Ruichun Liu, Shubo Sun, Jilei Lin
NetID: ruichun2, shubos2, jileil2

Final Project Written Report

Analysis of Geographic Differences of Average Education Resources in Mainland China

Contents

1 Introduction **2**

2 Data **2**

3 Bayesian Hierarchical Linear Model **3**

4 Analysis **4**

4.1 Computation 4

4.2 Convergence diagnostics 4

4.3 Inference and interpretation 5

4.3.1 Descriptive statistics 5

4.3.2 Geographic differences of average education resources 6

4.4 Model diagnostics and sensitivity analysis 8

4.4.1 Normality assumption checks 9

4.4.2 Model comparison 9

4.4.3 Sensitivity analysis 9

5 Summary and conclusions **10**

6 Appendix **11**

6.1 Tables 11

6.2 Figures 15

6.3 Codes 16

6.3.1 JAGS Code 16

6.3.2 R Code 17

References **22**

1 Introduction

In this project, we are interested in analyzing regional differences of high school enrollment in mainland China. First, we explain the intuition of high school enrollment. The high school enrollment rate of a province in a given year is the ratio between the number of students admitted by high school students and the number of students who graduate from middle school. Note that the number of students to admit depends on the number of teachers and high school of that province. Moreover, the number of students who graduate from middle school can also vary from provinces to provinces. Therefore, the high school enrollment rate actually measures the average education resources enjoyed by per person.

Typically, the abundance of education resources depends on the economy of a province, which may further depends on geographic locations or its administrative roles. For example, provinces in coastal areas tend to have more business activities and thus higher GDP, leading to better education resources. Or, for instance, Beijing, as capital of China, tend to have more education resources than other provinces but it has fewer students since it is a city certainly with population less than a province. Therefore, we are motivated to analyze the geographic differences of high school enrollment rate among different geographic locations of provinces in mainland China.

We will apply a Bayesian hierarchical linear model to infer the intercepts of six regions (i.e., coastal area, central area, southwestern area, northeastern area, northwestern area, and province-level municipality). By analyzing the different intercepts generated by the model, we can explain how hidden factors or fixed effects in different regions influence high school enrollment rate. This project may provide policymakers with important implications in terms of improving elementary education system in different geographic locations.

2 Data

We combined a panel dataset from the National Bureau of Statistics of China, CNKI database, and CEinet Statistics Database. Table 1 shows the description of this panel dataset and table 2 presents its descriptive statistics. It contains 8 variables, namely, high school enrollment rate, average education funding, high school number, teacher-student rate in middle school, average GDP, average consumption, average fiscal spending, and education-fiscal spending rate, from 34 provinces, municipalities, and autonomous administrative regions in mainland China between 2004 and 2017. Some variables are not directly available in the database, so we calculate them by using other available variables. The high school enrollment rate is calculated by the percentage of the number of students admitted by high schools to the number of students

who graduate from middle schools. Average education funding (i.e., education funding per student) is the quotient obtained by dividing the annual education funding in a province by its population. Average GDP (i.e., GDP per capita) is the quotient obtained by dividing the total GDP in a province by its population. Average consumption (i.e., consumption level per capita) is the quotient obtained by dividing the total household consumption in a province by its population. Average fiscal spending (i.e., fiscal spending per capita) is the quotient obtained by dividing the total fiscal spending in a province by its population. Education-fiscal spending rate calculated by the percentage of the total education spending in a province to its total fiscal spending.

We classify the provinces into 6 geographic areas, namely, coastal , central , southwestern , northeastern, northwestern areas, and province-level municipality. Table 3 shows the classification of each province. Note that province-level municipality is separated out due to its special role in administrative division. Province-level municipality includes Beijing, Shanghai, Tianjin, and Chongqing. Though they are in the same class as province, their population, economy, and education can be quite different from other provinces in China mainland. Here, geographic area is referred to a classification group.

3 Bayesian Hierarchical Linear Model

In this section, we describe the construction of our Bayesian hierarchical normal linear model mathematically. Let $z_{i,j,t}$ denote the enrollment rate $\in [0,1]$ of the jth province in ith geographic area at year t. Consider a transformation that maps $[0,1]$ to $\mathbb{R}$ with

$$y_{i,j,t} = \log\left(\frac{z_{i,j,t}}{1 - z_{i,j,t}}\right).$$

Suppose we are given a set of covariates $x_{i,j,t,k}$ for $i = 1, \ldots, I$, where the number of geographic areas is I, $j = 1, \ldots, J_i$, where J_i is the number of provinces or municipalities of geographic area i, $t = 1, \ldots, T$ is year, and $k = 1, \ldots, K$ is for the kth covariates. Then, consider the Bayesian hierarchical model

$$\begin{aligned}
y_{i,j,t} &\sim \text{indep. } \mathrm{N}(\alpha_i + X_{i,j,t}\beta, \sigma_y^2) \\
\beta_j &\sim \text{i.i.d. } \mathrm{U}(-1000, 1000) \\
\sigma_y^2 &\sim \mathrm{U}(0, 1000) \\
\alpha_i \mid \mu_\alpha, \sigma_\alpha^2 &\sim \text{i.i.d. } \mathrm{N}(\mu_\alpha, \sigma_\alpha^2) \\
\mu_\alpha &\sim \mathrm{U}(-1000, 1000) \\
\sigma_\alpha &\sim \mathrm{Expon}(0.001)
\end{aligned}$$

where $X_{i,j,t} = (x_{i,j,t,k})_k$ is the known row of covariates for the jth province in ith geographic area at year t, and β_i is $K \times 1$. Note that the hyper-prior distribution for μ_α and σ_α and the prior distribution for β_j are intended for approximating a noninformative prior.

