Brooklyn Bridge Pedestrian Estimation

STATS 451

Final Project

Group 14

Junfeng Luo/Dong Ding

Introduction

Topic:

Brooklyn Bridge is one of the most famous scenic spots of New York City. It connects Manhattan and Brooklyn and it is also one of the busiest bridges of New York City. Our team is quite interested in finding out the number of pedestrians that going through the bridge within a certain period and we try to find out what factors will influence the rate of pedestrians. We hope that our findings will be helpful to the department of traffic to regulate traffic on Brooklyn Bridge and to the peddlers who would like to know how many beverages or snacks they should prepare if they want to have some small businesses on Brooklyn Bridge.

Data:

The data[1] is from NYC OpenData, which is a public website devoted to providing data of New York City in order to help their citizens improve the life living in New York City. The data was collected by DOT (US Department of Transportation) by using automated technology. The data was recorded from 10/1/2017 to 7/31/2018. For each observation, the number of pedestrians was measured within one hour.

The dataset includes 12 variables which are hour_beginning (the date and time the counting process began), location (the location of the counting site), Pedestrians (the number of Pedestrians within one hour), Towards Manhattan (the number of Pedestrians towards Manhattan), Towards Brooklyn (the number of Pedestrians towards Brooklyn), weather_summary (weather description including cloudy, rain etc.), temperature (the hourly temperature, in Fahrenheit), precipitation (hourly precipitation), lat (latitude), long (longitude), events (including holidays etc.) and Location1. Below is an example of the dataset.

hour_beginning	location	Pedestrians	Towards Manhattan	Towards Brooklyn	weather_summary	temperature	precipitation	lat	long	events	Location1
10/1/2017 0:00	Brooklyn E	44	30	14	clear-night	52	0.0001	40.70816	-73.9995		(40.7081639691088, -73.9995087014816)
10/1/2017 1:00	Brooklyn E	30	17	13	partly-cloudy-night	53	0.0002	40.70816	-73.9995		(40.7081639691088, -73.9995087014816)
10/1/2017 2:00	Brooklyn E	25	13	12	partly-cloudy-night	52	0	40.70816	-73.9995		(40.7081639691088, -73.9995087014816)
10/1/2017 3:00	Brooklyn E	20	11	9	partly-cloudy-night	51	0	40.70816	-73.9995		(40.7081639691088, -73.9995087014816)
10/1/2017 4:00	Brooklyn E	18	10	8	partly-cloudy-night	51	0	40.70816	-73.9995		(40.7081639691088, -73.9995087014816)

Figure 1 Example of the dataset

Process

Data Cleaning:

There are 12 variables in the dataset. We only kept hour_beginning, Pedestrians, weather_summary, temperature for our analysis. Other variables, like events, might also have some influences on the number of pedestrians going through the bridge. However, the main purpose of our analysis is to predict the number of pedestrians generally. We hoped that our results to be more universal. Thus, here we were only considering two factors that would influence the number of pedestrians, weather and temperature.

For weather, we divided those weathers into two groups, good weather and bad weather. We have ‘clear-night’ and ‘clear-day’ for good weather; we have ‘rain’, ‘fog’, ‘snow’, ‘sleet’, ‘cloudy’, ‘wind’, ‘partly-cloudy-night’ and ‘partly-cloudy-day’ for bad weather.

Because the variable temperature is continuous, not factor. We decided to divide temperature into three groups. We considered that below 10 Celsius (50 Fahrenheit) as low temperature, between 10 Celsius (50 Fahrenheit) and 26 Celsius (78.8 Fahrenheit) as medium temperature and above 26 Celsius (78.8 Fahrenheit) as high temperature.

Then we checked the data of a single day. The example is shown below:

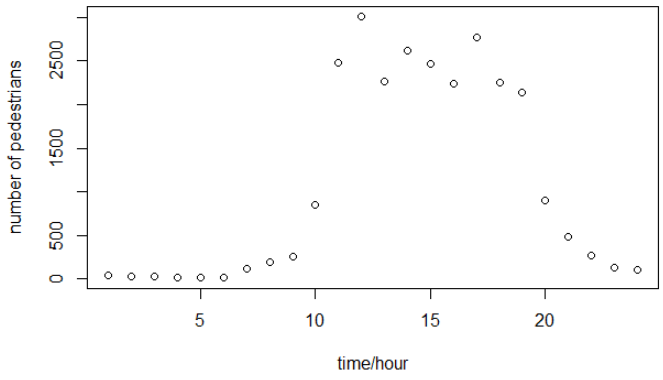

Figure 2 Histogram of a single day

From the plot, we can see that the number of pedestrians of a single day fluctuated and most people went through the bridge between 10am and 8pm. Thus, we would only use the data inside this interval.

