}_j|\ \boldsymbol{\theta}, \Sigma, \sigma^2, \boldsymbol{X}_j, \boldsymbol{y}_j] = \left(\Sigma^{-1} + \frac{\boldsymbol{X}_j^{\mathrm{T}}\boldsymbol{X}_j}{\sigma^2}\right)^{-1}\left(\Sigma^{-1}\boldsymbol{\theta} + \frac{\boldsymbol{X}_j^{\mathrm{T}}\boldsymbol{y}_j}{\sigma^2}\right)$$

3.2 Full conditional of θ

Considering the factorization of the posterior distribution as a function of $\boldsymbol{\theta}$:

$$p(\boldsymbol{\theta}|\ \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_m, \Sigma, \sigma^2, \boldsymbol{X}_1, ..., \boldsymbol{X}_m, \boldsymbol{y}_1, ..., \boldsymbol{y}_m) \propto p(\boldsymbol{\theta}) \prod_{j=1}^{m} p(\boldsymbol{\beta}_j|\ \boldsymbol{\theta}, \Sigma)$$

$$\propto exp\left\{-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\mu}_0)^{\mathrm{T}}\Lambda_0^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu}_0)\right\} exp\left\{-\frac{1}{2}\sum_{j=1}^{m}(\boldsymbol{\beta}_j - \boldsymbol{\theta})^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{\beta}_j - \boldsymbol{\theta})\right\}$$

$$= exp\left\{-\frac{1}{2}(\boldsymbol{\theta}^{\mathrm{T}}\Lambda_0^{-1}\boldsymbol{\theta} - 2\boldsymbol{\theta}^{\mathrm{T}}\Lambda_0^{-1}\boldsymbol{\mu}_0 + \boldsymbol{\mu_0}^{\mathrm{T}}\Lambda_0^{-1}\boldsymbol{\mu_0}) - \frac{1}{2}\sum_{j=1}^{m}(\boldsymbol{\beta}_j^{\mathrm{T}}\Sigma^{-1}\boldsymbol{\beta}_j - 2\boldsymbol{\theta}^{\mathrm{T}}\Sigma^{-1}\boldsymbol{\beta}_j + \boldsymbol{\theta}^{\mathrm{T}}\Sigma^{-1}\boldsymbol{\theta})\right\}$$

$$\propto exp\left\{-\frac{1}{2}(\boldsymbol{\theta}^{\mathrm{T}}\Lambda_0^{-1}\boldsymbol{\theta} - 2\boldsymbol{\theta}^{\mathrm{T}}\Lambda_0^{-1}\boldsymbol{\mu}_0) - \frac{1}{2}\sum_{j=1}^{m}(-2\boldsymbol{\theta}^{\mathrm{T}}\Sigma^{-1}\boldsymbol{\beta}_j + \boldsymbol{\theta}^{\mathrm{T}}\Sigma^{-1}\boldsymbol{\theta})\right\}$$

$$= exp\left\{-\frac{1}{2}(\boldsymbol{\theta}^{\mathrm{T}}\Lambda_0^{-1}\boldsymbol{\theta}) + \boldsymbol{\theta}^{\mathrm{T}}\Lambda_0^{-1}\boldsymbol{\mu}_0 - \frac{1}{2}(\boldsymbol{\theta}^{\mathrm{T}}m\Sigma^{-1}\boldsymbol{\theta}) + \boldsymbol{\theta}^{\mathrm{T}}m\Sigma^{-1}\bar{\boldsymbol{\beta}}\right\}$$

$$= exp\left\{-\frac{1}{2}\left(\boldsymbol{\theta}^{\mathrm{T}}(\Lambda_0^{-1} + m\Sigma^{-1})\boldsymbol{\theta}\right) + \boldsymbol{\theta}^{\mathrm{T}}(\Lambda_0^{-1}\boldsymbol{\mu}_0 + m\Sigma^{-1}\bar{\boldsymbol{\beta}})\right\}$$

It follows that

$$\{\boldsymbol{\theta}|\ \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_m, \Sigma\} \sim Multivariate\ Normal(\boldsymbol{\mu}_m, \Lambda_m)$$

where

$$\Lambda_m = (\Lambda_0^{-1} + m\Sigma^{-1})^{-1}$$

$$\boldsymbol{\mu}_m = (\Lambda_0^{-1} + m\Sigma^{-1})^{-1}(\Lambda_0^{-1}\boldsymbol{\mu}_0 + m\Sigma^{-1}\bar{\boldsymbol{\beta}})$$

3.3 Full conditional of Σ

From the factorization of the posterior distribution, the full conditional of Σ should be such that

$$p(\Sigma|\ \boldsymbol{\theta}, \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_m, \sigma^2, \boldsymbol{X}_1, ..., \boldsymbol{X}_m, \boldsymbol{y}_1, ..., \boldsymbol{y}_m) \propto p(\Sigma)\prod_{j=1}^{m} p(\boldsymbol{\beta}_j|\ \boldsymbol{\theta}, \Sigma)$$

$$\propto |\Sigma|^{-(\eta_0+p+1)/2} exp\left\{-\frac{1}{2}tr(\boldsymbol{S}_0\Sigma^{-1})\right\} |\Sigma|^{-m/2} exp\left\{-\frac{1}{2}\sum_{j=1}^{m}(\boldsymbol{\beta}_j - \boldsymbol{\theta})^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{\beta}_j - \boldsymbol{\theta})\right\}$$

Exploiting matrix algebra,

$$\sum_{j=1}^{m}(\boldsymbol{\beta}_j - \boldsymbol{\theta})^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{\beta}_j - \boldsymbol{\theta}) = tr(\boldsymbol{S}_\theta\Sigma^{-1})$$

where $\boldsymbol{S}_\theta = \sum_{j=1}^{m}(\boldsymbol{\beta}_j - \boldsymbol{\theta})(\boldsymbol{\beta}_j - \boldsymbol{\theta})^{\mathrm{T}}$. Therefore

$$|\Sigma|^{-(\eta_0+p+1)/2} exp\left\{-\frac{1}{2}tr(\boldsymbol{S}_0\Sigma^{-1})\right\} |\Sigma|^{-m/2} exp\left\{-\frac{1}{2}\sum_{j=1}^{m}(\boldsymbol{\beta}_j - \boldsymbol{\theta})^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{\beta}_j - \boldsymbol{\theta})\right\}$$

$$= |\Sigma|^{-(\eta_0+m+p+1)/2} exp\left\{-\frac{1}{2}tr(\boldsymbol{S}_0\Sigma^{-1})\right\} exp\left\{-\frac{1}{2}tr(\boldsymbol{S}_\theta\Sigma^{-1})\right\}$$

$$= |\Sigma|^{-(\eta_0+m+p+1)/2} exp\left\{-\frac{1}{2}tr([\boldsymbol{S}_0 + \boldsymbol{S}_\theta]\Sigma^{-1})\right\}$$

which is the kernel of an Inverse-Wishart distribution

$$\{\Sigma|\ \boldsymbol{\theta}, \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_m\} \sim Inverse\text{-}Wishart\left(\eta_0 + m, [\boldsymbol{S}_0 + \boldsymbol{S}_\theta]^{-1}\right)$$

3.4 Full conditional of σ^2

The full conditional distribution of σ^2 should be proportional to

$$p(\sigma^2|\ \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_m, \boldsymbol{X}_1, ..., \boldsymbol{X}_m, \boldsymbol{y}_1, ..., \boldsymbol{y}_m) \propto p(\sigma^2) \prod_{j=1}^{m} p(\boldsymbol{y}_j|\ \mathbf{X}_j\boldsymbol{\beta}_j, \sigma^2)$$

$$p(\sigma^2|\ \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_m, \boldsymbol{X}_1, ..., \boldsymbol{X}_m, \boldsymbol{y}_1, ..., \boldsymbol{y}_m) \propto p(\sigma^2) \prod_{j=1}^{m} \prod_{i=1}^{n_j} p(y_{i,j}|\ \boldsymbol{\beta}_j^{\mathrm{T}}\mathbf{x}_{i,j}, \sigma^2)$$

$$\propto (\sigma^2)^{-\frac{\nu_0}{2}-1} exp\left\{-\frac{\nu_0\sigma_0^2}{2\sigma^2}\right\} (\sigma^2)^{-\frac{1}{2}\sum_{j=1}^{m} n_j} exp\left\{-\frac{1}{2\sigma^2}\sum_{j=1}^{m}\sum_{i=1}^{n_j}(y_{i,j} - \ \boldsymbol{\beta}_j^{\mathrm{T}}\mathbf{x}_{i,j})^2\right\}$$

Defining the sum of the squared residuals, $SSR(\boldsymbol{\beta_j}) = \sum_{j=1}^{m}\sum_{i=1}^{n_j}(y_{i,j} - \ \boldsymbol{\beta}_j^{\mathrm{T}}\mathbf{x}_{i,j})^2$, we obtain

$$(\sigma^2)^{-\frac{\nu_0}{2}-\frac{1}{2}\sum_{j=1}^{m} n_j-1} exp\left\{-\frac{\nu_0\sigma_0^2}{2\sigma^2}\right\} exp\left\{-\frac{1}{2\sigma^2}SSR(\boldsymbol{\beta}_j)\right\}$$

$$(\sigma^2)^{-\frac{1}{2}(\nu_0+\sum_{j=1}^{m} n_j)-1} exp\left\{-\frac{1}{\sigma^2}\left(\frac{\nu_0\sigma_0^2 + SSR(\boldsymbol{\beta}_j)}{2}\right)\right\}$$

which is the kernel of an Inverse-Gamma distribution with parameters

$$\alpha = \frac{\nu_0 + \sum_{j=1}^{m} n_j}{2}$$

$$\beta = \frac{\nu_0\sigma_0^2 + SSR(\boldsymbol{\beta}_j)}{2}$$

Thus

$$\{\sigma^2|\ \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_m, \boldsymbol{X}_1, ..., \boldsymbol{X}_m, \boldsymbol{y}_1, ..., \boldsymbol{y}_m\}$$

$$\sim Inverse\text{-}Gamma\left(\frac{\nu_0 + \sum_{j=1}^{m} n_j}{2}, \frac{\nu_0\sigma_0^2 + SSR(\boldsymbol{\beta}_j)}{2}\right)$$

3.5 Gibbs Sampler

Given the current state of the parameters $\left\{\boldsymbol{\beta}_1^{(t)}, \boldsymbol{\beta}_2^{(t)}, ..., \boldsymbol{\beta}_m^{(t)}, \boldsymbol{\theta}^{(t)}, \Sigma^{(t)}, \sigma^{2(t)}\right\}$, at each iteration the new state is generated by:

- sampling $\boldsymbol{\beta}_j^{(t+1)} \sim p(\boldsymbol{\beta}_j|\ \boldsymbol{\theta}^{(t)}, \Sigma^{(t)}, \sigma^{2(t)}, \boldsymbol{X}_j, \boldsymbol{y}_j)$, for each $j \in \{1, ..., m\}$
- sampling $\sigma^{2(t+1)} \sim p(\sigma^2|\ \boldsymbol{\beta}_1^{(t+1)}, \boldsymbol{\beta}_2^{(t+1)}, ..., \boldsymbol{\beta}_m^{(t+1)}, \boldsymbol{X}_1, ..., \boldsymbol{X}_m, \boldsymbol{y}_1, ..., \boldsymbol{y}_m)$
- sampling $\boldsymbol{\theta}^{(t+1)} \sim p(\boldsymbol{\theta}|\ \boldsymbol{\beta}_1^{(t+1)}, \boldsymbol{\beta}_2^{(t+1)}, ..., \boldsymbol{\beta}_m^{(t+1)}, \Sigma^{(t)})$

- sampling $\Sigma^{(t+1)} \sim p(\Sigma|\ \boldsymbol{\beta}_1^{(t+1)}, \boldsymbol{\beta}_2^{(t+1)}, ..., \boldsymbol{\beta}_m^{(t+1)}, \boldsymbol{\theta}^{(t+1)})$

where each parameter gets updated conditional on the most current value of the other parameters. The prior parameters for which a value should be chosen are $(\boldsymbol{\mu}_0, \Lambda_0)$ for $p(\boldsymbol{\theta})$, $(\eta_0, \boldsymbol{S}_0^{-1})$ for $p(\Sigma)$, and (ν_0, σ_0^2) for $p(\sigma^2)$.