We select this model for its ease for interpretation and the nature of our response. Our inference objective lies mainly in the posterior distribution of α_i. However, the posterior distribution of β may aid the interpretation of α_i as well. The verification of normality assumption is in Section 4.4.1 and the sensitivity analysis for the prior is in Section 4.4.3.

4 Analysis

In this section, we present the computational details of our Bayesian analysis, and infer intuitive information from the results using Bayesian techniques.

4.1 Computation

The MCMC implementation of the Bayesian hierarchical normal linear model in Section 3 is through `rjags` package in R with JAGS software. The JAGS model code is listed in 6.3.1 and the R code for analysis is listed in Section 6.3.2.

To prevent from that the target distribution has different modes that are offset from each other, we consider four chains with over-dispersed starting points. Specifically, we consider $\beta = 10 \cdot 1_8$ and randomly sample α_i with replacement from $\{-10, 10\}$ repeatedly for four chains. We choose the burn-in stage iteration to be 5,000 and the simulation sample sample (after dropping posterior samples in burn-in stage) to be 25,000 for each chain.

4.2 Convergence diagnostics

The summary table for the Gelman-Rubin statistics is presented in Table 5 in Section 6.1. The point estimates and upper confidence intervals of Gelman-Rubin statistics of all the parameters is smaller than 1.1, the rule of thumb for judging convergence.

Moreover, the summary table for effective sample size is presented in Table 4 in Section 6.1. The effective sample sizes of all the parameters exceed 2,000. Thus, we consider it being adequate. Note that some of the parameters have effective sample size greater than 25,000, e.g., α_{central}, $\beta_{\text{high\_school\_TS\_ratio}}$, μ_α, σ_y^2. It implies that their estimation is super-efficient.

We also computed the Monte Carlo standard error using time series techniques for all the parameters. They are listed in Table 6 in Section 6.1. The trace-plots of all the parameters are

too cumbersome to present. Thus, we skip the presentation of them.

Overall, we think the convergence for MCMC approximation is achieved.

4.3 Inference and interpretation

In this section, we present our Bayesian inference and corresponding interpretation in the context of education resources. Section 4.3.1 describes the posterior quantiles and credible intervals of all the parameters and interpretation of β. Section 4.3.2 is dedicated to analysis of geographic differences of average education resources per person.

4.3.1 Descriptive statistics

We present the summary of the posterior quantile and corresponding 95% credible interval of all the parameters in Table 7 in Section 6.1. Note that the 95% credible intervals of α_i's are not meaningful since only comparisons of α_i's makes sense. Also, those of the hyper-parameters and σ_y^2 are not informative. Notably, the 95% credible intervals of β_{year}, $\beta_{\text{mid\_school\_TS\_ratio}}$, $\beta_{\text{ave\_GDP}}$, $\beta_{\text{edu\_fund\_fiscal\_rate}}$, $\beta_{\text{ave\_resident\_cons}}$, and $\beta_{\text{edu\_fund\_fiscal\_rate}}$ do not contain zeros whereas those of $\beta_{\text{ave\_edu\_fund}}$ and $\beta_{\text{mid\_school\_TS\_ratio}}$ contain zeros. Note that we assume all the geographic areas share the same β in the Bayesian hierarchical normal linear model in Section 3. Next, we explain the significance results as below.

First, from Table 7, we see that β_{year} is positive with high probability. It implies that average education resources per person increase by year. It is a good trend implying the educational policy of mainland China is working in benefiting citizens. Second, $\beta_{\text{mid\_school\_TS\_ratio}}$ is negative with high probability. It may not be intuitive at a first glance. However, if we assume the number of teacher is approximately a constant, it implies that if the number of students is larger, the enrollment rate is smaller, which makes sense. Third, $\beta_{\text{ave\_GDP}}$ is positive with high probability. It is intuitive in the sense that greater GDP implies strong economy and thus government receives more tax to invest on the education resources. Fourth, $\beta_{\text{ave\_resident\_cons}}$ is negative with high probability. It is not intuitive. It is possible that the the fact that costal provinces have higher average resident consumption but have lower average enrollment rate (explained in Section 4.3.2) induces the fitting of $\beta_{\text{ave\_resident\_cons}}$ to be negative with high probability. Fifth, $\beta_{\text{edu\_fund\_fiscal\_rate}}$ is positive with high probability, which also is intuitive. With the proportion of education funding in the fiscal funding is larger, government will construct more high schools or relevant policies will motivate more individuals in investing in high school education such that the enrollment rate should increase.

4.3.2 Geographic differences of average education resources

Based on our posterior samples from four chains, we constructed a boxplot in Figure 1 in Section 6.2. Figure 1 provides a clear view of the relationship between different geographic areas. By many non-overlapping boxes, we may conclude that significant differences of average education resources per person exist for different areas. In terms of the posterior median, the rankings from the highest to the lowest is, (1) province-level municipality, (2) northwestern provinces, (3) northeastern provinces, (4) central provinces, (5) coastal provinces, and (6) southwestern provinces. Next, we analyze the α_i's in more detail using approximated posterior probabilities as below.

