Course project for ISyE 6420

Bayesian Regression Analysis on Factors Influencing Number of Covid-19 Cases in Different Provinces of Mainland China

Jingyu Li

Apr 20, 2020

Content

COVID-19, an on going pandemic all over the world

2 245 872

Confirmed cases

Last update: 18 April 2020, 20:00 GMT-4

152 707

Confirmed deaths

Last update: 18 April 2020, 20:00 GMT-4

213

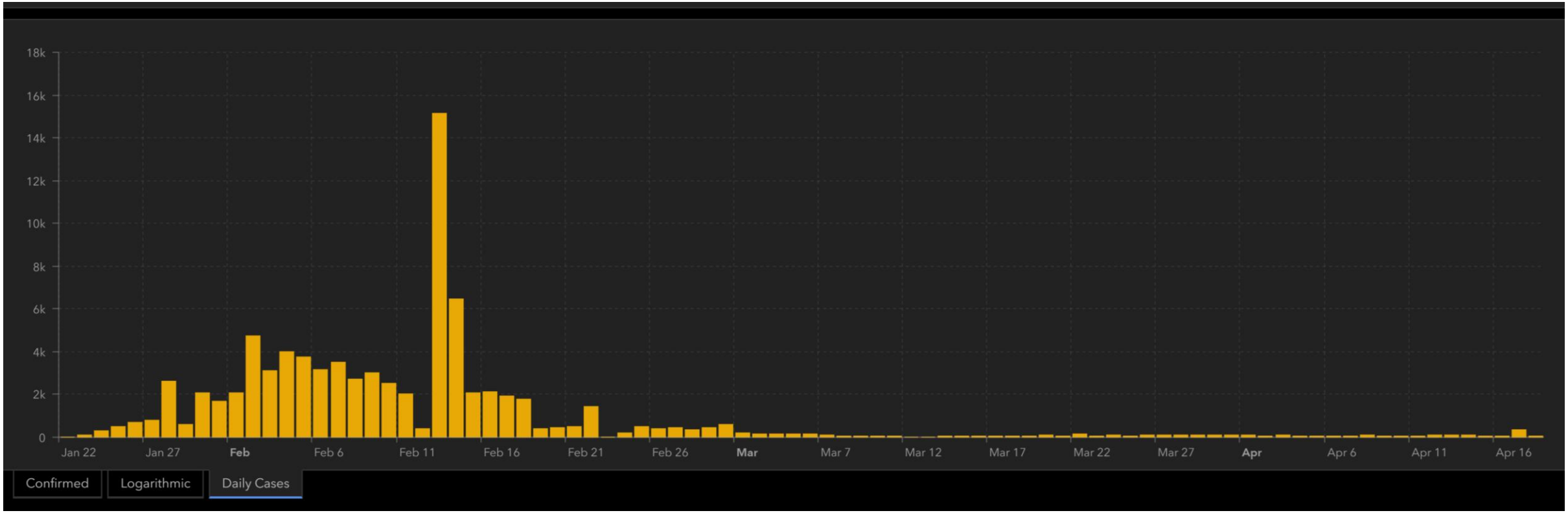
Countries, areas or territories with cases