A unit-information-prior-inspired approach is used to select the prior values for $\boldsymbol{\mu_0}$ and σ_0^2. The prior expectation of $\boldsymbol{\theta}$, $\boldsymbol{\mu}_0$, is assumed to be equal to the average of the OLS coefficients estimated for each school. The prior distribution of σ^2 is centered on the average of the within-school variances estimated by OLS, and it is reasonably diffuse by taking $\nu_0 = 1$. It is common in literature to impose $\Lambda_0 = g\sigma_0^2(X^{\mathrm{T}}X)^{-1}$, for some positive value g that reflects the amount of information in the data. The use of such prior value ensures the invariance to changes in the scale of the regressors. In the first specification, a $g = 300$ is chosen to naturally induce a diffuse prior. The prior matrix $\boldsymbol{S}_0$ is assumed to be equal to Λ_0, but the degrees of freedom η_0 are set equal to $p + 2 = 5$ to ensure that the prior distribution of Σ is just vaguely concentrated around its mean. A Gibbs sampler with 10,000 burn-in iterations and 10,000 actual iterations is run. The initial values of $\boldsymbol{\theta}$, Σ, and σ^2 are set by randomly sampling from their prior distributions. The updated parameters are saved any 10 iterations, generating a sequence of 1000 values for each parameter.

Figure 9 depicts the trace plots, autocorrelation plots and boxplots for $\boldsymbol{\theta}$, whereas Figure 10 and Figure 11 depict the trace plots, autocorrelation plots and boxplots of $\boldsymbol{\beta}$ for the school with the most observations (school 37) and the school with the fewest observations (school 67). Each boxplot represents 200 MCMC samples.

According to the trace plots and boxplots, all parameters appear to have converged. The sample distribution in any of the boxplots is the same as it is in the others, suggesting that the chain has achieved stationarity. Furthermore, each sequence seems to be characterized by a fairly low autocorrelation, indicating that the Gibbs Sampler is moving quickly enough in the parameter space.

The notebook also shows the convergence of the components of the var-cov matrix Σ and of the within-school variance σ^2, but in this report we mainly focus on analysing $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$.

Instead of the more general g-prior (where g could be any positive number), a unit-information prior can be also used for the matrix Λ_0.

A unit-information prior is one that contains the same amount of information as a single observation. $(X^{\mathrm{T}}X)/\sigma^2$ can be thought of as the precision of an OLS estimate obtained considering all n observations. Because this matrix represents the amount of

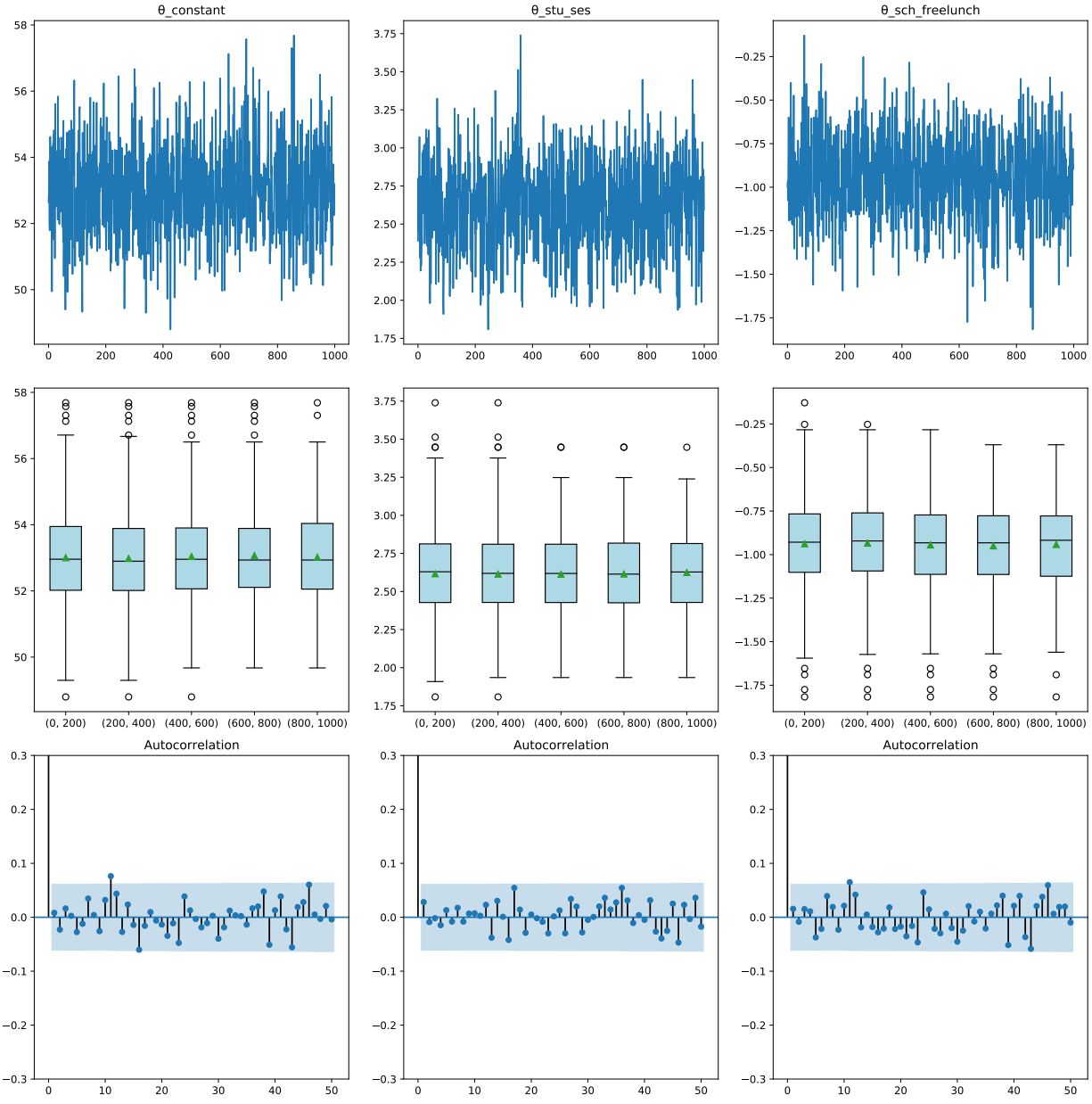

Figure 9: Trace Plots, Box Plots, and Autocorrelation Plots of $\boldsymbol{\theta}$

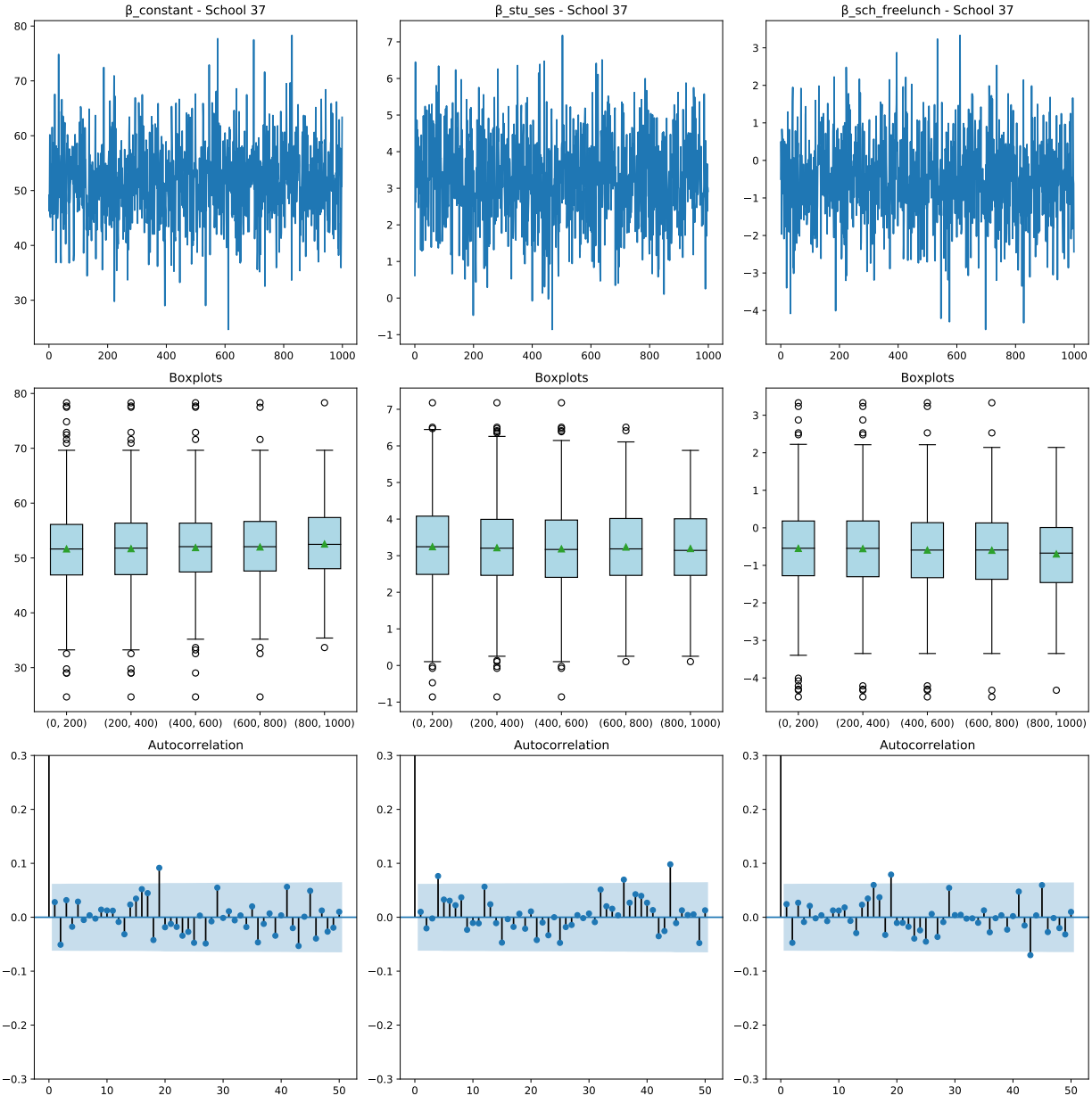

Figure 10: Trace Plots, Box Plots, and Autocorrelation Plots of $\boldsymbol{\beta}_{37}$

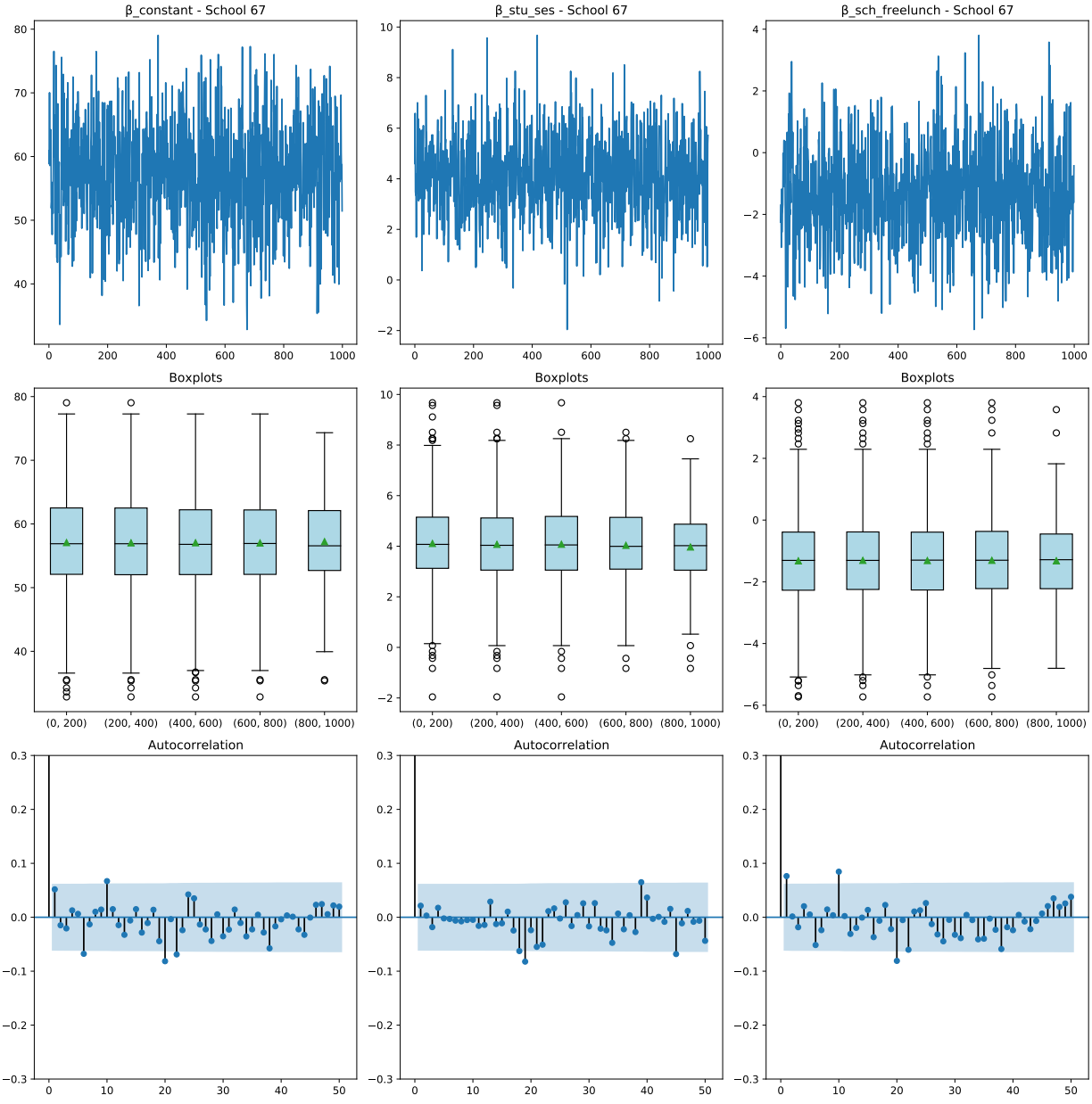

Figure 11: Trace Plots, Box Plots, and Autocorrelation Plots of $\boldsymbol{\beta}_{67}$

information in n observations, $(X^{\mathrm{T}}X)/n\sigma^2$ represents the amount of information in a single observation. Using a unit-information prior for Λ_0 implies considering $\Lambda_0^{-1} = (X^{\mathrm{T}}X)/n\sigma^2$, which means $\Lambda_0 = n\sigma^2(X^{\mathrm{T}}X)^{-1}$.

