Monte Carlo Markov Chain (MCMC) and Multilevel Modeling: A Bayesian Approach to *Airbnb* Price Prediction

ST308 Project

Candidate Number: 20840

Contents

1 Introduction

1.1 Motivation

Price, by construction, should be some artificial reflection of the value of some good or service that it represents (Fetter, 1912). Using features of the good or service, price prediction is thus a natural subject to apply statistical techniques on. Successful and accurate models can have substantial business implications in many areas – market planning, demand forecasting, etc. The aim of this article is to predict prices for *Airbnb*[1] listings, specifically in Vancouver, Canada, and to determine how the various features of a listed property can predict its price from a Bayesian perspective. Notably, MCMC simulation will be the primary tool employed to sample and estimate models in this empirical analysis.

1.2 Data

The data was accessed from the public online database of *Inside Airbnb*, where the owner of it scrapped the data from the *Airbnb* website[2] (Cox, 2021). After subseting, it contains 279 listings in Vancouver, and it records information associated with listings' price, location, room type, etc[3]. There are two key aspects of the data that demand some attention. Firstly, the outcome variable, i.e. listing price, is positively skewed (skewness = 1.76). Although this is not detrimental to the forthcoming analysis, for linear models that we will employ, a normally distributed outcome variable is desirable. Hence, for the outcome variable, we will instead use the price log-transformed (skewness = −0.07). Secondly, the data exhibits a hierarchical structure in the observations. To be precise, the data consists of $N = 279$ listings, which are listed on *Airbnb* by a total of $H = 20$ different hosts.

To further understand the data, Fig. 1 presents plots that illustrate the relationship between the log of price and various property-level predictors, where each property is coloured according to its host and can be matched by plot(2, 3). By informal eyeballing, among the categorical predictors, *room type* is a categorical variable that shows difference in the log of price for the two groups; There is less of an obvious relationship among the groups in the other two categorical predictors. For the continuous predictors, the OLS fitted lines suggest that log of price does correlate with *latitude* and *longitude*, while its relationship with the other two is relatively ambiguous. Moreover, the correlation matrix among the continuous predictors show weak correlation between most features. A few relationships that are relatively correlated include the one between *latitude* & *longitude*, and also relationships that are associated to *host listings count*[4].

2 Pooled Linear Regression

2.1 Model Framework

A direct initial model to consider would be a pooled model. In this model framework, we consider all observations of listed properties together having a single distribution in price regardless of their host. Specifically, we will only include information on listings' property-level features, i.e. information that is independent of the host.

As a basis of the analysis, we will formulate the price prediction problem with a linear model framework. Namely, for property i, given the listed price y_i and the vector of the property-level features x_i , we have that:

$$\log y_i = x_i^\top \beta + \varepsilon_i \qquad\qquad \varepsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2) \tag{1}$$

Hence, given $K \in \mathbb{N}$ features, the goal is to accurately, estimate for the vector of parameters, $\beta \in \mathbb{R}^K$, such that we may predict prices via $\log \hat{y}_i = x_i^\top \hat{\beta}$.

In the context of the empirical problem, most of the predictors will be modeled in the conventional way, given no strong beliefs of any peculiarities that need to be captured among them. However, the *latitude* and *longitude* of a property's location do require sophisticated specification to fully resemble the possibly intricate spatial and geographical dynamics. Therefore, we will add 5th order polynomials for those predictors[5]. Moreover, given that there is a substantial number of predictors, we will be proceed with a Bayesian linear regression set-up to combat against over-fitting. This will be analogous to a ridge regression from a frequentist's perspective with regularisation term λ.

[1]An online platform for hosts to list their properties for short-term renting

[2]To the date of writing, the most recent version of the data file was downloaded: The 9th February, 2021 version of *listings.csv* in Vancouver

[3]The raw data provided by the source includes 4340 listings and a random subset in choosing 20 different hosts was necessary to avoid excessive computational burden

[4]The scatter plot between *host listings count* and log of price is omitted as the former is a host-level feature, thus not directly related to the latter

[5]Interaction term between *latitude* and *longitude* is arguable important in capturing the underlying relationship. The reason why it is excluded is explained in 4.3

