STA723 Case Study 2 - Group 5 Report

Youngsoo Baek, Phuc Nguyen, Irene Ji

Executive Summary

1. Introduction

2. Materials and Methods

To understand which factors influence the price and popularity of a listing, we fit a bivariate response linear regression that has varying intercept across different neighbourhoods and boroughs. For the i-th listing in neighbourhood j, within borough k, the model can be written as

$$\begin{pmatrix} \text{Log Price}_{k[j[i]]} \\ \text{Log Monthly reviews}_{k[j[i]]} \end{pmatrix} = \begin{pmatrix} \beta_1^T \mathbf{X}_i \\ \beta_2^T \mathbf{X}_i \end{pmatrix} + \eta_{k[j]} + \theta_j + \epsilon_{k[j[i]]},$$

where $\eta_{k[j]}$ is the 2×1 vector of neighborhood-level random effect, θ_j is the borough-level random effect, and ϵ is an observation error that has $N(0, \sigma^2 \mathbf{I}_2)$ distribution. The random intercepts are assumed to have a bivariate normal distribution centered at zero, so we can estimate the between-response correlation between boroughs and between neighborhoods, within boroughs. The model is fit through a standard MLE procedure implemented in `lme4` package for `R`. The slopes remain fixed across different groups, and the predictors for the two responses remain identical for the two responses.

3. Results

3.1 Exploratary Data Analysis and Preprocessing

...It is unclear what quantity `availability_365` variable is measuring, or how precise a measure it can serve as for whatever quantity inherent to a listing.

3.2 Main Results

Our model does not explicitly account for possible spatial structure within neighborhoods. A useful quantity to model is the semivariogram γ of the response Y, which is defined as

$$\gamma(||\mathbf{h}||) \equiv \frac{1}{2}\mathrm{E}[Y(\mathbf{s}+\mathbf{h}) - Y(\mathbf{s})].$$

Here, $\mathbf{s}$ indexes the listing's location, and $\mathbf{h}$ is the displacement vector. We assume that the variogram only depends on the distance, $d \equiv ||\mathbf{h}||$, and not on the direction. In such setting, a customary, simple nonparametric estimator for semivariogram is (BCG, 2011)

$$\hat{\gamma}(d) = \frac{1}{2|N(d)|} \sum_{(\mathbf{s}_i, \mathbf{s}_j) \in N(d)} [Y(\mathbf{s}_i) - Y(\mathbf{s}_j)]^2,$$

where $N(d)$ consists of all pairs of locations that have a pairwise Euclidean distance d. The semivariogram estimators for each of the residuals are plotted on increasing pairwise distance.

It is clear that for monthly review rates, we observe a negative spatial autocorrelation: closer things have more different response rather than similar. While seemingly counterintuitive, the negative autocorrelation can be explained by strong competition between listings when in close proximity. Within the same neighborhoods, a potential customer is always being sapped away from one listing to another, so the latter has at least one more review. Such effect will be strongly visible when listings are closer, but as distance increases, the dependence will become weak to none, and the variability of the difference between two monthly review rates will stabilize. That we do not see a similar negative autocorrelation in price is informative, as it suggests the hosts' pricing policy remains indifferent to their neighboring hosts, conditional on the neighborhood they belong to. The negative spatial correlation for monthly review rates raises cause for concern in applying standard spatial models like conditional autoregressive (CAR) models, which assumes positive autocorrelation.