Model Selection:

The basic idea of the analysis is Bayesian Analysis. We wanted to learn the rate of pedestrians going through the bridge. The rate was our unknown parameter. We made histograms for the observations:

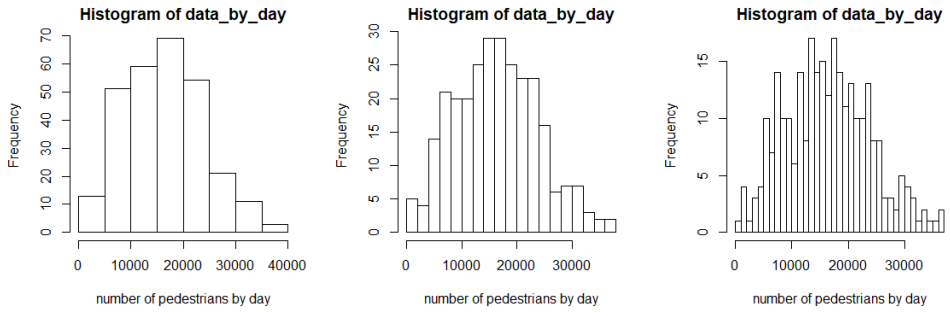

Figure 3 Histogram of observations

We used different bars and from the results we found that the observations were close to normal distribution. Thus, we predicted that normal distribution might be a good likelihood function in this case. Otherwise, because we wanted to predict the rate of pedestrians. We thought Poisson distribution might also be good here. Thus, we used normal distribution and Poisson distribution as our likelihood functions.

Analysis:

We separated the data by two different ways, which one of them was by weather and the other one was by temperature. We used the Stan package in R to run the whole analysis.

Results

Results about weather:

```
           mean se_mean    sd      2.5%       25%       50%       75%
mu1      1741.04    0.36 22.85   1694.61   1725.82   1741.24   1756.40
sigma1    735.22    0.26 16.32    703.89    723.95    734.81    746.14
mu2      1368.22    0.33 21.38   1326.85   1353.32   1368.21   1383.25
sigma2    737.88    0.22 15.11    709.36    727.67    737.42    747.90
lp__   -15951.28    0.03  1.42 -15954.61 -15952.00 -15950.99 -15950.23
           97.5% n_eff Rhat
mu1      1785.37  3984    1
sigma1    767.61  4025    1
mu2      1409.07  4240    1
sigma2    767.59  4564    1
lp__   -15949.51  1980    1
```

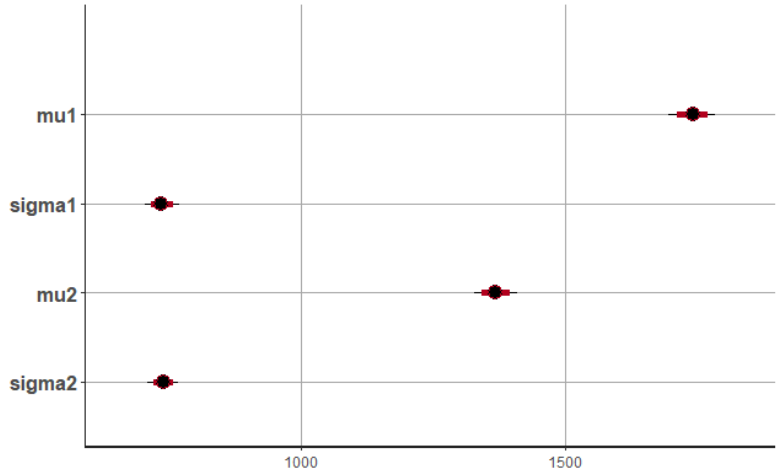

Figure 4 Parameters of normal likelihood function (weather)

	mean	se_mean	sd	2.5%	25%	50%
lambda1	1741.27	0.02	1.27	1738.79	1740.42	1741.25
lambda2	1367.71	0.02	1.09	1365.61	1366.98	1367.71
lp__	21980667.95	0.02	1.01	21980665.23	21980667.59	21980668.24