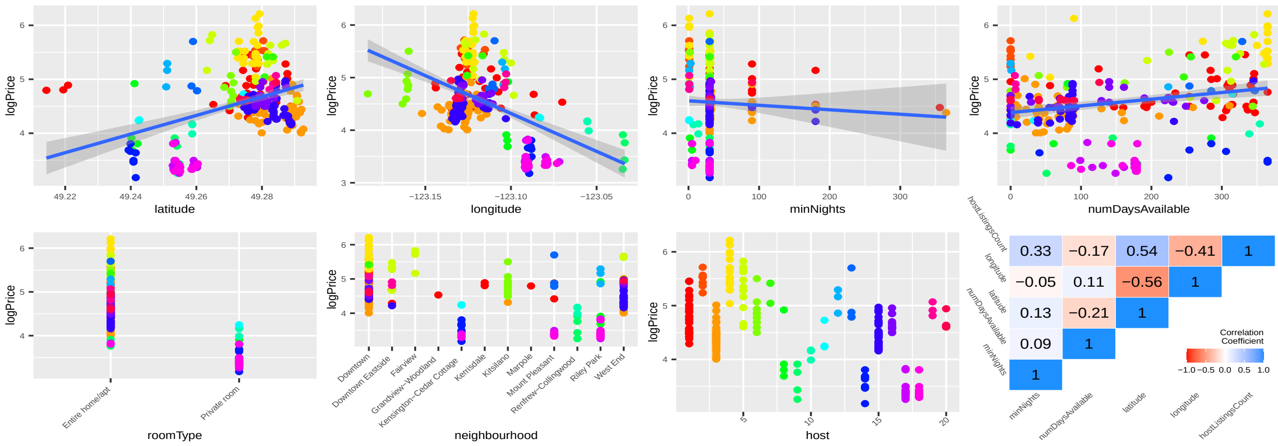

Figure 1: Scatter Plots between Log of Price & Property-Level Predictors and Correlation Matrix Among Continuous Predictors

2.2 Prior and Hyper-prior Distributions

In order to perform Bayesian analysis, it is critical to specify the priors for the parameters of interest. We would require prior distributions for β and σ to fulfill the linear regression framework. However, not being an expert in the Vancouver real estate market, the only relevant information available is the data. Given the absence of strong evidence present regarding the relationship among the various car features and its price, we will proceed with the analysis by considering weakly-informative and non-informative prior distributions. Also, a prior for the additional hyper-parameter λ, which serves to be the strength of regularisation, is necessary under a full Bayesian implementation. We will choose a Half-Cauchy distribution for it due it being vague – it has fat-tails – and that regularisation will not be too extreme – density decreases towards $+\infty$.

2.3 Results

We ran the model with priors specified under Model A in table 1. After 3000 iterations (2000 warm-up), the MCMC sampling implemented through *rstan*[6] found no divergence in during sampling. Chains mixed well with $\hat{R}$ being very close to 1 for all the parameters. Examples of trace plots that are well mixed are shown in Fig. 2.

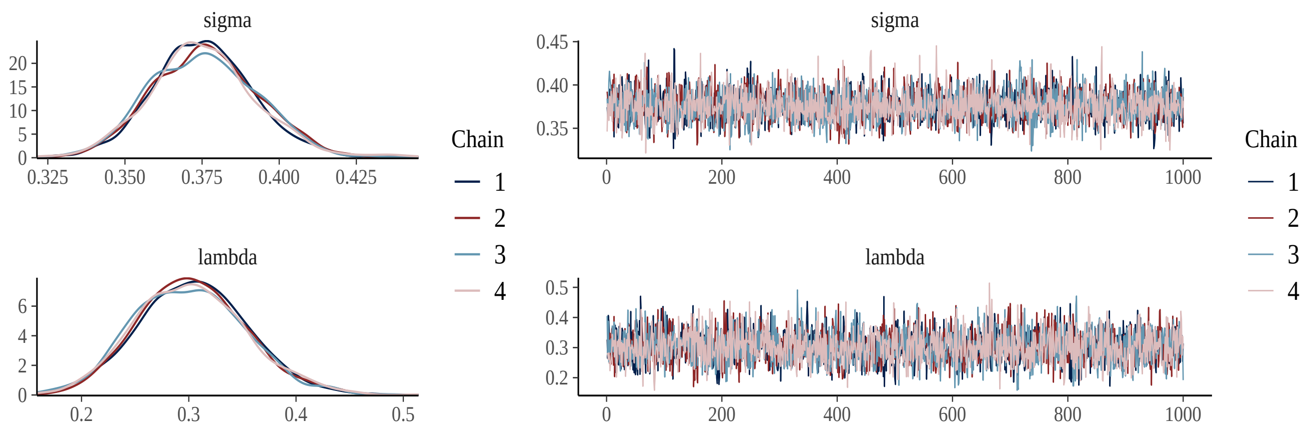

Figure 2: Selection of MCMC Sampled Density & Trace Plots (Model A)

Using the MCMC result, we can apply it to our core price prediction problem. By generating fitted values for the log of price from the posterior draws of the parameters and the given features, Fig. 3 compares it to the true log of price in the data. The sampled values y_{rep} gives us a vague geometric sketch of the posterior predictive distribution for our problem.