Thus, another Gibbs Sampler is run considering this new Λ_0, with n = 1993 (total number of observations). This change results in an even more diffuse prior for $\boldsymbol{\theta}$. All other prior parameter values are preserved, with the exception of S_0, which is set equal to the new Λ_0. Figures 12, 13, and 14 show the trace plots, autocorrelation plots and boxplots for $\boldsymbol{\theta}$, $\boldsymbol{\beta}_{37}$, and $\boldsymbol{\beta}_{67}$, respectively.

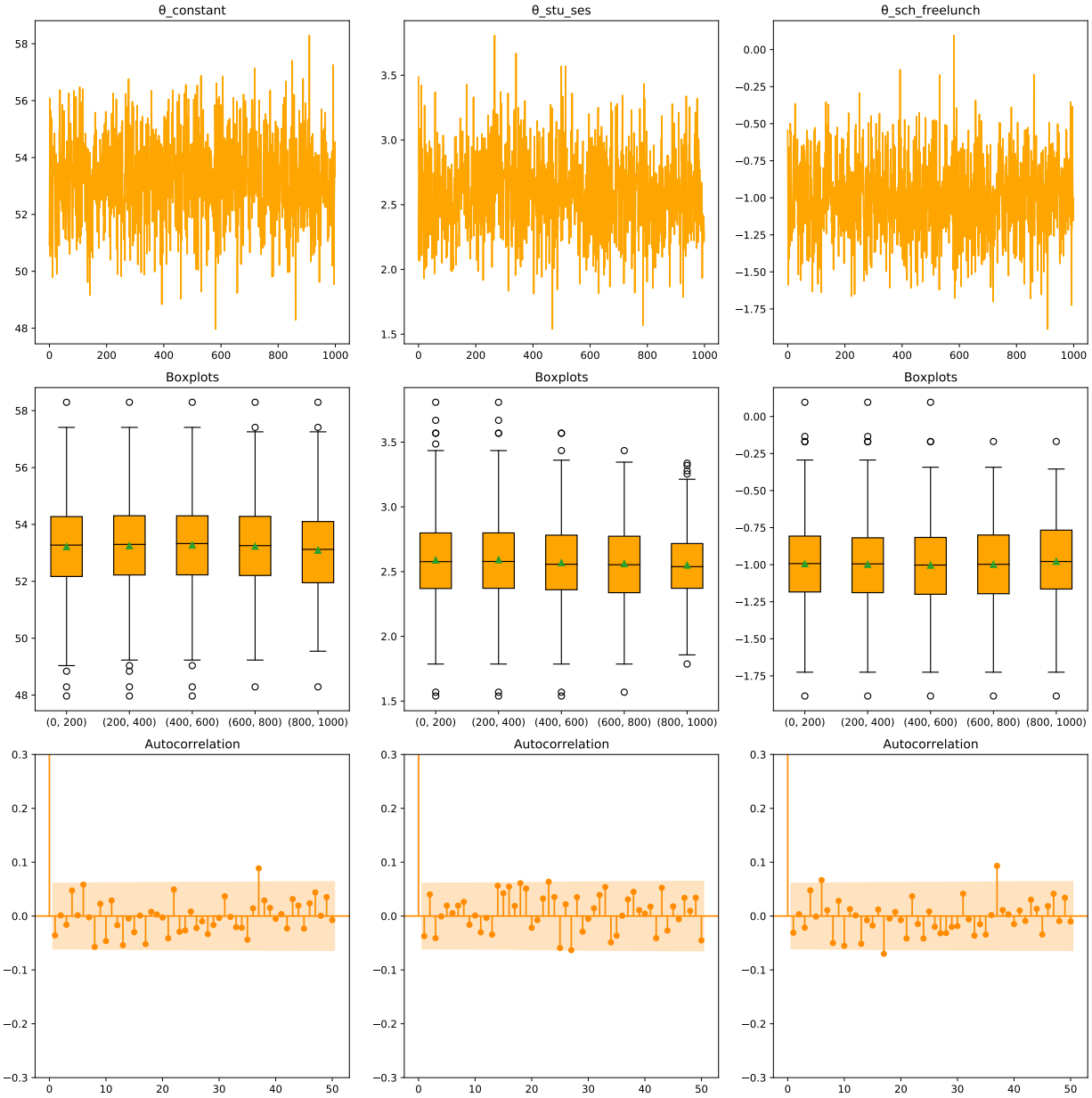

Figure 12: Trace Plots, Box Plots, and Autocorrelation Plots of $\boldsymbol{\theta}$

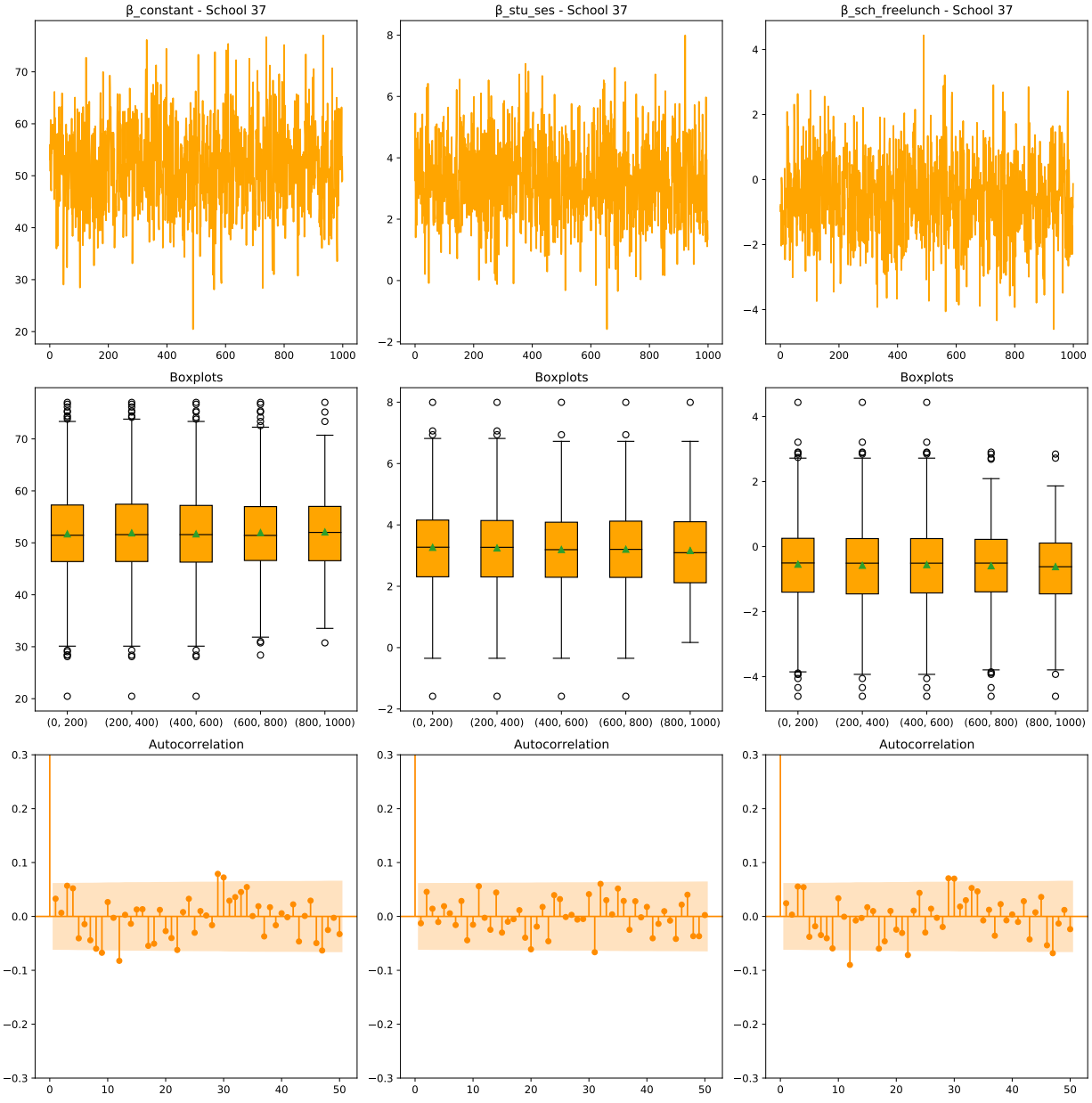

Figure 13: Trace Plots, Box Plots, and Autocorrelation Plots of $\boldsymbol{\beta}_{37}$

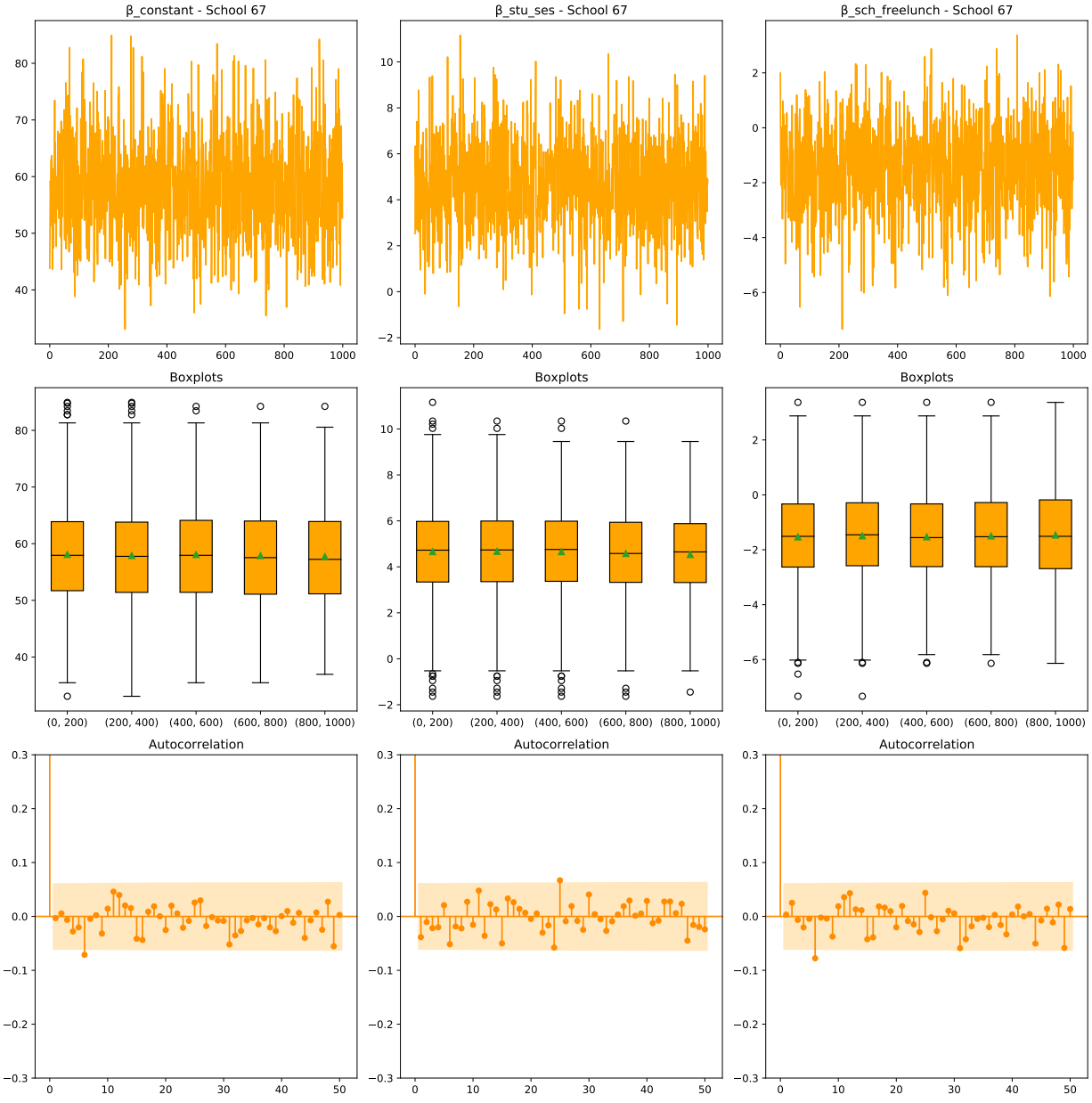

Figure 14: Trace Plots, Box Plots, and Autocorrelation Plots of $\boldsymbol{\beta}_{67}$

The Gibbs Sampling is then repeated with a different value for $\boldsymbol{\mu}_0$. Rather than using the mean of the estimated OLS coefficients for each school, we have used the vector of coefficients obtained by regressing schools' average math-scores on average SES and number of free lunches. All other parameters of the first Gibbs Sampler have been kept. Figure 15 depicts the trace plots, autocorrelation plots and boxplots for $\boldsymbol{\theta}$, whereas Figure 16 and Figure 17 depict the trace plots, autocorrelation plots and boxplots of $\boldsymbol{\beta}$ for schools 37 and 67. It is worth noting that with this Gibbs Sampler some components of Σ show some relevant autocorrelation until lag-10, implying that this Gibbs Sampler is moving slower in the parameter space of Σ than the previous ones.