Last update: 18 April 2020, 20:00 GMT-4

Source: WHO

Due to strict lockdown policy, COVID-19 has been controlled in China

Daily Confirmed Cases in China

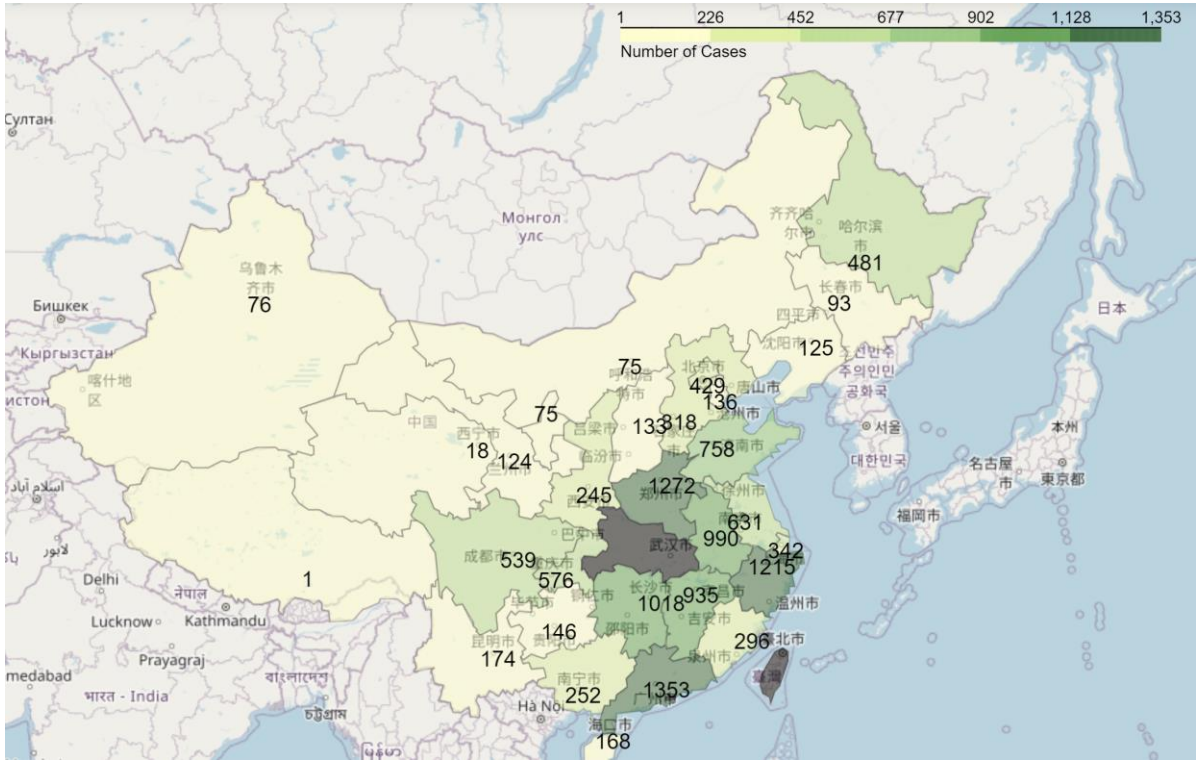


Source: Johns Hopkins University

Problem: exploring the factors that influence the number of confirmed cases in each province of mainland China

Total Confirmed Cases in Mainland China

(Hubei excluded)



- Did the provinces adjacent to Hubei geographically have more cases?
- Was the number of cases different between more developed and less developed provinces?

Source: National Health Commission of China

Content

Data collection

30 provinces in Mainland China. Hubei, Taiwan, Hongkong and Macau are excluded.

Response variable	• **Cases:** Total number of confirmed cases by the end of March 10
Explanatory variables	• **Population:** Resident population by the end of 2018 (Unit: 10,000). • **GDP:** 2019 annual GDP (Unit: 100 Million RMB). • **Distance:** Direct distance between each province's capital city with Wuhan (Unit: kilometer). • **PassengerTurnover:** 2018 annual railway passenger turnover – total number of passenger X average travel distance per passenger (Unit: 100 million passengers * kilometer). • **TravelConnection:** Percentage of people travelled from Wuhan on January 15.

Source: National Health Commission, National Bureau of Statistics, Baidu Map

Data collection

Source: National Health Commission, National Bureau of Statistics, Baidu Map

Data exploration

Correlation Matrix

	Cases	Population	GDP	Distance	PassengerTurnover	TravelConnection
Cases	1.000000	0.670640	0.694976	-0.589030	0.682056	0.784423
Population	0.670640	1.000000	0.842882	-0.501326	0.859728	0.647999
GDP	0.694976	0.842882	1.000000	-0.507243	0.688214	0.494538
Distance	-0.589030	-0.501326	-0.507243	1.000000	-0.574257	-0.586913
PassengerTurnover	0.682056	0.859728	0.688214	-0.574257	1.000000	0.712156
TravelConnection	0.784423	0.647999	0.494538	-0.586913	0.712156	1.000000

Data exploration

Linear Regression

```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)      -6.863e-17  9.925e-02   0.000 1.000000
Population       -2.964e-01  2.769e-01  -1.071 0.294973
GDP               5.556e-01  1.978e-01   2.810 0.009711 **
Distance         -4.386e-02  1.351e-01  -0.325 0.748183
PassengerTurnover 9.726e-02  2.223e-01   0.438 0.665604
TravelConnection  6.067e-01  1.541e-01   3.936 0.000619 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5436 on 24 degrees of freedom
Multiple R-squared:  0.7554,    Adjusted R-squared:  0.7045
F-statistic: 14.83 on 5 and 24 DF,  p-value: 1.131e-06
```