The posterior probability that the province-level municipality has the greatest education resources per person is

$$\hat{\Pr}(\max_{i=1,\ldots,I} \alpha_i = \alpha_{\text{PLM}}) = \frac{1}{4 \times \#_{\text{sim}}} \sum_{j=1}^{4\times\#_{\text{sim}}} I(\alpha_{\text{PLM}}^{(j)} = \max_{i=1,\ldots,I} \alpha_i^{(j)}) = 0.93115,$$

where $\alpha_i^{(j)}$ denotes the jth observation from posterior sample of the α_i collected from the four chains. Let $R(\cdot)$ denote the rank operator with 1 being the largest. Similarly, we compute the posterior probability as below.

$$\hat{\Pr}(R(\alpha_{\text{costal}}) = 1) = 0, \quad \hat{\Pr}(R(\alpha_{\text{costal}}) = 5) = 0.8726, \quad \text{and} \quad \hat{\Pr}(R(\alpha_{\text{coastal}}) = 6) = 0.0697.$$

It is very interesting to note that the provinces in the coastal areas, whose economies are considered being very strong among all the provinces in mainland china, to have such a high posterior probability of being the second smallest in terms of education resources per person. It is to be expected that the posterior probability that province-level municipality has the highest education resources per person as explained in the introduction. Actually, these two observations are not a coincidence. From 1990, the economy of coastal areas and province-level municipality has been growing in an unprecedented speed. This attracts Chinese citizens from other provinces to study, work, and do businesses in those areas. It further motivates them to settle down and live there. Note that by educational policy, only residents in the local city can take the high school entrance examinations and thus be admitted by local high schools. So, this requires the parents of those middle school students to be local residents. Nevertheless, the formal transfer of residency to province-level municipalities is actually very difficult and more difficult than coastal provinces. Even if there are a lot of citizens working and living in province-level municipality, only a limited number of them are regarded local residents. More importantly, due to the administrative role of province-level municipality, more education resources tend to concentrate there, say for example, more high schools. This relative difficulty between province-level municipality and coastal provinces drives more transfer of residency

in coastal provinces. However, education resources of coastal province may not be enough to achieve the same level of average education resources per person as other provinces.

Additionally, we investigated the posterior probability for the rank of education resources of northeastern provinces as below.

$$\hat{\Pr}(R(\alpha_{\text{NE}}) = 1) = 0.0109, \quad \hat{\Pr}(R(\alpha_{\text{NE}}) = 2) = 0.2959, \quad \text{and} \quad \hat{\Pr}(R(\alpha_{\text{NE}}) = 3) = 0.6932.$$

Note that it is with high probability that northeastern provinces have the relatively high education resources among other provinces. This is beyond our expectations due to the slow GDP growth rate of the northeastern area these years; see for example, the 2019 GDP growth rates of Liaoning province, Heilongjiang province, and Jilin province are 3.5%, 4.5%, and 5.8% (Sina Finance, 2020) compared to 6.1%, the 2019 GDP growth rate of mainland China. In fact, northeastern three provinces is famous for their heavy industry (e.g., manufacturing, steel, energy, chemical industry). From 1940s to 1990s, the economies of northeastern provinces are among the strongest ones in mainland China. As a result, the education resources of northeastern provinces got improved. However, since 1990s, mainland China moves its economic focus to service industry, technology, finance, etc. Those development occurs in costal areas as overseas trade and transportation are easier in costal areas. Since northeastern provinces did not reform their industry structure, their economy went downhill gradually. Consequently, more and more residents in northeastern provinces selected to migrate to province-level municipality or costal areas such that the number of local residents is decreasing every year. For example, the amount of local residents is 10.696 million in 2013, which decreased to 10.414 million in 2017 (Sina Finance, 2020). As the number of residents decreases but gross education resources stay relatively changed, it is reasonable that northeastern provinces enjoy higher average education resources per person.

Next, we will talk about northwestern provinces. Due to adverse weather conditions and geographic features, many of the places of northwestern provinces are not suitable for people to leave; and thus their economies fall behind those of other provinces. For example, in the first three quarters of 2019, the GDP of Qinghai, Ningxa, Gansu and Xinjiang provinces are 204.64, 299.68, 642.60 and 912.71 billion yuan; and their rank are 28^{th}, 29^{th}, 31^{st} and 33^{rd} respectively among all 34 provinces, province-level municipalities, and autonomous administrative regions (World Economic Net, 2021). However, surprisingly, from Figure 1, we see that northwestern provinces enjoy relatively high average education resources per person. Similarly, we computed the posterior probability for the rank of education resources of northwestern provinces as below.

$$\hat{\Pr}(R(\alpha_{\text{NW}}) = 1) = 0.0579, \quad \hat{\Pr}(R(\alpha_{\text{NW}}) = 2) = 0.6451, \quad \text{and} \quad \hat{\Pr}(R(\alpha_{\text{NW}}) = 3) = 0.2970.$$

It implies that northwestern provinces have a high probability of being the second largest for average education resources per person. As illustrated above, the low GDP leads to insufficient investment in education resources. Nevertheless, Chinese government has increased the education investment on northwestern provinces from 2012. For example, the education investment of northwestern provinces in 2017 is 1.5 times that in 2012, which is significantly higher than the average increase of national one (Chinese Education Journal, 2018). Note also that the population of northwestern provinces is small than their eastern counterparts. Accordingly, with a small population and more education investment, northwestern provinces enjoy higher average education resources per person.