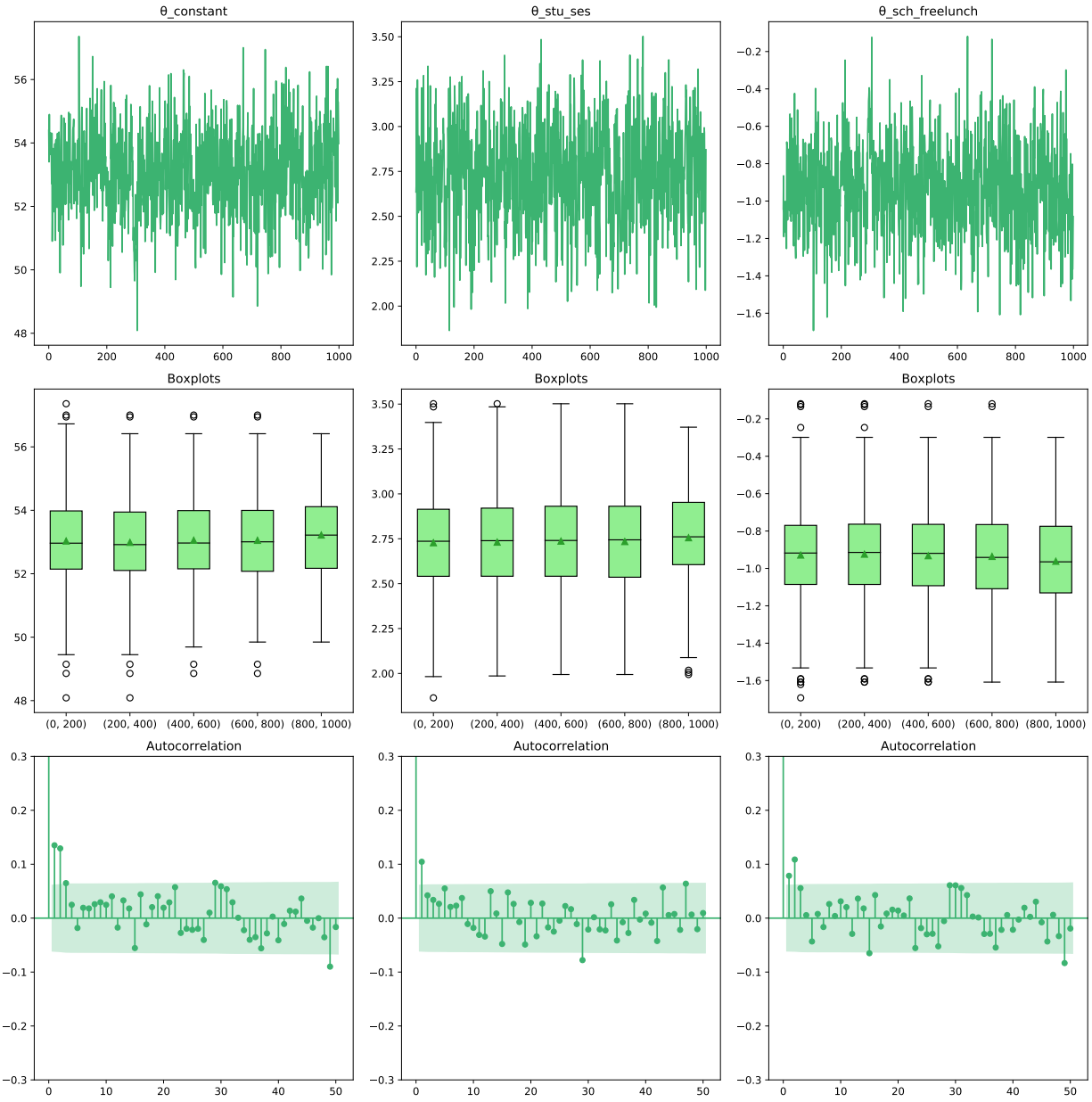

Figure 15: Trace Plots, Box Plots, and Autocorrelation Plots of $\boldsymbol{\theta}$

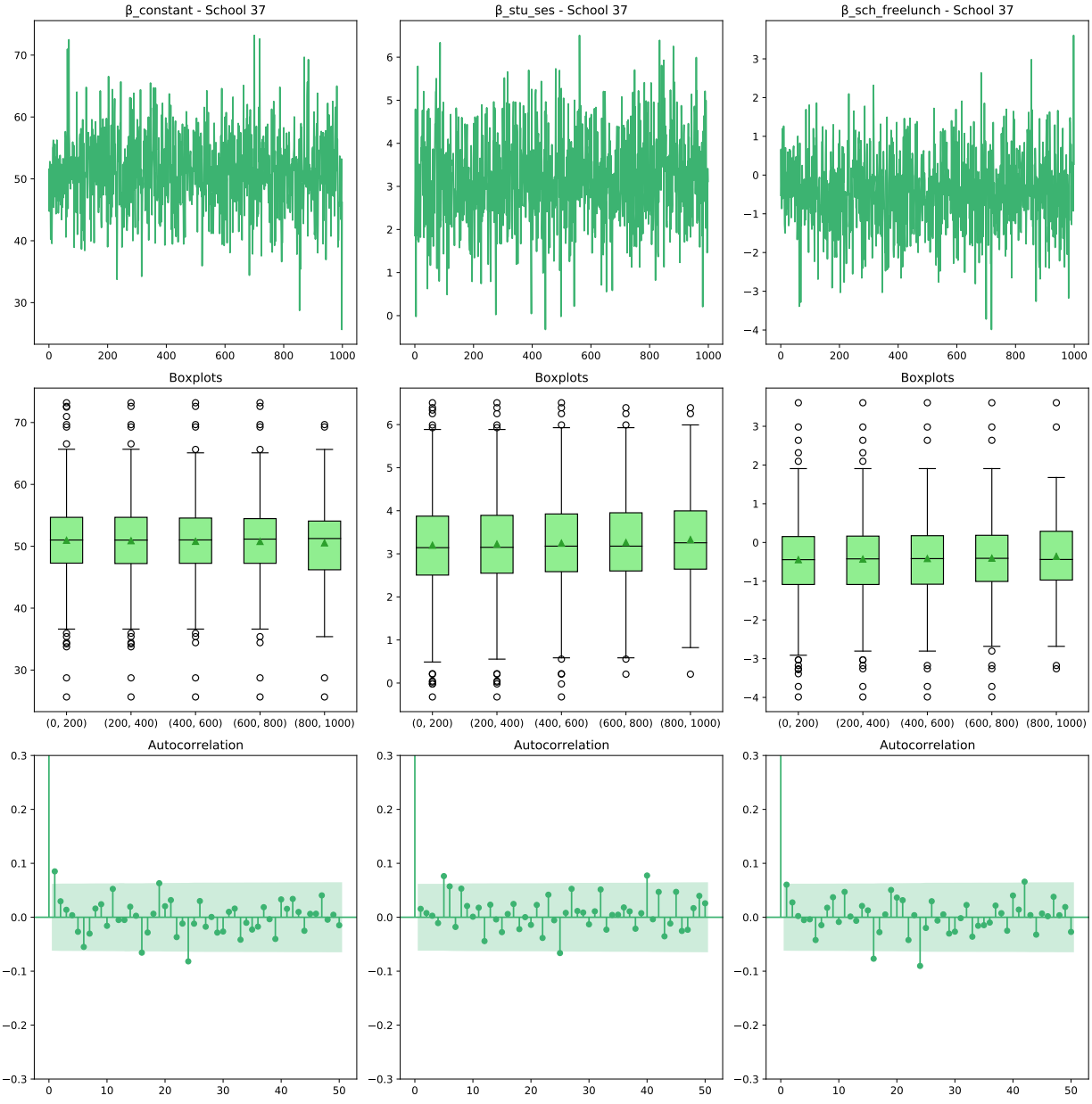

Figure 16: Trace Plots, Box Plots, and Autocorrelation Plots of $\boldsymbol{\beta}_{37}$

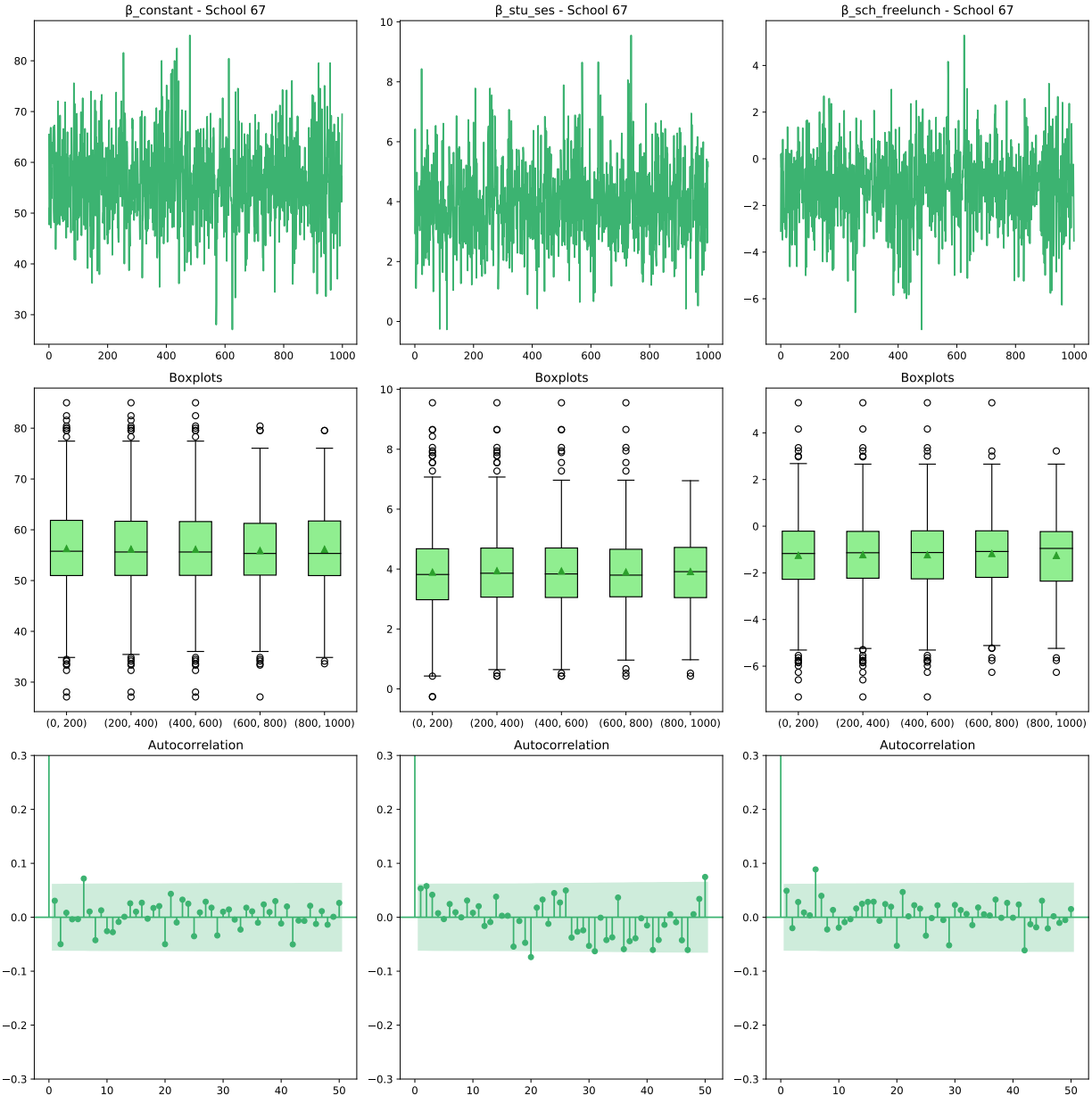

Figure 17: Trace Plots, Box Plots, and Autocorrelation Plots of $\boldsymbol{\beta}_{67}$

Another Gibbs Sampler is run with a different starting point, but with the same parameter values as the first Gibbs Sampler. Prior expectations were used as initial values for $\boldsymbol{\theta}$, Σ, σ^2. In particular, $\boldsymbol{\theta}$ was initially set equal to $\boldsymbol{\mu}_0$, Σ was initially set equal to S_0, σ^2 was initially set equal to σ_0^2.