Content

Theoretical Analysis

Model settings

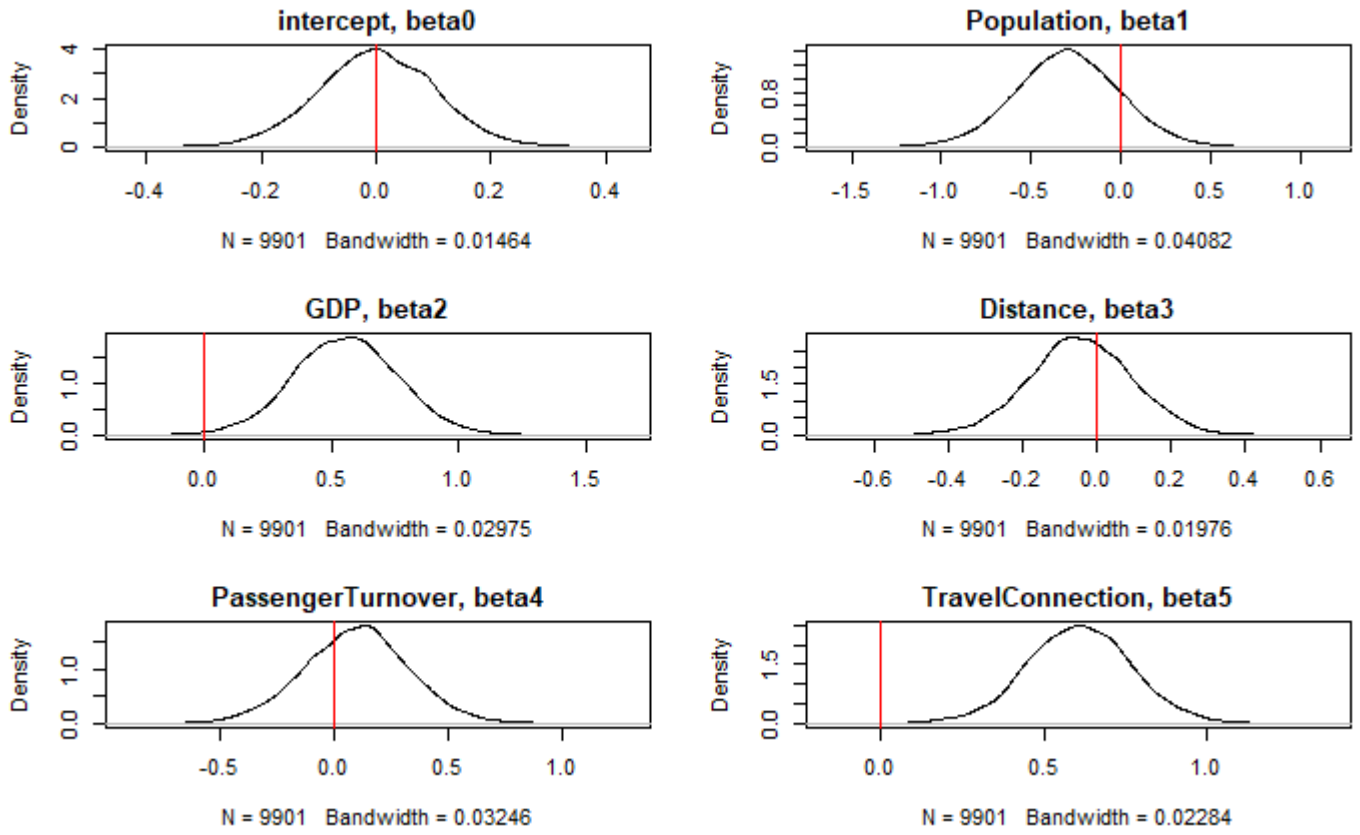
$$y = number\ of\ confirmed\ cases$$
$$x_1 = Population, x_2 = GDP, x_3 = Distance,$$
$$x_4 = PassengerTurnover, x_5 = TravelConnection$$
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \epsilon, \epsilon \backsim^{iid} N(0, \sigma^2)$$
$$y\,|\beta, \sigma^2 \backsim N(X\beta, \sigma^2 I)$$
$$P(\beta, \sigma^2) \propto \frac{1}{\sigma^2}$$

Conditional posteriors

$$\beta|\sigma^2, y \backsim N((X^T X)^{-1} X^T y, \sigma^2 (X^T X)^{-1})$$
$$\sigma^2|\beta, y \backsim Inver - Gamma(\frac{n}{2}, \frac{e^T e}{2}), e = y - X\beta$$