Last, we explained the reason why southwestern provinces were ranked as the lowest in terms of posterior median α_i from Figure 1. Likewise, we computed the posterior probability of the rank as below.

$$\hat{\Pr}(R(\alpha_{\text{SW}}) = 5) = 0.0696 \quad \text{and} \quad \hat{\Pr}(R(\alpha_{\text{SW}}) = 6) = 0.9300.$$

It can be possible due to the fact that the economy of southwestern provinces is weak. Let $T_{i,j,t}$ denote the average GDP per person of province j at year t. We computed the sample mean of the average GDP per person across geographic areas. We obtain

$$\begin{aligned} \bar{T}_{\text{central}} &= 26,800.17, \quad \bar{T}_{\text{coastal}} = 44,180.80, \quad \bar{T}_{\text{NE}} = 38,237.64, \\ \bar{T}_{\text{NW}} &= 26,464.40, \quad \bar{T}_{\text{PLM}} = 67,617.31, \quad \bar{T}_{\text{SW}} = 20,790.29. \end{aligned}$$

As we can see, southwestern provinces have the lowest sample mean of average GDP per person. This observation coincides with the fact that $\beta_{\text{ave\_GDP}}$ is positive with high probability. Unfortunately, when the income of a family is not sufficient to support the family, in China, it is often the case that teens who just finished their middle school education (required by law) in those poor areas may do not continue their high school education, which is optional according to law. Instead, they stay at local to work or move to areas where income is higher for work.

4.4 Model diagnostics and sensitivity analysis

In this section, we perform some model diagnostics regarding our Bayesian hierarchical normal linear model and sensitivity analysis of our prior. We first use Bayesian posterior predictive p-value to verify the normality assumption of the conditional sampling distribution. Next, we perform model comparison using DICs to see whether model can be further simplified. Last, we use a different prior to see whether our results are sensitive to the prior or our approximation to the noninformative prior is desirable.

4.4.1 Normality assumption checks

First, we define the residual vector as

$$\varepsilon_{i,j,t} \equiv (y_{i,j,t} - \alpha_i - X_{i,j,t}\beta)/\sigma_y.$$

Based on the Bayesian hierarchical normal linear model presented in Section 3, we should have

$$\varepsilon^{\text{rep}}_{i,j,t} \equiv (y^{\text{rep}}_{i,j,t} - \alpha_i - X_{i,j,t}\beta)/\sigma_y \,|\, (\alpha_i, \beta, \sigma_y) \sim \text{N}(0,1),$$

where $y^{\text{rep}}_{i,j,t}$ is the replicated response. To check for outlier, we consider the discrepancy measure

$$T(y_{i,j,t}, \alpha_i, X, \beta, \sigma_y) = \max_{i,j,t} |\varepsilon_{i,j,t}/\sigma_y|.$$

Based on our simulated posterior sample, we compute the posterior predictive p-value as

$$\hat{\Pr}(T(y^{\text{rep}}_{i,j,t}, \alpha_i, X, \beta, \sigma_y) \geq T(y_{i,j,t}, \alpha_i, X, \beta, \sigma_y) \,|\, y_{i,j,t}) = 0.5507.$$

By a rule of thumb of .05, we conclude that there is no outlier and thus and normality assumption is not violated.

4.4.2 Model comparison

Recall that in Section 4.3.1, we found that the credible intervals of $\beta_{\text{ave\_edu\_fund}}$ and $\beta_{\text{mid\_school\_TS\_ratio}}$ contain zero. It is natural to think whether we can reduce the model. Thus, we constructed a second model that uses the same model as in Section 3 but with a different covariate $X_{i,j,t}$, which does not include average education funding and middle school teacher-student ratio. The resulting DIC for the full model is -241.3 whereas the DIC for the reduce model is -216.9. It implies that we should prefer the full model.

4.4.3 Sensitivity analysis

Though sensitivity analysis is often implemented for a Bayesian model with informative prior, we conduct sensitivity analysis to see whether our approximation to noninformative prior is acceptable. Our alternative model for sensitivity analysis is as below

$$\begin{aligned}
y_{i,j,t} &\sim \text{indep. } \text{N}(\alpha_i + X_{i,j,t}\beta, \sigma_y^2) \\
\beta_j &\sim \text{i.i.d. } \text{U}(-100, 100) \\
\sigma_y^2 &\sim \text{U}(0, 100) \\
\alpha_i \,|\, \mu_\alpha, \sigma_\alpha^2 &\sim \text{i.i.d. } \text{N}(\mu_\alpha, \sigma_\alpha^2) \\
\mu_\alpha &\sim \text{U}(-100, 100) \\
\sigma_\alpha &\sim \text{U}(0, 100)
\end{aligned}$$

We implemented this model with the same computational details as described in Section 4.1. We computed the DIC for this new model, and it is -241.3, which is the same as that of the model in Section 3 as computed in Section 4.4.2. Therefore, we consider our approximation to noninformative prior being adequate.

5 Summary and conclusions

In this section, we present a summary of our analysis of the geographic differences of average education resources per person.

