Bayesian Analysis Project Report

Group D 2021H1540840P 2021H1540843P 2021H1540833P 2019A5PS1075P

Problem Statement : The airbnb_small data in the bayes rules package contains information on AirBnB rentals in Chicago. This data was originally collated by Trinh and Ameri (2016) and distributed by Legler and Roback (2021). In this open-ended exercise, build, interpret, and evaluate a model of the number of reviews an AirBnB property has with respect to its rating, district, room_type, and the number of guests it accommodates.

Approach: We first checked the data for missing data, business abnormalities. Then we selected priors (default). After that, we built the model using a sequential approach. After the model building, We observed the Markov Chain diagnostics followed by the posterior analysis. Then we checked for the accuracy of the predictability of the model.

Observations:

Categorical Variables : Room_type(I), district (I)
Continuous variables : Rating(I) , Accommodates(I) , Reviews (Dependant)

```
     district
Far North:336
North    :484
Northwest: 49
```

```
     rating
district
 Min.   :2.500
North:336
 1st Qu.:4.500
:484
 Median :5.000
49
 Mean   :4.801
 3rd Qu.:5.000
 Max.   :5.000
```

```
          room_type
Entire home/apt:467
Private room   :370
Shared room    : 32
```

```
 accommodates
Min.   : 1.000
1st Qu.: 2.000
Median : 2.000
Mean   : 3.522
3rd Qu.: 4.000
Max.   :16.000
```

Every value of the variables present have business sense. There are no spurious values as of our knowledge.

Model Building :

Test train split, Prior :

- We have built a posterior based on Prior 1 (the default prior).
- This posterior is then taken as Prior 2 and the models are simulated.
- We have used the sequential analysis approach. We have splitted the data into 3 sets in the ratio of 1:1:3.
- Two equal parts are used for prior and test, whereas the largest part is for training purposes.

Distributions:

- Dependant variable - Negative binomial distribution
- Prior - Gaussian

Model : (depicting sequential approach)

```
a1 <- stan_glm(reviews ~ rating + accommodates+ district+room_type,
               data = start,
               family = neg_binomial_2,
               prior_intercept = normal(2, 0.5,autoscale=TRUE),
               prior = normal(0, 2.5, autoscale = TRUE),
               prior_aux = exponential(1, autoscale = TRUE),
               chains = 4, iter = 5000*2, seed = 84735,
               prior_PD = TRUE)
```

```
a2 <- stan_glm(reviews ~ rating + accommodates+ district+room_type,
                                  data = train,
                                  family = neg_binomial_2,
                                  prior_intercept = normal(t_df$estimate[1], t_df$std.error[1]),
                                  prior=normal(c(t_df$estimate[2],
                                                 t_df$estimate[3],
                                                 t_df$estimate[4],
                                                 t_df$estimate[5],
                                                 t_df$estimate[6],
                                                 t_df$estimate[7]),
                                                 c(t_df$std.error[2],
                                                   t_df$std.error[3],
                                                   t_df$std.error[4],
                                                   t_df$std.error[5],
                                                   t_df$std.error[6],
                                                   t_df$std.error[7])),
                                  prior_aux = exponential(1, autoscale = TRUE),
                                  chains = 4, iter = 5000*2, seed = 84735)
```

Results :

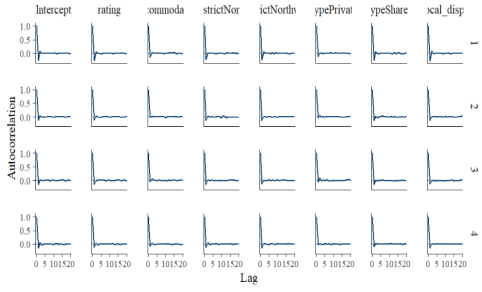
A tibble: 7 x 5

term	estimate	std.error
<chr>	<dbl>	<dbl>
(Intercept)	1.63531882	0.61562896
rating	0.27527951	0.12539785
accommodates	0.05024106	0.01834141
districtNorth	0.17497808	0.07273202
districtNorthwest	-0.20542806	0.16213498
room_typePrivate room	0.20500735	0.08320858
room_typeShared room	-0.44511758	0.19332951

7 rows

Important observations(MCMC Diagnostics):

- **Negative autocorrelation :** causing Neff to be greater than 1 (property of antithetic markov chain - needs further conditioning of model)




```
      (Intercept)            rating      accommodates      districtNorth
         1.318767          1.287467          1.015733          1.185100
districtNorthwest room_typePrivate room  room_typeShared room reciprocal_dispersion
         1.203600          0.998600          1.294633          1.315967
```

Posterior predictability measures :

Estimate **SE**

mae	mae_scaled	within_50	within_95
<dbl>	<dbl>	<dbl>	<dbl>
17.51518	0.6923427	0.5017561	0.9597142

elpd_loo	-2206.96322	28.422147
p_loo	7.91684	1.297097
looic	4413.92644	56.844294

Conclusion :

Based on ELPD and pp_check , we can infer that our model is good.Our assumption of dependent variable distribution to be negative binomial is proven to be correct. However, based on the MAE values ,we can infer that there is a lot of scope for improvement in the model. We intend to try with other family of distributions for priors, also explore other packages for even better predictability.