Gibbs Sampling

Density Plot

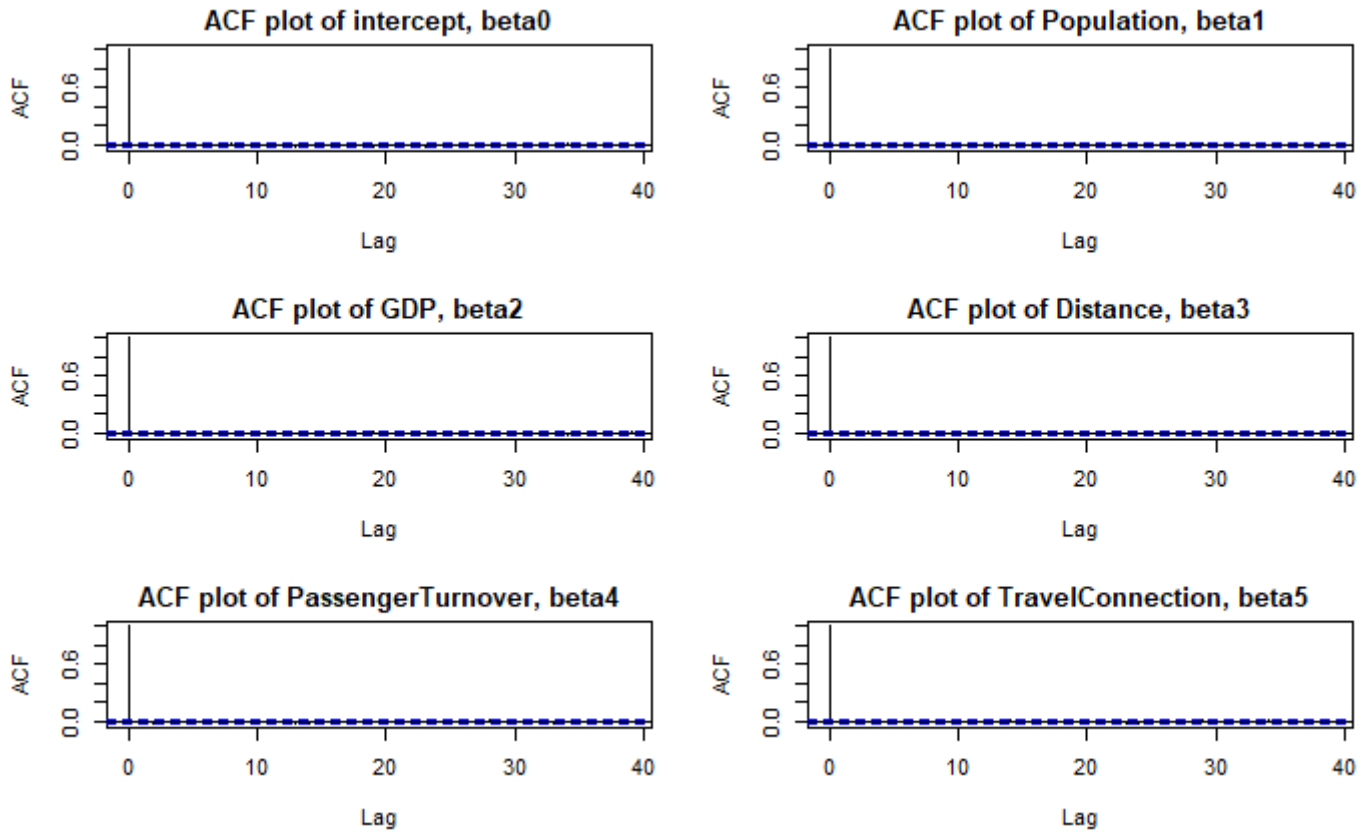


95% HPD Interval

Variable	Lower	Upper
(intercept)	-0.200	0.204
Population	-0.870	0.288
GDP	0.128	0.958
Distance	-0.313	0.240
PassengerTurnover	-0.377	0.552
TravelConnection	0.300	0.946

Gibbs Sampling: MCMC Diagnostics

ACF plot



Effective Size

Parameters	Lower
β_0	9,408
β_1	10,000
β_2	10,000
β_3	9,679
β_4	10,000
β_5	10,000
σ^2	6,360

Content

Conclusion and Discussion

Positive influence of TravelConnection

- Explain: reflects the number of people who travelled from Wuhan to other provinces and how close the two places are related with each other in terms of population migration before Spring Festival.

Positive influence of GDP

- Explain: economic activities in China are highly correlated among provinces. People in provinces with higher GDP may have more travel needs. As a result, their exposure risk can also be higher.

Population showed no effect

- Explain: one of the main reasons is that locking-down and self-quarantine orders were strictly executed, significantly decreasing inter-personal contact and slowing down local infection.

Thank you!