	75%	97.5%	n_eff	Rhat
lambda1	1742.13	1743.76	3437	1
lambda2	1368.45	1369.80	3218	1
lp__	21980668.66	21980668.93	1766	1

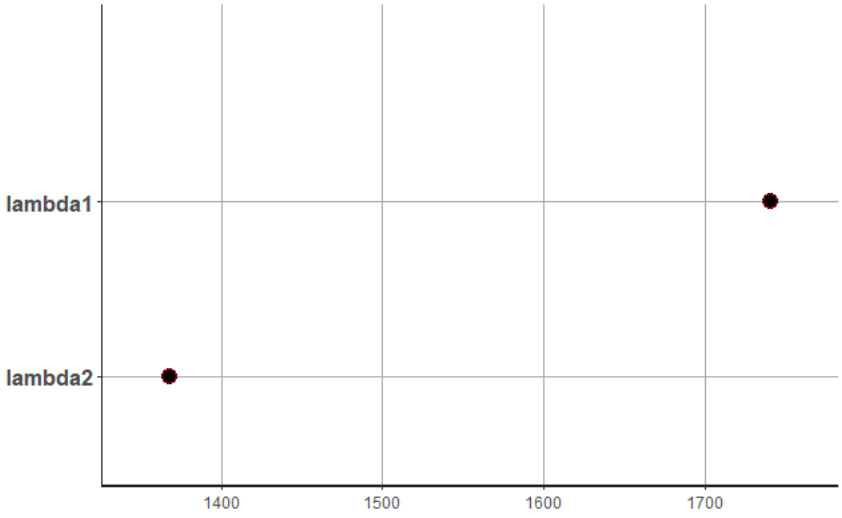

Figure 5 Parameters of Poisson likelihood function (weather)

From the results and graphs above, we can see that when using normal distribution as likelihood function the variances of the two groups are pretty close to each other, and the difference between the mean of two groups is distinct. When using Poisson distribution as likelihood function, the predicted mean of two groups are pretty close to what we got from normal distribution. However, we know that for Poisson distribution, the variance should be equal to the mean. The normal distribution shows that the variance of the model is not close to the mean. Hence, Poisson distribution might not be a good choice here. But the difference between the parameters is still huge. Thus, we have reasons to believe

that the number of pedestrians going through the bridge when the weather is good is more than the number of pedestrians going through the bridge when the weather is bad.

Results about temperature:

In general, we can see that the higher the temperature, the more the pedestrians going through the bridge:

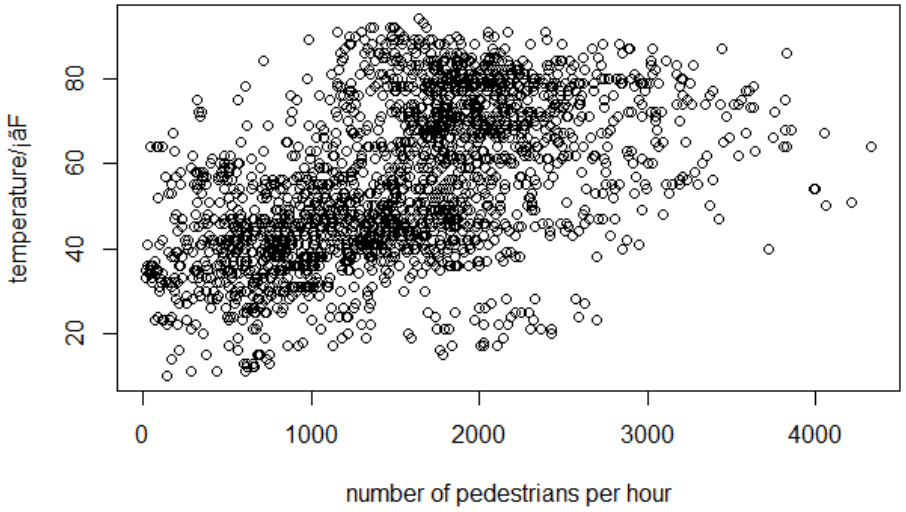

Figure 6 Scatterplot between Pedestrians and temperature

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
mu1	2000.52	0.39	28.82	1943.85	1980.69	2000.51	2019.86	2058.93	5468	1
sigma1	507.32	0.26	20.06	470.60	493.61	506.40	520.87	548.30	5924	1
mu2	1811.55	0.32	24.28	1765.57	1794.67	1811.65	1828.55	1858.10	5908	1
sigma2	749.26	0.25	17.59	714.55	737.23	748.94	760.77	784.69	4990	1
mu3	1127.16	0.26	20.17	1087.20	1113.48	1127.30	1140.98	1167.06	5915	1
sigma3	622.94	0.18	13.60	597.37	613.40	622.66	631.78	650.69	5615	1
lp__	-15673.45	0.04	1.72	-15677.75	-15674.37	-15673.15	-15672.17	-15671.06	1927	1

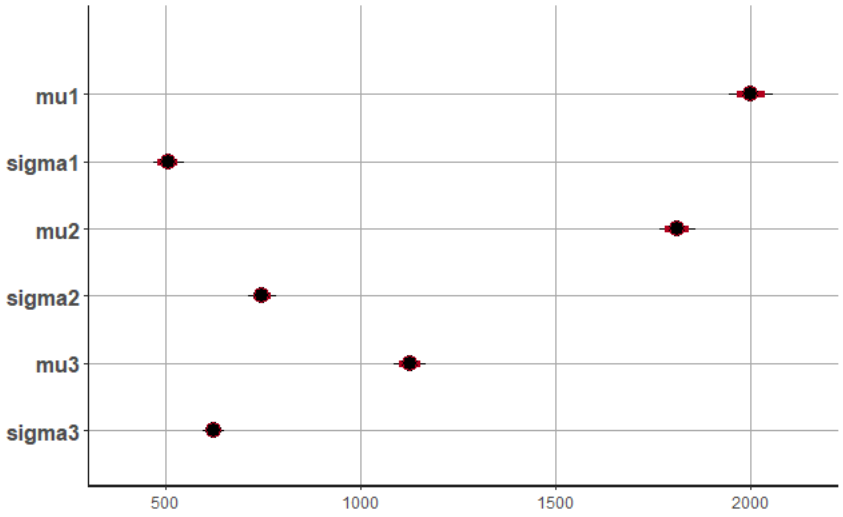

Figure 7 Parameters of normal likelihood function (temperature)

	mean	se_mean	sd	2.5%	25%	50%
lambda1	2000.53	0.05	2.46	1995.69	1998.87	2000.55
lambda2	1811.73	0.02	1.37	1809.07	1810.80	1811.76
lambda3	1127.15	0.02	1.08	1125.03	1126.43	1127.14
lp__	22057277.49	0.03	1.23	22057274.26	22057276.92	22057277.79

	75%	97.5%	n_eff	Rhat
lambda1	2002.22	2005.33	2774	1
lambda2	1812.63	1814.45	5238	1
lambda3	1127.88	1129.22	3018	1
lp__	22057278.40	22057278.88	1895	1

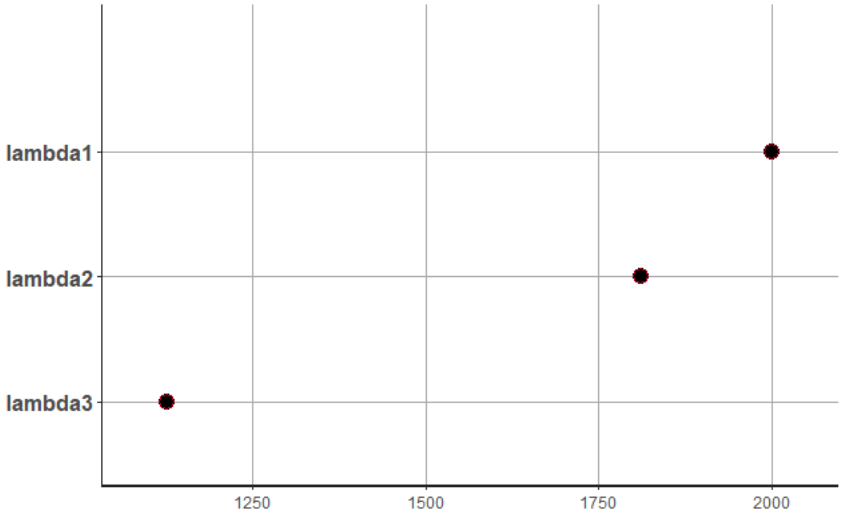

Figure 8 Parameters of Poisson likelihood function (temperature)