Recall in Section 1 that our response is the high school enrollment rate. It is a proxy for the average education resource per person. We used a logit transformation of the transformation and constructed a Bayesian hierarchical normal linear model as in Section 3 for analysis. We diagnosed the convergence and assumption of our models and they appear to be adequate. We found with high probability that average GDP per person, middle school teacher-student ratio, and education funding proportion of fiscal funding rate controbute positively, negatively, and positively to enrollment rate, respectively; and that the enrollment rate is increasing by years.

We analyzed the geographic differences of enrollment rate using posterior samples of α_is from a socio-economic perspective. We found that due to difficult formal transfer of residency and strong economy, province-level municipality enjoys the largest average education per person with highest posterior probability. In contrast, due to easier formal transfer of residency and strong economy, coastal provinces have relatively lower average education resources per person among all the provinces, where it has high posterior probability of being ranked the second smallest. Additionally, we find that the average education resources per person of northeastern provinces have high posterior probability of being ranked as be second largest due to its declining population but relatively sufficient gross education sources due to past strong economy. Moreover, we conclude that increased education investment of Chinese government in northwestern provinces is effective in increasing its average education resources per person; because of its relatively smaller population, its posterior probability of being ranked as the second is high. Last, we find that southwestern provinces have the lowest average education resources per person with high probability due to its lower average GDP per person.

In this project, we found many interesting results and tried to make sense of them by connecting them to the real life. The Bayesian approach is very intuitive for interpretation in this kind of study, where the practical meaning of the parameters is important.

Ruichun Liu, Shubo Sun, Jilei Lin
NetID: ruichun2, shubos2, jileil2

6 Appendix

In this section, we present the tables, figures, and codes of our analysis.

6.1 Tables

In this section, we present the tables. Table 1 describes the variables used. Table 2 describes the descriptive statistics of the variables used. Table 3 describes the detailed classification of geographic areas by provinces. Table 4 details the effective sample sizes for all the parameters. Table 5 provides the Gelman-Rubin statistics for all the parameters in MCMC approximation. Table 6 details the Monte Carlo standard error for all the parameters. Table 7 details the posterior quantiles for all the parameters.

Table 1: Variable description

Variable	Description	Unit
enroll_rate	Enrollment rate	100%
ave_edu_fund	Average education funding	RMB per student
#_high_school	High school number	1
mid_school_TS_ratio	Middle school teacher-student ratio	%
ave_GDP	Average GDP	1
ave_resident_cons	Average resident consumption	RMB per person
edu_fund_fiscal_rate	Education funding-fiscal rate	100%
high_school_TS_ratio	Middle school teacher-student ratio	%

Table 2: Descriptive statistics (sample size $n = 434$)

Variable	Mean	St. Dev.	Min	25%	75%	Max
enroll_rate	0.510	0.091	0.289	0.452	0.572	0.704
ave_edu_fund	12,305.00	10,192.57	1,495.08	4,842.22	16,365.27	63,206.70
#_high_school	464.919	235.796	22	289	603	1,031
high_school_TS_ratio	15.607	2.827	7.640	13.672	17.588	23.180
mid_school_TS_ratio	14.462	3.142	7.730	12.282	16.640	24.570
ave_GDP	36,443.87	24,061.78	4,317.00	18,215.500	46,601.00	128,994.10
ave_resident_cons	12,735.170	8,777.813	1,946	6,197.5	16,123.9	53,617
edu_fund_fiscal_rate	0.243	0.063	0.109	0.203	0.274	0.500

Table 3: Area classification

Area	Provinces
SW (Southwest)	Sichuan, Guangxi, Yunnan, Guizhou, Xizang,
PLM (province-level municipality)	Beijing, Shanghai, Tianjin, Chongqing
NW (Northwest)	Shaanxi, Gansu, Qinghai, Ningxia, Xinjiang,
NE (Northeast)	Liaoning, Neimeng, Jilin, Heilongjiang
Coastal	Hebei, Shandong, Jiangsu, Zhejiang, Fujian, Guangdong, Hainan
Central	Henan, Hubei, Hunan, Anhui, Jiangxi, Shanxi

Table 4: Effective simulation sample size ($\#_{\text{burn-in}} = 5,000$ and $\#_{\text{sim}} = 25,000$)

Parameter	Effective sample size
α_{central}	29,005.186
α_{coastal}	12,045.567
α_{NE}	10718.573
α_{NW}	40,803.277
α_{PLM}	9,920.299
α_{SW}	11,136.843
β_{year}	6,280.704
$\beta_{\text{ave\_edu\_fund}}$	5,818.175
$\beta_{\text{\#\_high\_school}}$	9,953.530
$\beta_{\text{mid\_school\_TS\_ratio}}$	4,809.531
$\beta_{\text{ave\_GDP}}$	2,426.461
$\beta_{\text{ave\_resident\_cons}}$	2,053.137
$\beta_{\text{edu\_fund\_fiscal\_rate}}$	16,079.617
$\beta_{\text{high\_school\_TS\_ratio}}$	5,684.776
μ_α	53,330.615
σ_α	15,404.283
σ_y^2	81,406.664

Table 5: Gelman-Rubin statistics ($\#_{\text{burn-in}} = 5,000$ and $\#_{\text{sim}} = 25,000$)