Notice that σ_0^2 is not the actual expectation of σ^2, as the expectation of an Inverse-Gamma distribution is not defined when the shape parameter is less than 1. Indeed, setting $\nu_0 = 1$, $\nu_0/2$ is < 1. Figures 18, 19, and 20 show the trace plots, autocorrelation plots and boxplots for $\boldsymbol{\theta}$, $\boldsymbol{\beta}_{37}$, and $\boldsymbol{\beta}_{67}$, respectively.

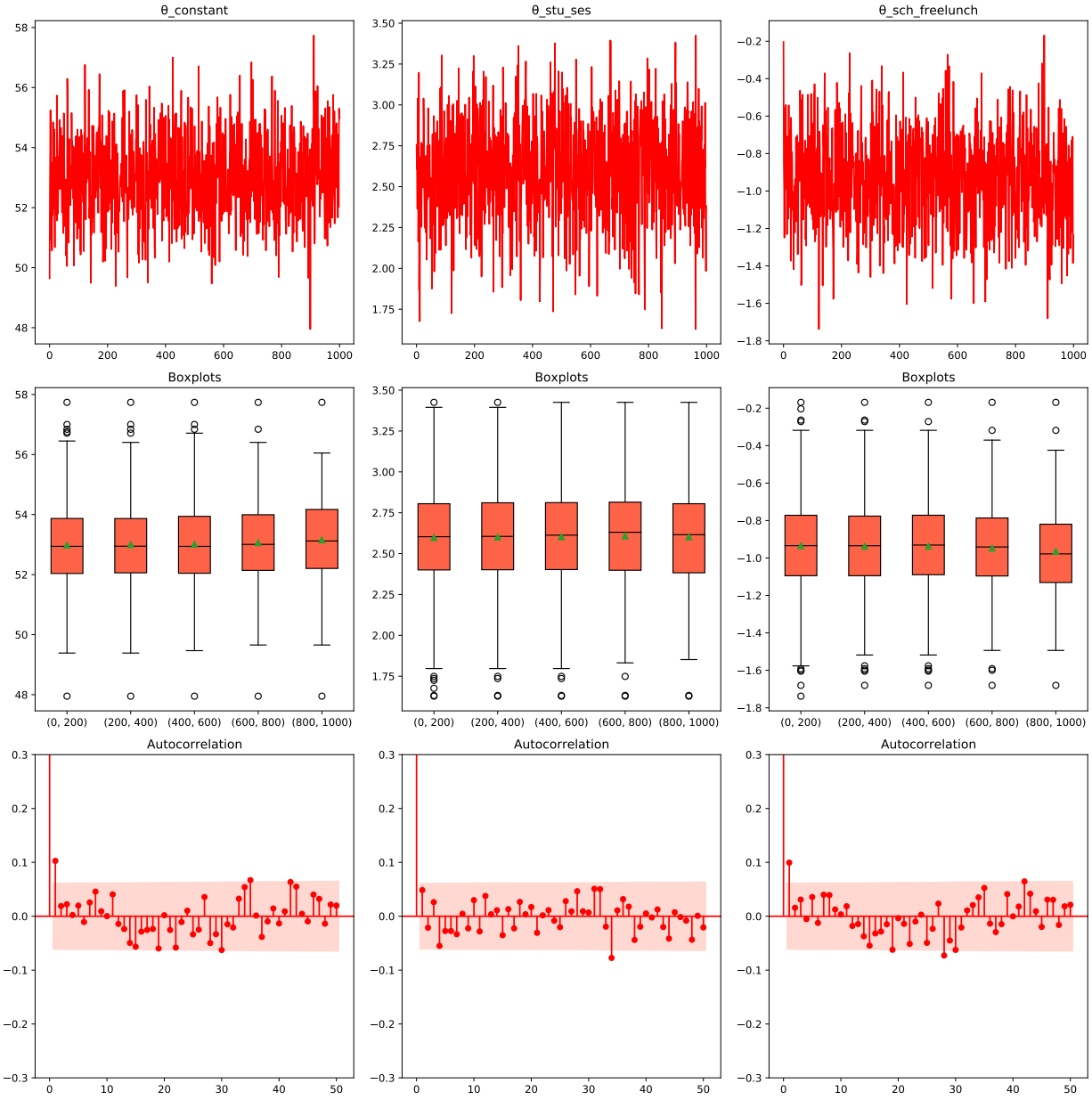

Figure 18: Trace Plots, Box Plots, and Autocorrelation Plots of $\boldsymbol{\theta}$

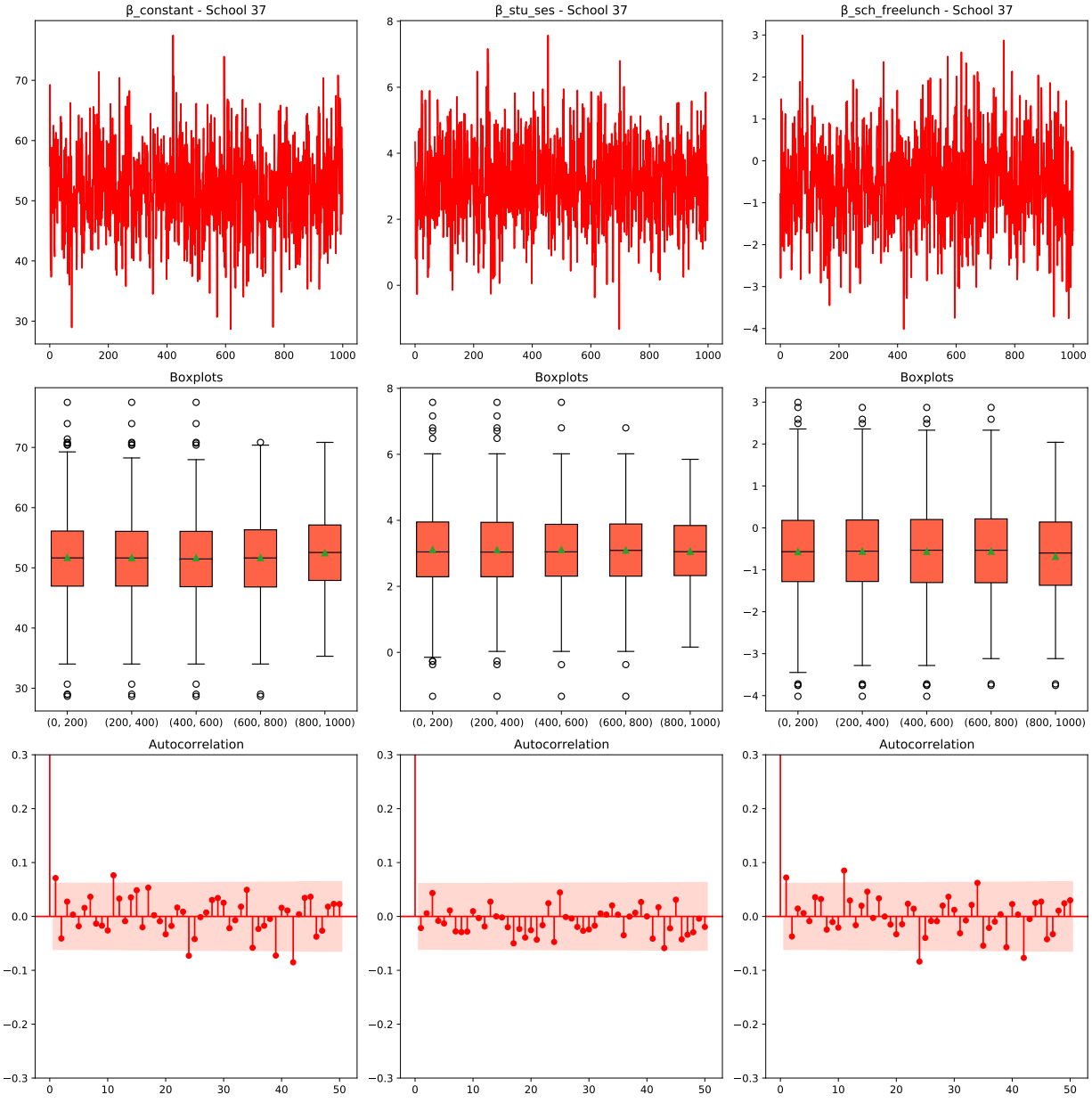

Figure 19: Trace Plots, Box Plots, and Autocorrelation Plots of $\boldsymbol{\beta}_{37}$

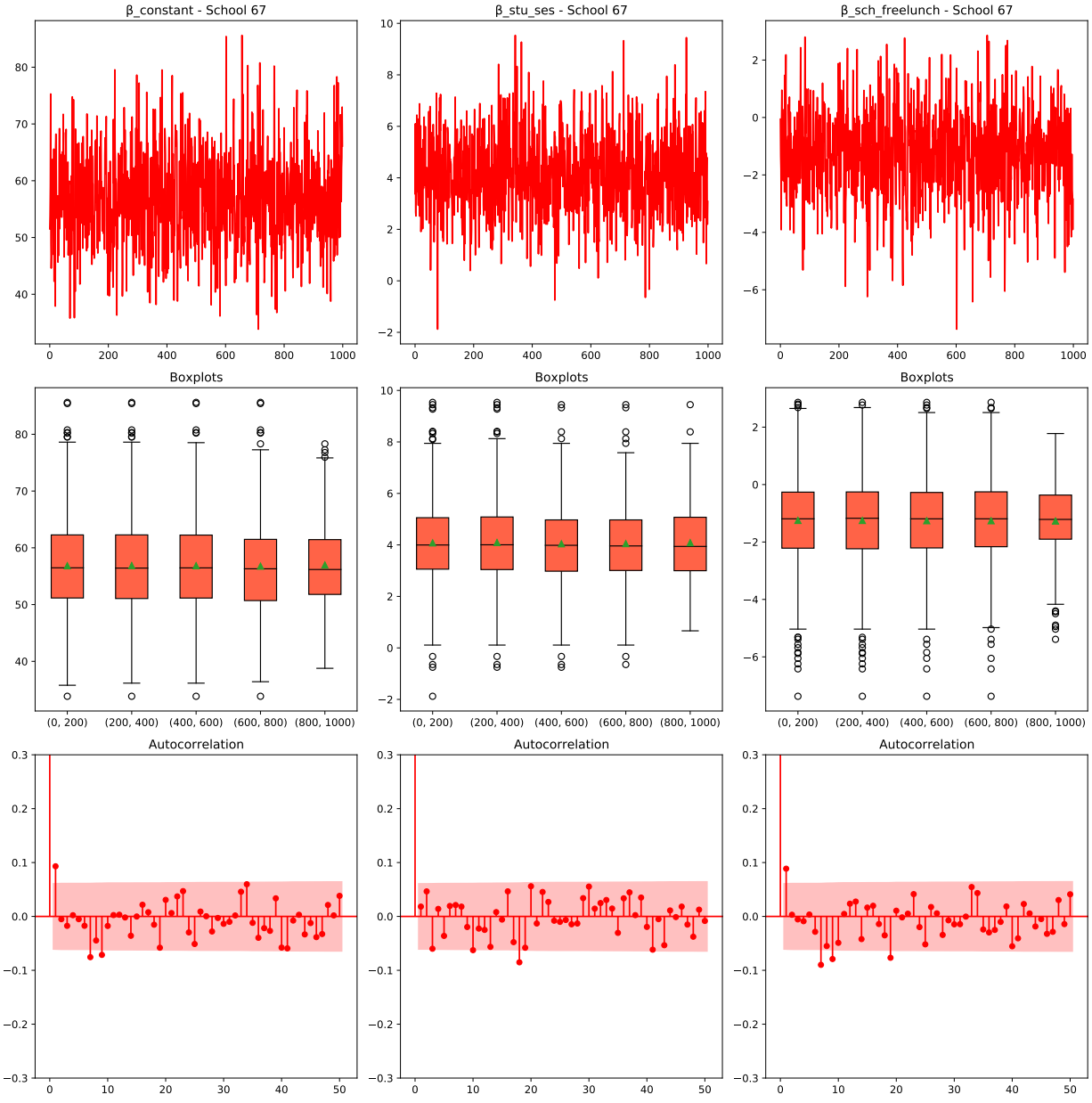

Figure 20: Trace Plots, Box Plots, and Autocorrelation Plots of $\boldsymbol{\beta}_{67}$

Finally, another Gibbs Sampler is run starting with θ equal to the zero vector, Σ equal to the identity matrix of size 3 (I_3), and σ^2 equal to σ_0^2. Figure 21 depicts the trace plots, autocorrelation plots and boxplots for $\boldsymbol{\theta}$, whereas Figure 22 and Figure 23 depict the trace plots, autocorrelation plots and boxplots of $\boldsymbol{\beta}$ for schools 37 and 67.