4. Discussion

References

Sarthak, N., Post: "Availability_365=0?", Discussion thread: New York City Airbnb Open Data, Kaggle. https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data/discussion/111835

Banrjee, S., Carlin, B. P., and Gelfand, A. E. (2011). *Hierarchical Modeling and Analysis for Spatial Data.*

Appendix: Figures and Tables

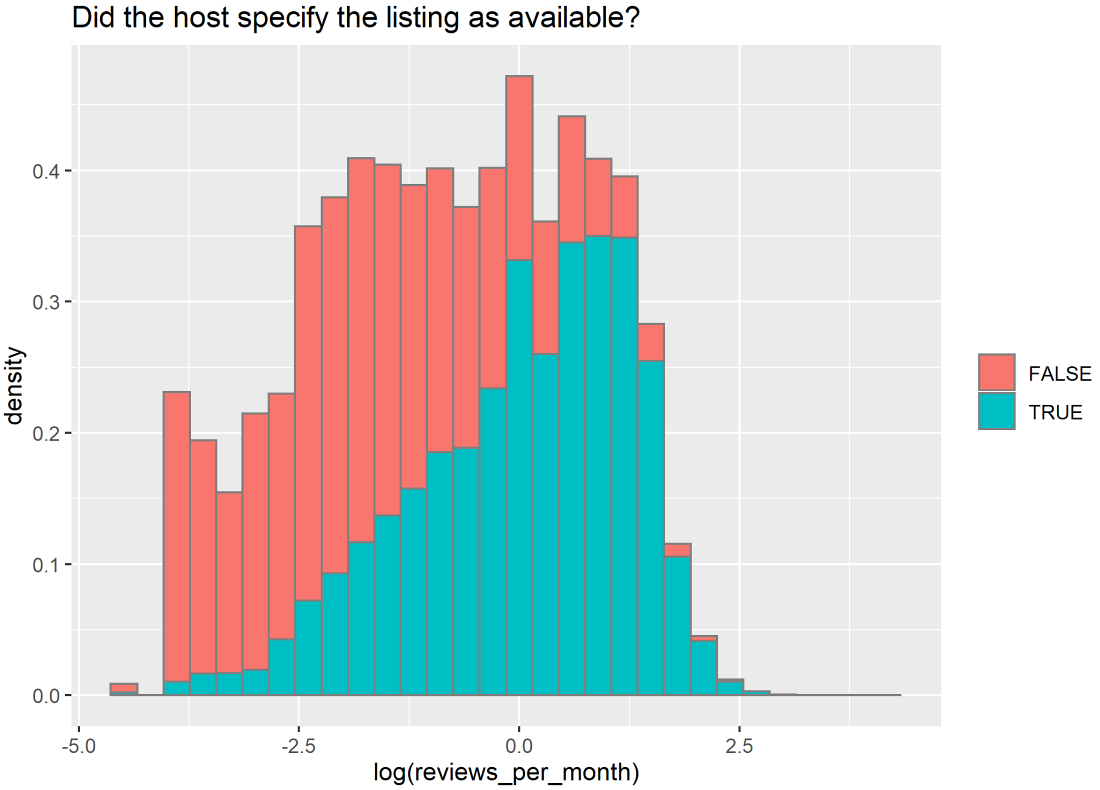

Figure 1: Distribution of log monthly review rates for listings with zero/non-zero availability feature.