When considering temperature as factor, we can see that the estimations we got from normal distribution and Poisson distribution are almost the same. In general, we can see that the higher the temperature, the more pedestrians going through the bridge. For high temperature and medium temperature, though there is a difference, the estimations are close to each other. However, for low temperature, the estimation is much smaller. Thus, we can see that with low temperature, there will be few pedestrians going through the bridge.

Simulations for weather:

Because the number of pedestrians going through the bridge should not be negative, we used truncated normal distribution to run simulations.

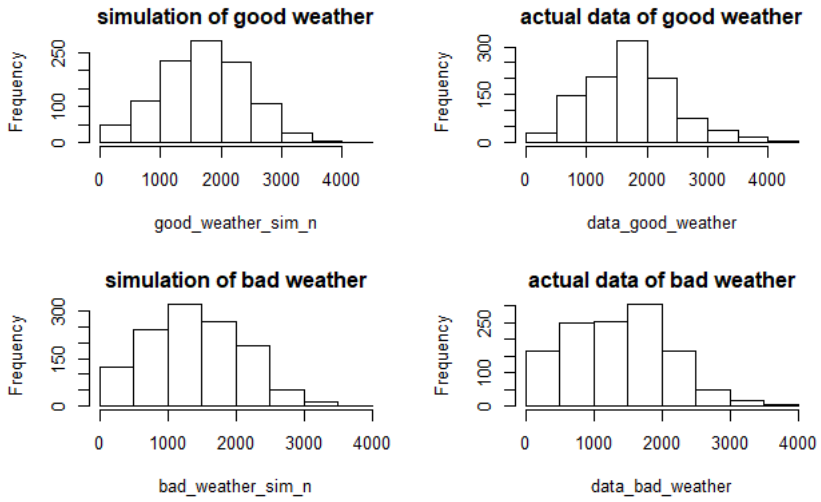

Figure 9 Simulations of weather using normal likelihood function

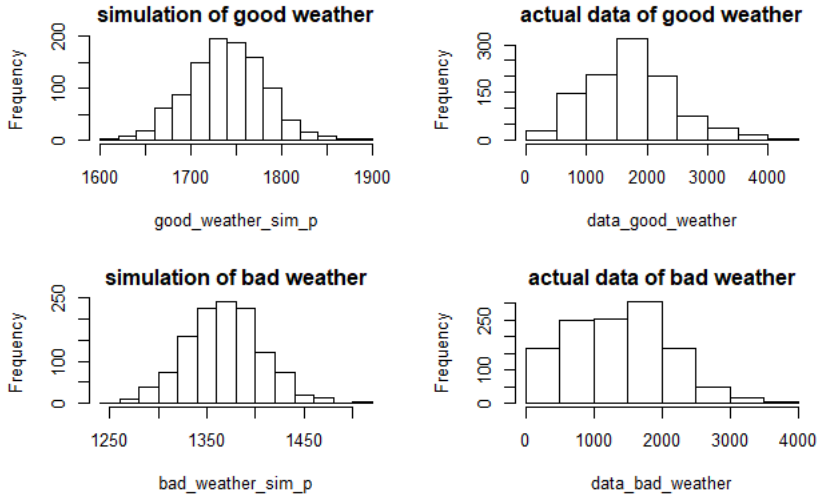

Figure 10 Simulations of weather using Poisson likelihood function

From the histograms above, we can see that when using normal likelihood function, the histograms of simulations are pretty similar with what we have for real data. However, when using Poisson likelihood function, the histograms are much more concentrated. Thus, we can conclude that Poisson distribution is not a very good model for this analysis. Normal distribution is a good choice for estimating the number of pedestrians going through the bridge with different weathers.

Simulations for temperature:

With the same reason above, we used truncated normal distribution for simulations instead of regular normal distribution.