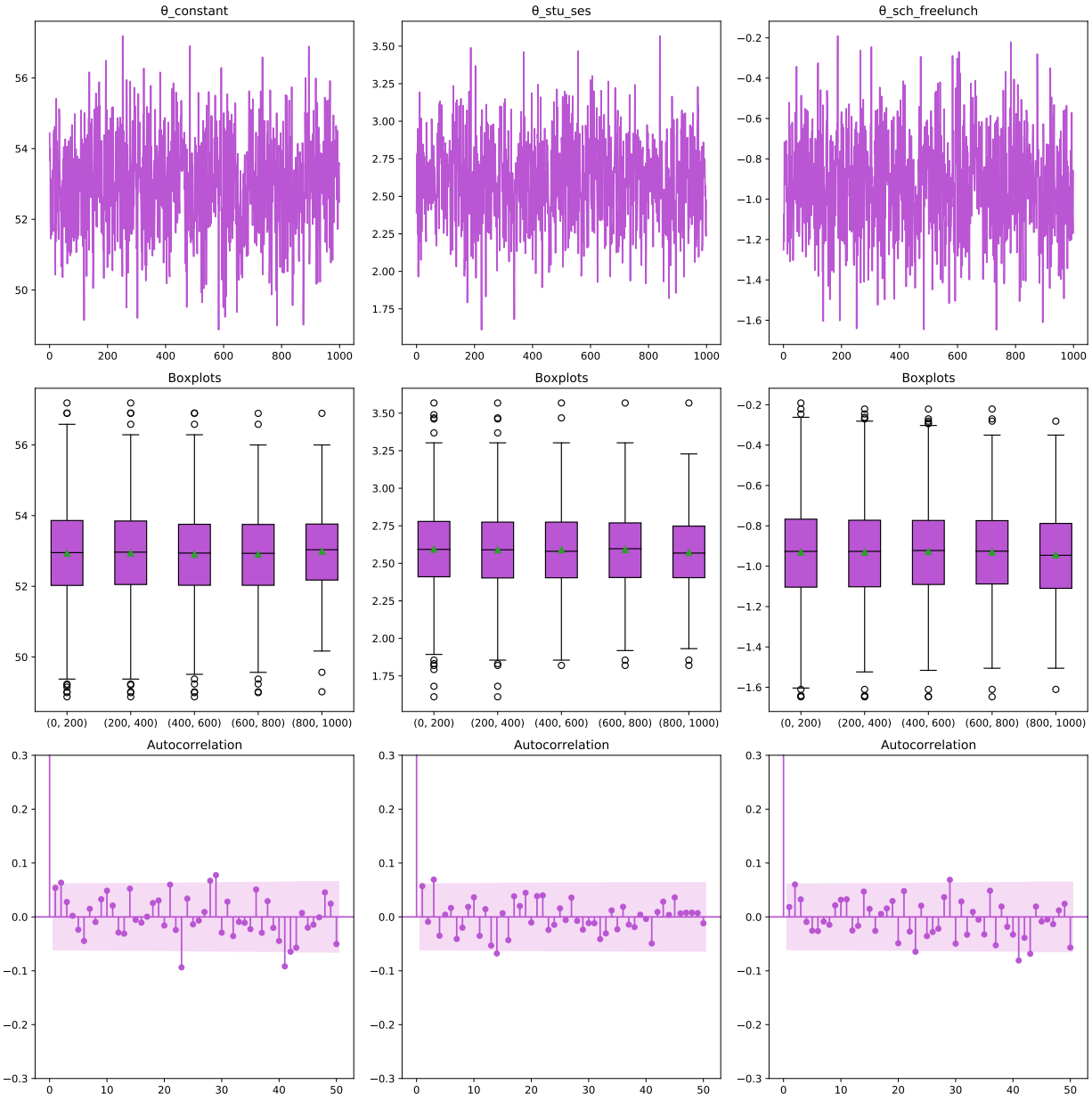

Figure 21: Trace Plots, Box Plots, and Autocorrelation Plots of $\boldsymbol{\theta}$

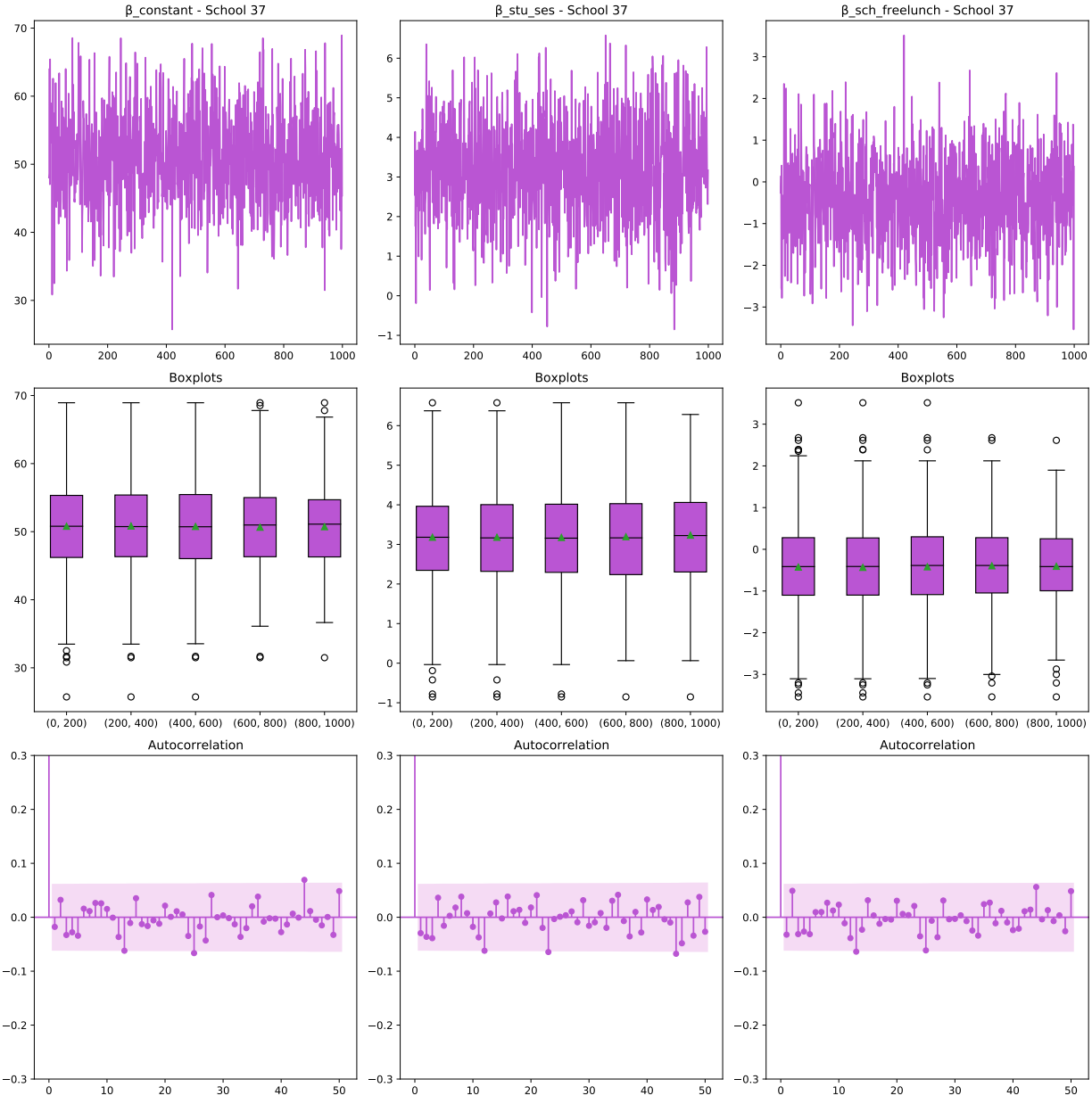

Figure 22: Trace Plots, Box Plots, and Autocorrelation Plots of $\boldsymbol{\beta}_{37}$

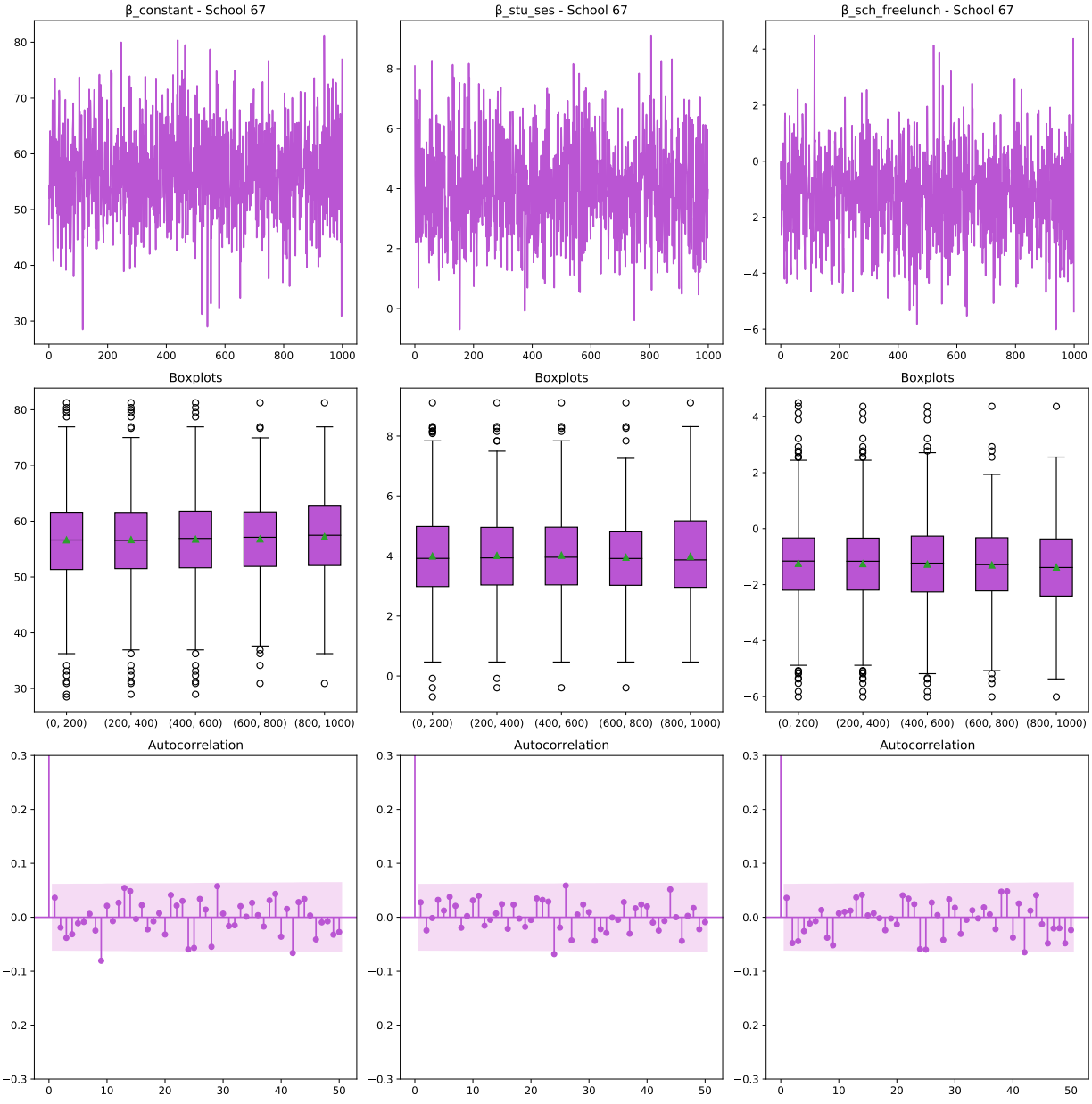

Figure 23: Trace Plots, Box Plots, and Autocorrelation Plots of $\boldsymbol{\beta}_{67}$

All of the Gibbs Samplers seem to mix well, and the five chains appear to oscillate around the same average value for each parameter. Figures 24, 25, and 26 show the comparison among the trace plots of the five Gibbs Samplers for $\boldsymbol{\theta}$, $\boldsymbol{\beta}_{37}$ and $\boldsymbol{\beta}_{67}$.