Parameter	Point Estimate	Upper Confidence Interval
α_{central}	1.00	1.00
α_{coastal}	1.00	1.00
α_{NE}	1.00	1.00
α_{NW}	1.00	1.00
α_{PLM}	1.00	1.00
α_{SW}	1.00	1.00
β_{year}	1.00	1.00
$\beta_{\text{ave\_edu\_fund}}$	1.00	1.00
$\beta_{\text{\#\_high\_school}}$	1.00	1.00
$\beta_{\text{mid\_school\_TS\_ratio}}$	1.00	1.00
$\beta_{\text{ave\_GDP}}$	1.00	1.01
$\beta_{\text{ave\_resident\_cons}}$	1.00	1.01
$\beta_{\text{edu\_fund\_fiscal\_rate}}$	1.00	1.00
$\beta_{\text{high\_school\_TS\_ratio}}$	1.00	1.00
μ_α	1.00	1.00
σ_α	1.00	1.00
σ_y^2	1.00	1.00
Multivariate Gelman-Rubin statistics		1.00

Table 6: Monte Carlo standard error ($\#_{\text{burn-in}} = 5,000$ and $\#_{\text{sim}} = 25,000$)

Parameter	Mean	SD	Naive SE	Time-series SE
α_{central}	-0.0323	0.0232	0.0001	0.0001
α_{coastal}	-0.0866	0.0252	0.0001	0.0002
α_{NE}	0.1757	0.0313	0.0001	0.0003
α_{NW}	0.1956	0.0241	0.0001	0.0001
α_{PLM}	0.2697	0.0388	0.0001	0.0004
α_{SW}	-0.1456	0.0288	0.0001	0.0003
β_{year}	0.1651	0.0193	0.0001	0.0002
$\beta_{\text{ave\_edu\_fund}}$	-0.0308	0.0256	0.0001	0.0003
$\beta_{\text{\#\_high\_school}}$	-0.0141	0.0148	0.0000	0.0001
$\beta_{\text{mid\_school\_TS\_ratio}}$	-0.1198	0.0221	0.0001	0.0003
$\beta_{\text{ave\_GDP}}$	0.2283	0.0395	0.0001	0.0008
$\beta_{\text{ave\_resident\_cons}}$	-0.1528	0.0403	0.0001	0.0009
$\beta_{\text{edu\_fund\_fiscal\_rate}}$	0.0393	0.0152	0.0000	0.0001
$\beta_{\text{high\_school\_TS\_ratio}}$	0.0371	0.0230	0.0001	0.0003
μ_α	0.0627	0.1115	0.0004	0.0005
σ_α	0.2416	0.1225	0.0004	0.0010
σ_y^2	0.0327	0.0023	0.0000	0.0000

Table 7: Posterior Quantiles ($\#_{\text{burn-in}} = 5,000$ and $\#_{\text{sim}} = 25,000$)

	Quantiles					
Parameter	2.5%	25%	50%	75%	97.5%	95% credible interval
α_{central}	-0.077	-0.048	-0.032	-0.017	0.013	$(-0.077, 0.013)$
α_{coastal}	-0.136	-0.103	-0.086	-0.069	-0.035	$(-0.136, -0.035)$
α_{NE}	0.113	0.154	0.175	0.196	0.236	$(0.113, 0.236)$
α_{NW}	0.149	0.179	0.195	0.211	0.242	$(0.149, 0.242)$
α_{PLM}	0.194	0.243	0.270	0.296	0.346	$(0.194, 0.346)$
α_{SW}	-0.202	-0.165	-0.145	-0.126	-0.089	$(-0.202, -0.089)$
β_{year}	0.127	0.152	0.165	0.178	0.202	$(0.127, 0.202)$
$\beta_{\text{ave\_edu\_fund}}$	-0.080	-0.048	-0.030	-0.012	0.021	$(-0.080, 0.021)$
$\beta_{\text{\#\_high\_school}}$	-0.043	-0.024	-0.014	-0.004	0.015	$(-0.042, 0.015)$
$\beta_{\text{mid\_school\_TS\_ratio}}$	-0.164	-0.136	-0.121	-0.106	-0.078	$(-0.164, -0.078)$
$\beta_{\text{ave\_GDP}}$	0.153	0.201	0.227	0.253	0.303	$(0.153, 0.303)$
$\beta_{\text{ave\_resident\_cons}}$	-0.231	-0.178	-0.152	-0.126	-0.073	$(-0.231, -0.073)$
$\beta_{\text{edu\_fund\_fiscal\_rate}}$	0.009	0.029	0.039	0.049	0.069	$(0.009, 0.069)$
$\beta_{\text{high\_school\_TS\_ratio}}$	-0.006	0.023	0.038	0.054	0.083	$(-0.009, 0.082)$
μ_α	-0.160	0.004	0.063	0.121	0.276	$(-0.160, 0.276)$
σ_α	0.112	0.166	0.211	0.280	0.550	$(0.112, 0.550)$
σ_y^2	0.028	0.031	0.033	0.034	0.037	$(0.028, 0.037)$

6.2 Figures

In this section, we present the figure used. Figure 1 provides a clear view for group differences of posterior α_is across different geographic areas.