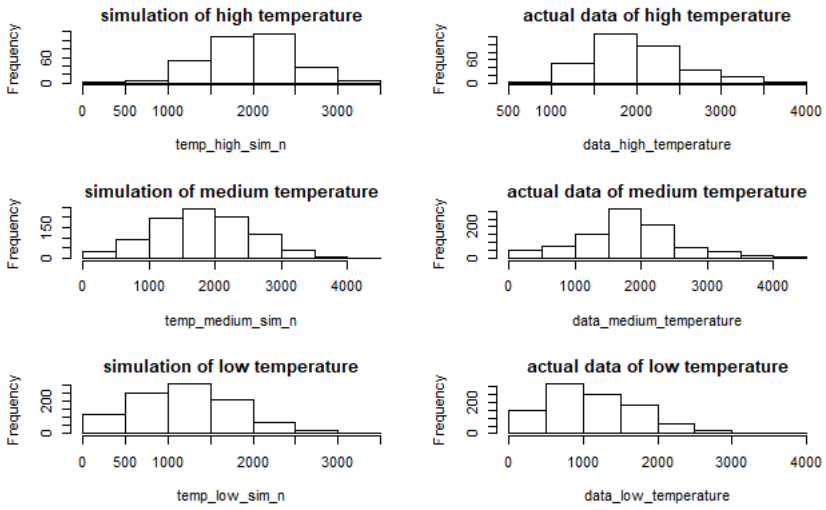

Figure 11 Simulations of temperature using normal likelihood function

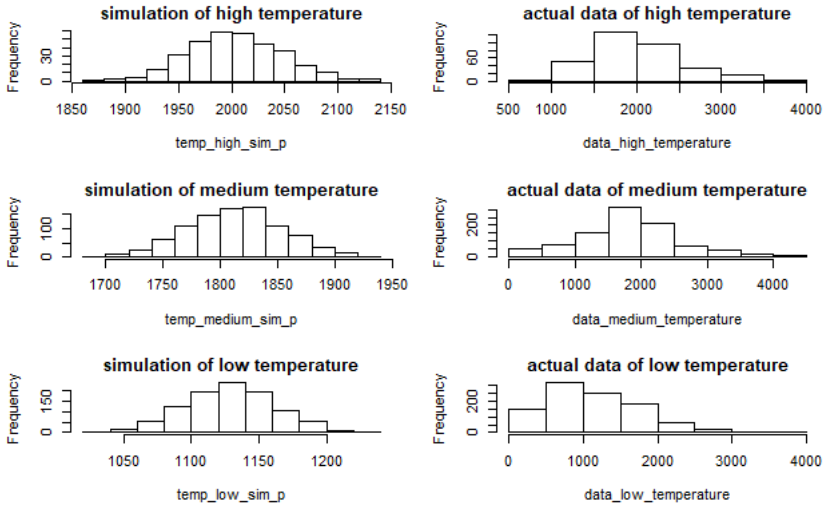

Figure 12 Simulations of temperature using Poisson likelihood function

For temperature, we can see that the simulations are still better when using normal distribution than using Poisson distribution. However, even when using normal distribution, the simulations are not as good as what we have for weather. A possible answer to the difference should be that for temperature, there are not as many observations in each group as for weather. Because when doing simulations, we chose the same numbers as we had in our observations for each group. We had three groups of temperature instead of two groups of weather. Hence, the simulations for temperature did not perform as good as for weather.

Conclusions:

- In general, more people go through the bridge when the weather is good. About 25% more people go through the bridge when the weather is good.
- Low temperature indeed will reduce the number of pedestrians going through the bridge. About 40% fewer people go through the bridge when it is cold outside.
- The simulations we got from Bayesian analysis are accurate, but we should have enough numbers of sampling.

Improvement:

- We only have two factors here. Other factors, like time of the day and events, can also have some influences.
- Perhaps we can run linear regression about the number of pedestrians with all other factors and compare the coefficients we get with what we have from Bayesian analysis.

Appendix:

Reference:

1. The data is from https://data.cityofnewyork.us/Transportation/Brooklyn-Bridge-Automated-Pedestrian-Counts-Demons/6fi9-q3ta/data.

Contribution of teammates:

Junfeng Luo: Cleaning data, selecting models, planting in Stan package, studying the influence of weather, data visualization

Dong Ding: Searching data, making plans, studying the influence of temperature, making simulations, writing report