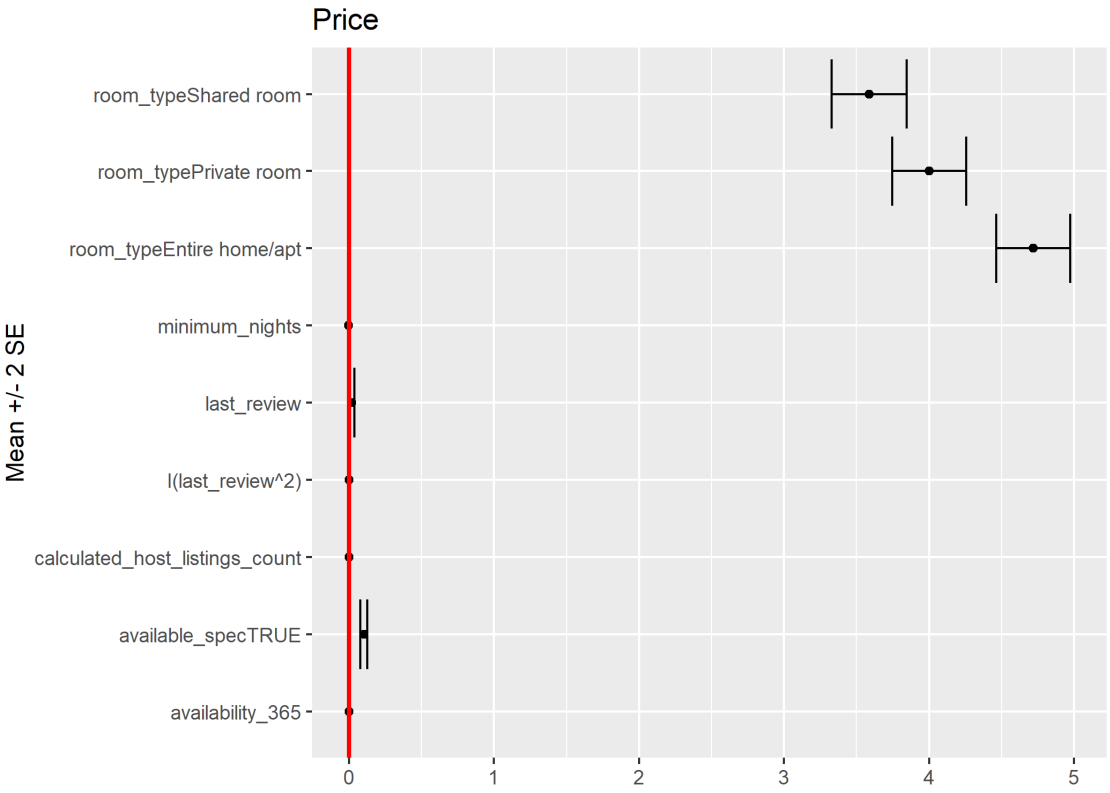

Figure 2: Fixed effect estimates for log price.