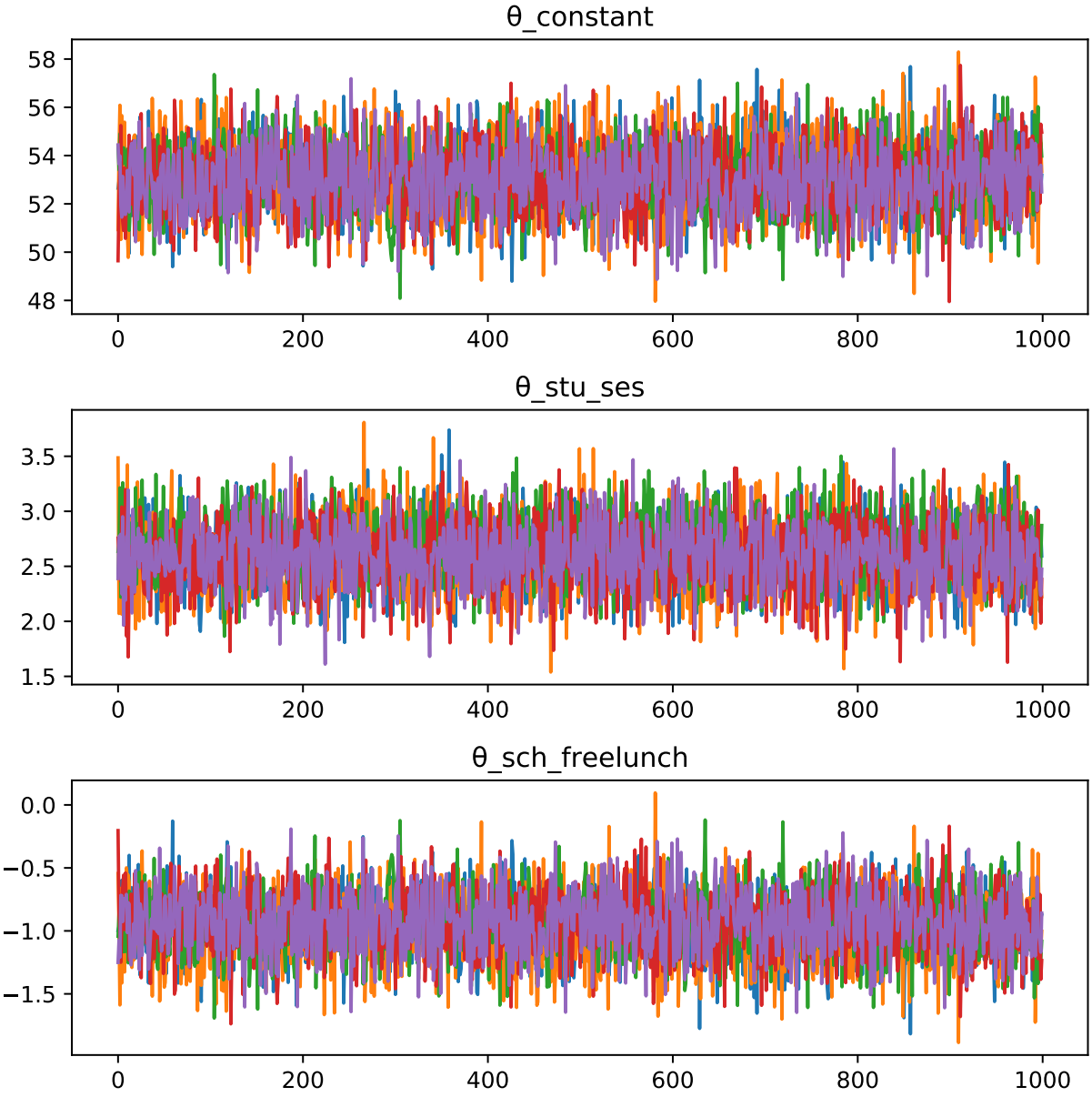

Figure 24: Comparison among the five Gibbs Samplers - Trace Plots of $\boldsymbol{\theta}$

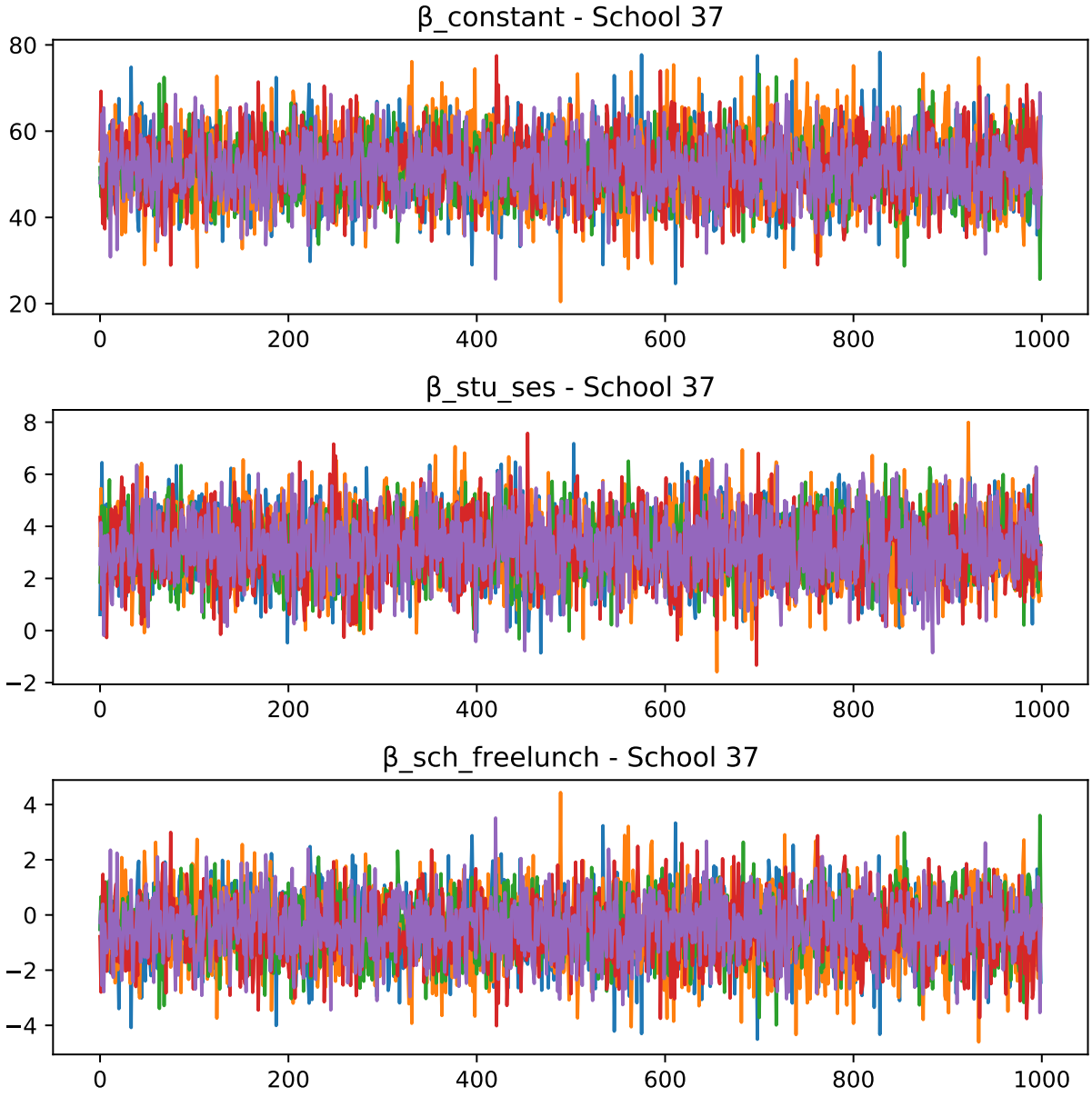

Figure 25: Comparison among the five Gibbs Samplers - Trace Plots of $\boldsymbol{\beta}_{37}$

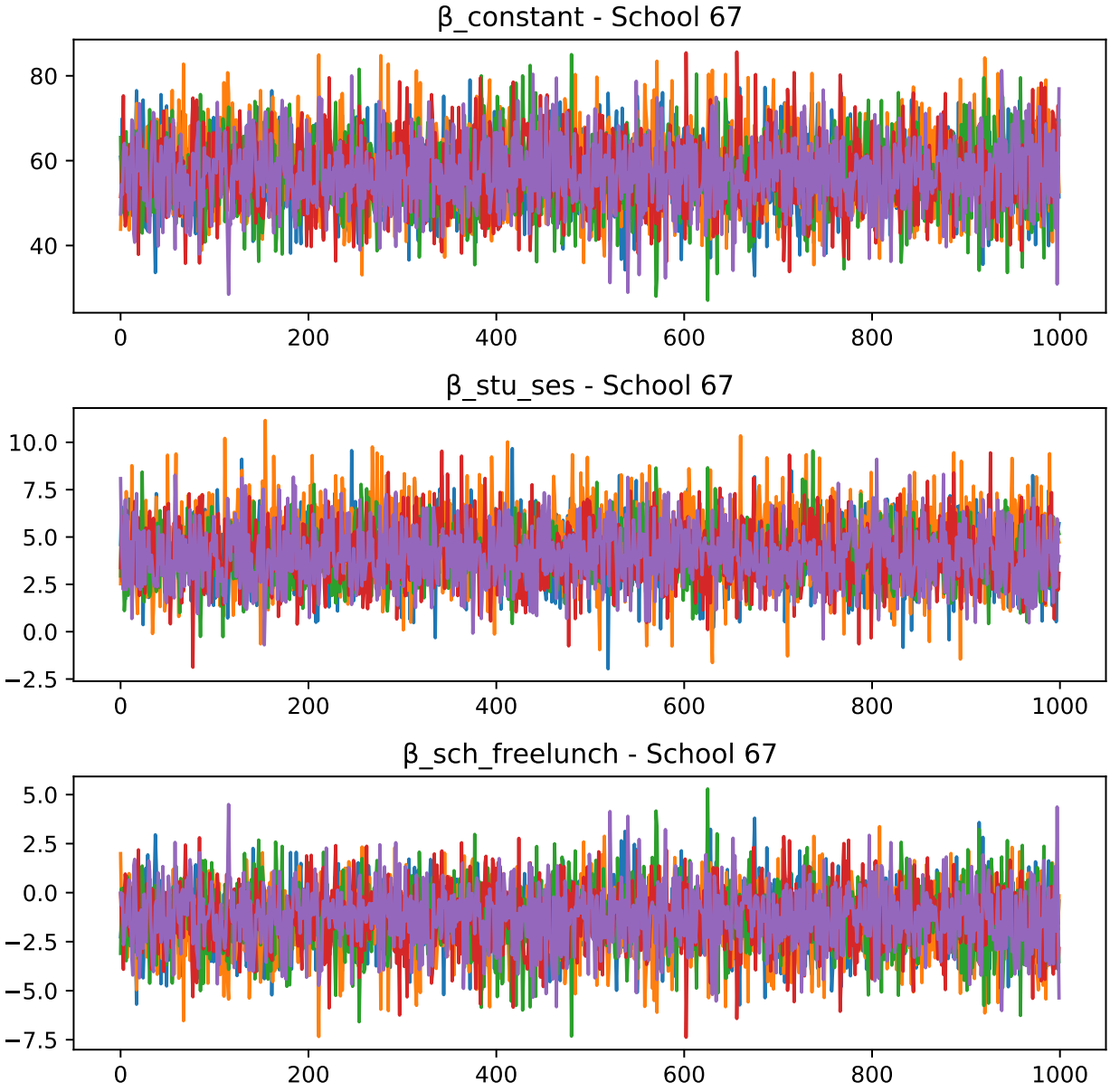

Figure 26: Comparison among the five Gibbs Samplers - Trace Plots of $\boldsymbol{\beta}_{67}$