Ruichun Liu, Shubo Sun, Jilei Lin
NetID: ruichun2, shubos2, jileil2

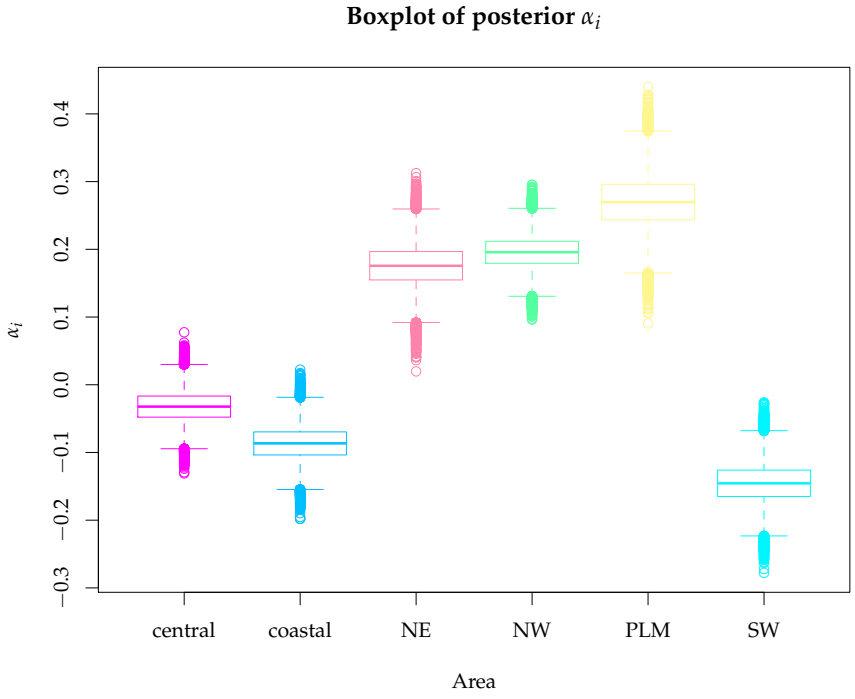

Figure 1: Boxplot of posterior α_is (40,000 posterior samples)

6.3 Codes

In this section, we present the codes of JAGS model and R for anlaysis

6.3.1 JAGS Code

The code for the model in Section 3 is

```
model{
     for (i in 2:(I + 1)) {
          for (s in (nis[i - 1] + 1):nis[i]) {
       y[s] ~ dnorm(alpha[i-1] + sum(X[s, ] * beta), 1 / sigmasqy)
     }
     alpha[i-1] ~ dnorm(mu_alpha, 1 / (sigma_alpha ^ 2))
     }


     for (j in 1:p) {
```

```
          beta[j] ~ dunif(-1000, 1000)
     }

     sigmasqy ~ dunif(0, 1000)
     mu_alpha ~ dunif(-1000, 1000)
     sigma_alpha ~ dexp(0.001)
}
```

The code for the model in Section 4.4.3 for sensitivity analysis is

```
model{
     for (i in 2:(I + 1)) {
          for (s in (nis[i - 1] + 1):nis[i]) {
       y[s] ~ dnorm(alpha[i-1] + sum(X[s, ] * beta), 1 / sigmasqy)
     }
     alpha[i-1] ~ dnorm(mu_alpha, 1 / (sigma_alpha ^ 2))
     }


     for (j in 1:p) {
          beta[j] ~ dunif(-100, 100)
     }

     sigmasqy ~ dunif(0, 100)
     mu_alpha ~ dunif(-100, 100)
     sigma_alpha ~ dunif(0, 100)
}
```