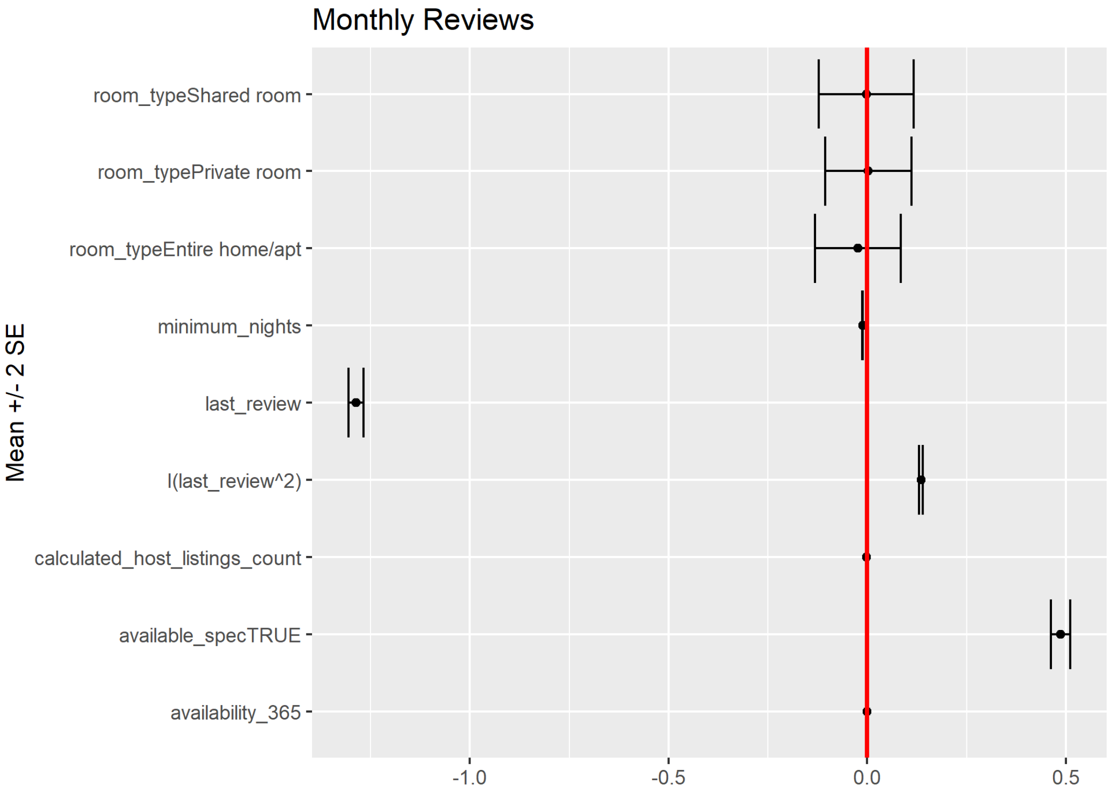

Figure 3: Fixed effect estimates for log monthly review rates.

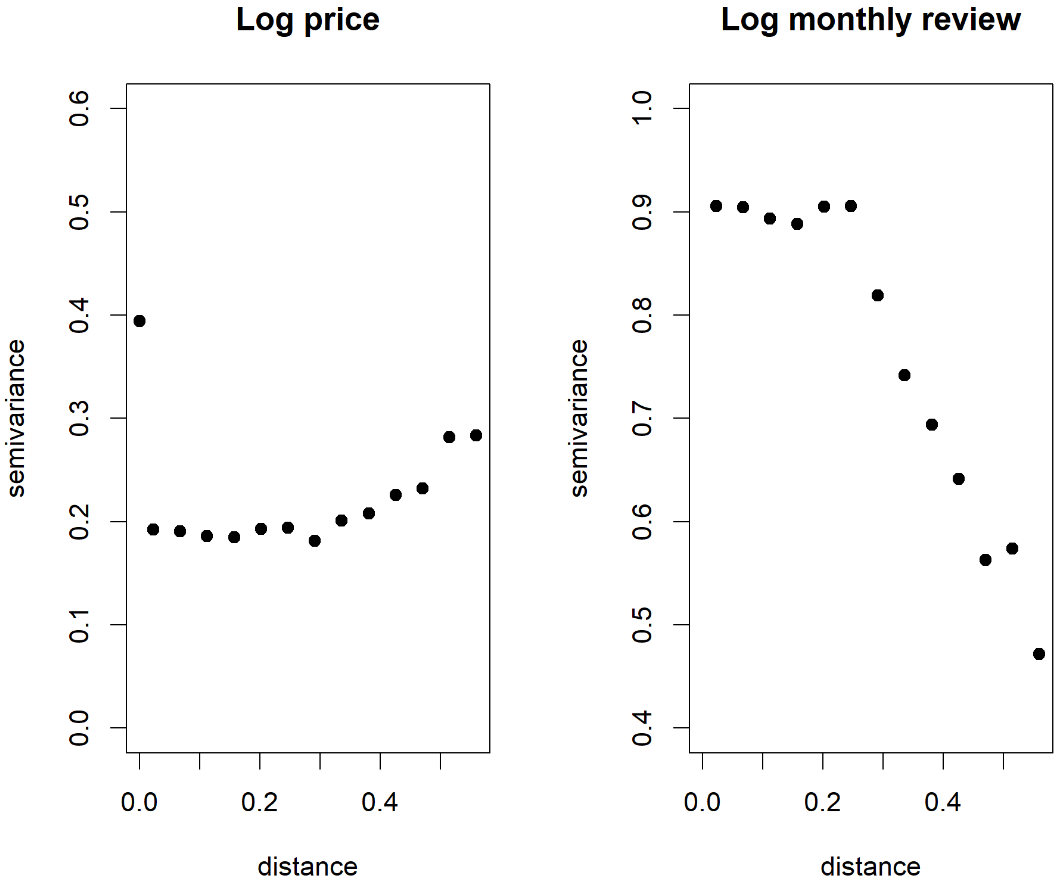

Figure 4: Semivariogram estimators calculated from model residuals for price and monthly review rates.