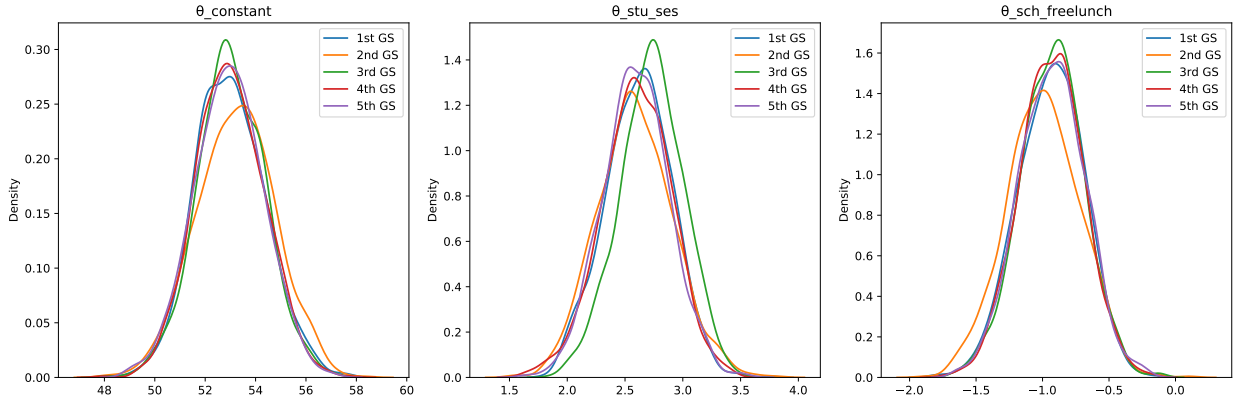

Figure 27: Comparison among the five Gibbs Samplers-Estimated posterior density of $\boldsymbol{\theta}$

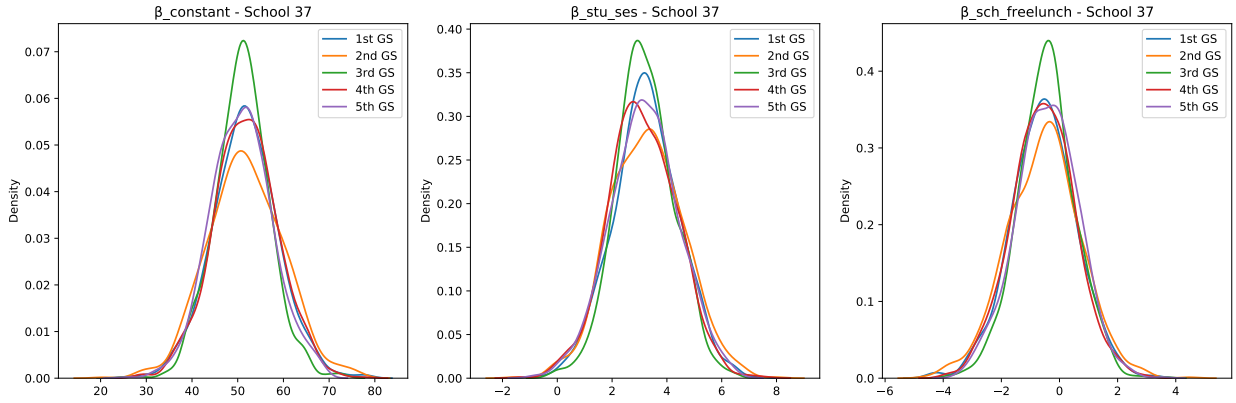

Figure 28: Comparison among the five Gibbs Samplers-Estimated posterior density of $\boldsymbol{\beta}_{37}$

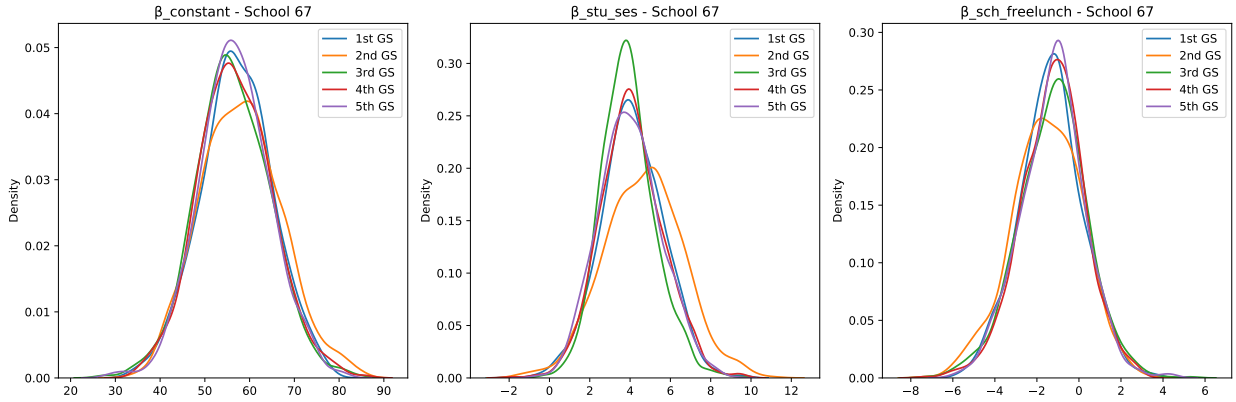

Figure 29: Comparison among the five Gibbs Samplers-Estimated posterior density of $\boldsymbol{\beta}_{67}$

Similarly, for each parameter, all of the estimated densities appear to be centered on a single value. Figures 27, 28, and 29 depict the comparison among the estimated posterior densities of the five Gibbs Samplers for $\boldsymbol{\theta}$, $\boldsymbol{\beta}_{37}$ and $\boldsymbol{\beta}_{67}$. The estimated density of the second Gibbs Sampler has thicker tails than the others due to the higher variability of the prior for Λ_0 and S_0.

The posterior density obtained by the first Gibbs Sampler is considered for the following analysis. Figure 30 compares the prior density of θ with its estimated posterior density.

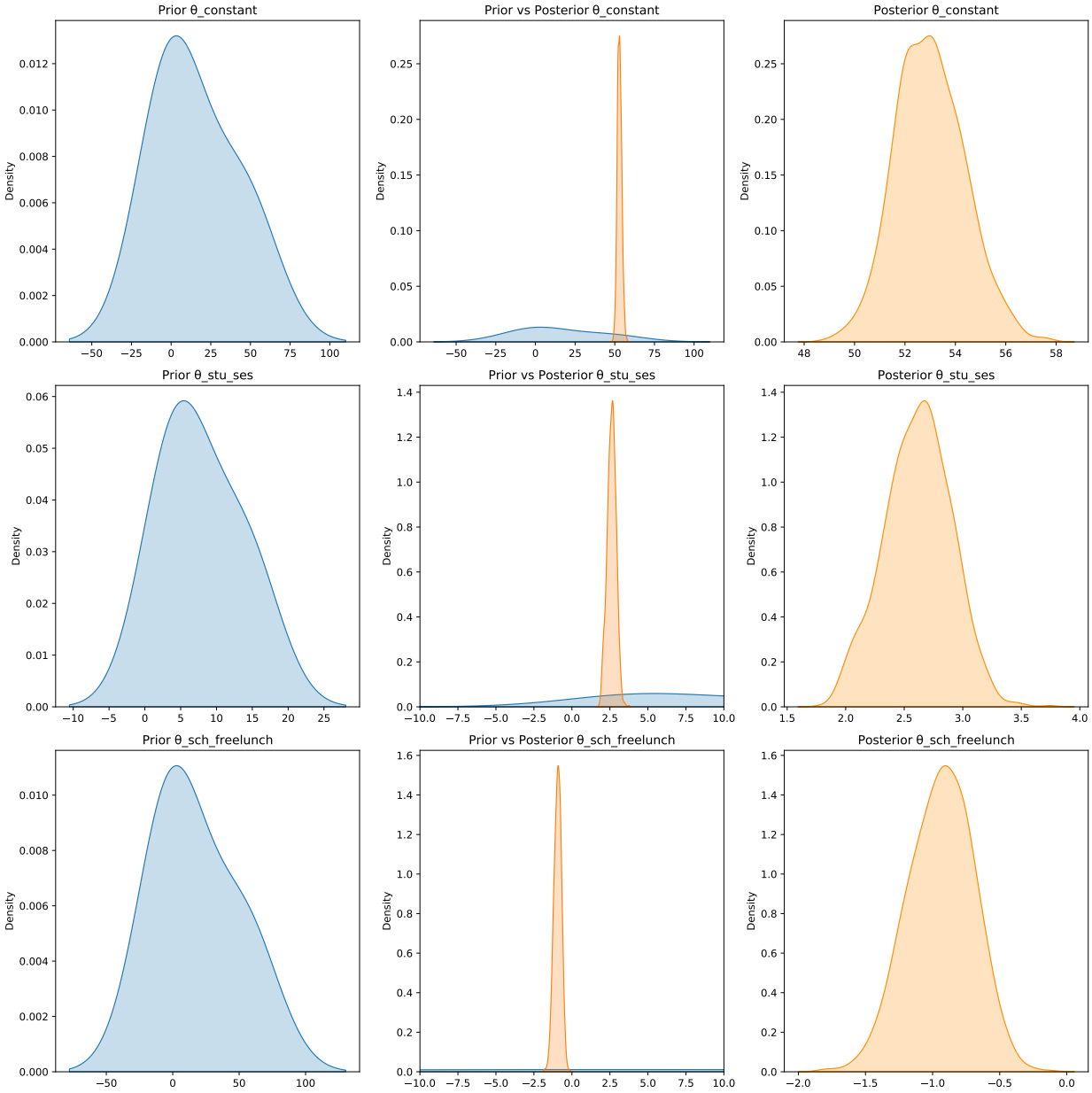

Figure 30: Prior and Estimated Posterior density of $\boldsymbol{\theta}$

From a very diffused prior, through iterations, the estimated density concentrates more and more around what seems to be the true value of the parameter. The estimate of the

between-school coefficient of student's SES is 2.617, with a 95% quantile-based posterior confidence interval equal to (2.047, 3.153). This suggests that the population average of the school-level impact of students' SES on students' math score is positive. However, the estimated coefficient of student's SES can be negative inside single schools, such as in schools 44 and 79. Nonetheless, for these two schools the 95% quantile-based posterior confidence intervals for the students' SES coefficient are, respectively, (-3.334, 2.514) and (-4.383, 2.329), which include both positive and negative values.

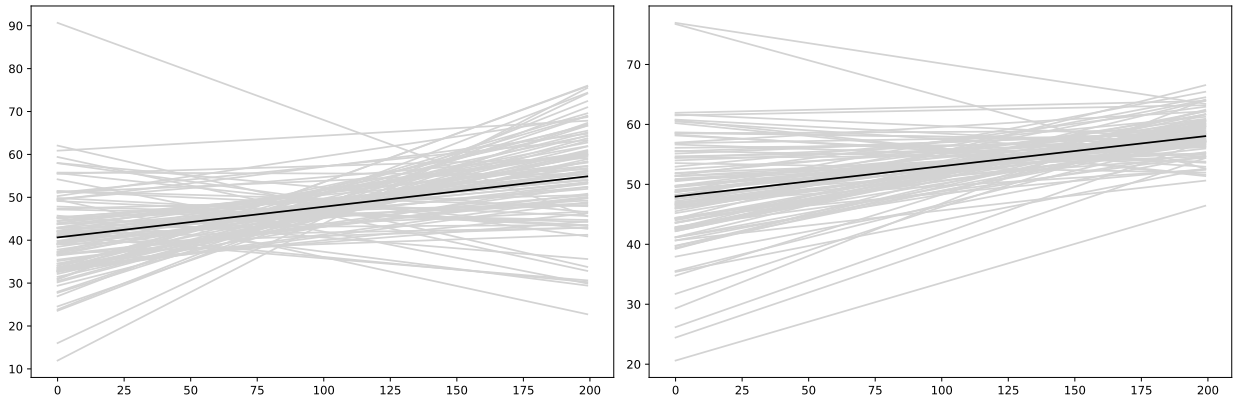

Figure 31: Comparison school-specific regression lines OLS vs Hierarchical Model

Figure 31 compares running a simple OLS in each school to using a hierarchical model. The left figure shows the 100 school-specific regression lines obtained using OLS, while the right figure shows the posterior expectations of the regression lines obtained using a hierarchical model. The regression line relative to $\boldsymbol{\theta}$ is shown in black. Hierarchical models, which enable information sharing between groups, make the extreme regression lines to shrink towards the across-group average. As a result, only a few regression lines still have a negative slope. Furthermore, a hierarchical model allows to take into consideration schools' number of free lunches, which is a factor that does not vary across groups but may affect students' math scores.

The point estimate of the between-school coefficient of school's number of free lunches is -0.938, with a 95% quantile-based posterior confidence interval equal to (-1.424, -0.487). Thus, the estimated average of the group-level effect of the schools' number of free lunches on students' math score is negative.

To conclude, the analysis suggests that schools that provide fewer free lunches and have students with a higher socioeconomic status seem to perform better in terms of math scores.