6.3.2 R Code

The code for Bayesian analysis in R is

```
# load library
library("rjags")
library("dplyr")
library('readxl')

# set working directory
setwd('~/Desktop/Courses/STAT 578 Bayesian Analysis & Computation/')
```

```
# load data
enroll <- read_excel("enroll.xlsx")
enroll <- as.data.frame(enroll)
enroll$area <- factor(enroll$area)
enroll <- enroll[order(enroll$area), ]

# construct data
z <- enroll$enroll_rate
y <- log(z / (1 - z))
X <- scale(enroll[, c(3, 9, 17, 19, 20, 21, 23, 18)])
nis <- as.numeric(table(enroll$area))
nis <- c(0, cumsum(nis))

# define inputs and initial values
input <- list(I = 6, y = y, X = X, nis = nis, p = 8)
# intialization
inits <- list(list(alpha = sample(c(-10, 10), 6, replace = TRUE),
                   beta = rep(10, 8)),
              list(alpha = sample(c(-10, 10), 6, replace = TRUE),
                   beta = rep(10, 8)),
              list(alpha = sample(c(-10, 10), 6, replace = TRUE),
                   beta = rep(10, 8)),
              list(alpha = sample(c(-10, 10), 6, replace = TRUE),
                   beta = rep(10, 8)))
# jags model m1
mod <- jags.model("final.bug", data = input, inits = inits,
                  n.chains = 4, n.adapt = 1000)
# burn-in
update(mod, 5000)
x <- coda.samples(mod, c('beta', 'alpha', 'sigmasqy', 'mu_alpha', 'sigma_alpha
                  n.iter = 25000)

# effective sample size
xtable(cbind(effectiveSize(x)), digits = 3)

# SE
```

```
require('xtable')
xtable(summary(x)[[1]], digits = 4)

# quantiles
xtable(summary(x)[[2]], digits = 3)

# gelman-rubin statistics
gelman.diag(x, autoburnin = FALSE)

# box-plot
alphas <- as.matrix(x[, paste0('alpha[', 1:6, ']')])
alphas <- matrix(alphas)
areas <- unique(enroll$area)
box <- data.frame(area = matrix(sapply(areas,
                                       FUN = function(x) rep(x, 1e+05))),
           alpha = alphas)
# load package
require('tikzDevice')
require('randomcoloR')
options(tikzLatexPackages
        = c(getOption( "tikzLatexPackages" ),
            "\\usepackage{amsmath,amsfonts,amsthm, palatino, mathpazo}"))
# image
tikz('box.tex', standAlone = TRUE, width = 6, height = 5)
boxplot(alpha ~ area, data = box,
        xlab = 'Area', ylab = '$\\alpha_i$',
        main = 'Boxplot of posterior $\\alpha_i$',
        col = 'white',
        border = c('magenta', 'deepskyblue', 'palevioletred1',
                   'seagreen1', 'khaki1', 'turquoise1'))
dev.off()

# posterior probability calculation
pos.x <- as.matrix(x)
# posterior probability for PLM to be the maximum
mean(apply(pos.x[, 1:6], 1, FUN = function(y)
  ifelse(which.max(y) == 5, yes = 1, no = 0)))
```

```
# posterior probability for costal
mean(apply(pos.x[, 1:6], 1, FUN = function(y)
  ifelse(rank(-y)[2] == 5, yes = 1, no = 0)))
mean(apply(pos.x[, 1:6], 1, FUN = function(y)
  ifelse(rank(-y)[2] == 6, yes = 1, no = 0)))
mean(apply(pos.x[, 1:6], 1, FUN = function(y)
  ifelse(rank(-y)[2] == 1, yes = 1, no = 0)))
# posterior probability for NE
mean(apply(pos.x[, 1:6], 1, FUN = function(y)
  ifelse(rank(-y)[3] == 1, yes = 1, no = 0)))
mean(apply(pos.x[, 1:6], 1, FUN = function(y)
  ifelse(rank(-y)[3] == 2, yes = 1, no = 0)))
mean(apply(pos.x[, 1:6], 1, FUN = function(y)
  ifelse(rank(-y)[3] == 3, yes = 1, no = 0)))
# posterior probability for NW
mean(apply(pos.x[, 1:6], 1, FUN = function(y)
  ifelse(rank(-y)[4] == 1, yes = 1, no = 0)))
mean(apply(pos.x[, 1:6], 1, FUN = function(y)
  ifelse(rank(-y)[4] == 2, yes = 1, no = 0)))
mean(apply(pos.x[, 1:6], 1, FUN = function(y)
  ifelse(rank(-y)[4] == 3, yes = 1, no = 0)))
# posterior probability for NW
mean(apply(pos.x[, 1:6], 1, FUN = function(y)
  ifelse(rank(-y)[6] == 6, yes = 1, no = 0)))
mean(apply(pos.x[, 1:6], 1, FUN = function(y)
  ifelse(rank(-y)[6] == 5, yes = 1, no = 0)))

# assumption check
post.sigmasqy <- as.matrix(x)[, 'sigmasqy']
post.alpha <- as.matrix(x)[, paste0('alpha[', 1:6, ']')]
post.beta <- as.matrix(x)[, paste0('beta[', 1:8, ']')]
err.std.sim <- matrix(NA, 40000, length(z))
indic.mat <- model.matrix(~ enroll$area - 1)
for (s in 1:40000) {
  err.std.sim[s, ] <- (y - (indic.mat %*% cbind(post.alpha[s, ]) +
                              X %*% cbind(post.beta[s, ]))) /
    sqrt(post.sigmasqy[s])
```

```
}
ref.std.normal <- matrix(rnorm(40000 * length(z)), 40000, length(z))
mean(apply(abs(ref.std.normal), 1, max) >=
       apply(abs(err.std.sim), 1, max))

# model 2
X2 <- scale(enroll[, c(3, 17, 20, 21, 23, 18)])

# define inputs and initial values
input <- list(I = 6, y = y, X = X2, nis = nis, p = 6)
# intialization
inits2 <- list(list(alpha = sample(c(-10, 10), 6, replace = TRUE),
                    beta = rep(10, 6)),
               list(alpha = sample(c(-10, 10), 6, replace = TRUE),
                    beta = rep(10, 6)),
               list(alpha = sample(c(-10, 10), 6, replace = TRUE),
                    beta = rep(10, 6)),
               list(alpha = sample(c(-10, 10), 6, replace = TRUE),
                    beta = rep(10, 6)))
# jags model m1
mod2 <- jags.model("final.bug", data = input, inits = inits2,
                   n.chains = 4, n.adapt = 1000)
update(mod2, 5000)
# DICs
dic.samples(mod, 25000) # -241.3
dic.samples(mod2, 25000) # -216.6

# sensitivity
alt.mod <- jags.model("final2.bug", data = input, inits = inits,
                      n.chains = 4, n.adapt = 1000)
update(alt.mod, 5000)
dic.samples(alt.mod, 25000)
```

References

Chinese Education Journal. (2018). *The education investment align against central and western areas; the education investment in 2018 exceed 13 billion yuan.* https://www.sohu.com/a/233068906_498091. (Accessed on 2021-05-06)

Sina Finance. (2020). *Lower than national average! the 2019 gdp growth rate of northeastern three provinces are 3.5%, 4.5%, and 5.8%.* https://baijiahao.baidu.com/s?id=1656081840282998355&wfr=spider&for=pc. (Accessed on 2021-05-05)

World Economic Net. (2021). *The 2020 gdp of northwestern provinces and the gdp ranking of national provinces in 2020.* https://www.shijiejingji.net/rediantupian/20210317/316690.html. (Accessed on 2021-05-06)