The left plot in Fig. 4 shows the Bayes estimate of the all the parameters in the form of posterior means. Since all the (continuous) predictors are standardised before the sampling, it is possible to identify a few predictors that are impactful in predicting the log of price of a property: the *room type*, *longitude*, *latitude*, selected neighborhoods, etc. Features like *room type* is easy to interpret as its (95%) credible interval does not contain 0, i.e. a private room is less expensive than a full apartment/home. However, conclusions on predictors like *longitude* and *latitude* are less definite due to their complicated polynomial structure and heavy-tailed sampled posterior marginal distribution. These findings are generally consistent with the elementary observations made when we interpreted Fig. 1.

[6] Apart from *rstan*, the list of R non-base packages used throughout the analysis includes: *dplyr*, *ggplot2*, *reshape2*, *gridExtra*, *moments*, *bayesplot*, *loo*

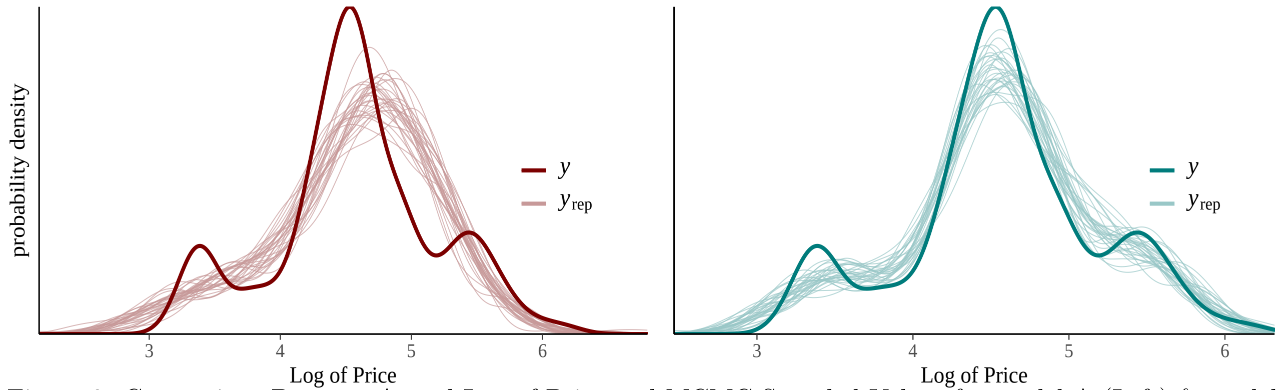

Figure 3: Comparison Between Actual Log of Price and MCMC Sampled Values for model A (Left) & model D (Right)

2.4 Sensitivity Analysis

The results that were obtained are dependent on the priors that were chosen. To an extent, they were arbitrary choices. Without any strong beliefs, we cannot plausibly evaluate other priors that are more informative. Hence, we consider two other specifications with vague priors of the pooled model. To avoid confusion, let the previous specification be model A, and the new specifications be model B and C. Their priors can be summarised in table 1.

	Model A	Model B	Model C
Prior:	$\beta \sim \mathcal{N}(0, \sigma^2\lambda^{-2}I_K)$	$\beta \sim \mathcal{N}(0, \sigma^2\lambda^{-2}I_K)$	$\beta \sim \mathcal{N}(0, \sigma^2 N(X^\top X)^{-1})$
Hyper-Prior:	$\sigma^2 \sim IG(0.01, 0.01)$ $\lambda \sim$Half-Cauchy$(0, 1)$	$\sigma^2 \sim IG(0.001, 0.001)$ $\lambda \sim$Half-Cauchy$(0, 2)$	$\sigma^2 \sim IG(0.01, 0.01)$
LOO-ELPD Difference (Std. Error):	0	-0.1 (0.2) **relative to A*	-11.9 (4.7) **relative to A*

Table 1: Comparison Between Pooled Model with Different Priors

Model B uses priors that are even less informative than the priors in model A. Again from Fig. 4, the comparison between the specifications shows virtually no difference in the Bayes estimates. Crucially, after taking into account of their credible intervals, the posterior means of lambda between the two specifications remain similar. This provides some confidence to the degree of regularisation applied to the parameters. On the other hand, Model C uses the unit information prior for β, and does not involve the hyper-parameter λ. Although having a normally distributed prior for β centered at 0 already enforces some regularisation, it is no surprise that Fig. 4 shows that the magnitude of the posterior estimates are much larger. Despite this difference, we still largely see consistency in the direction of the posterior means.

All in all, if we compare the various model specifications using Leave-One-Out Cross Validation (LOOCV), we find that the LOO estimate of the expected log pointwise predictive density (ELPD) is the greatest for Model A. This is shown in the comparison done in table 1. In layman's terms, model A is more accurate compared to the rest when predicting 'unseen' property listings[7], but is only marginally better than model B. Its superiority over model C is expected since over-fitting may be an issue for the latter model.

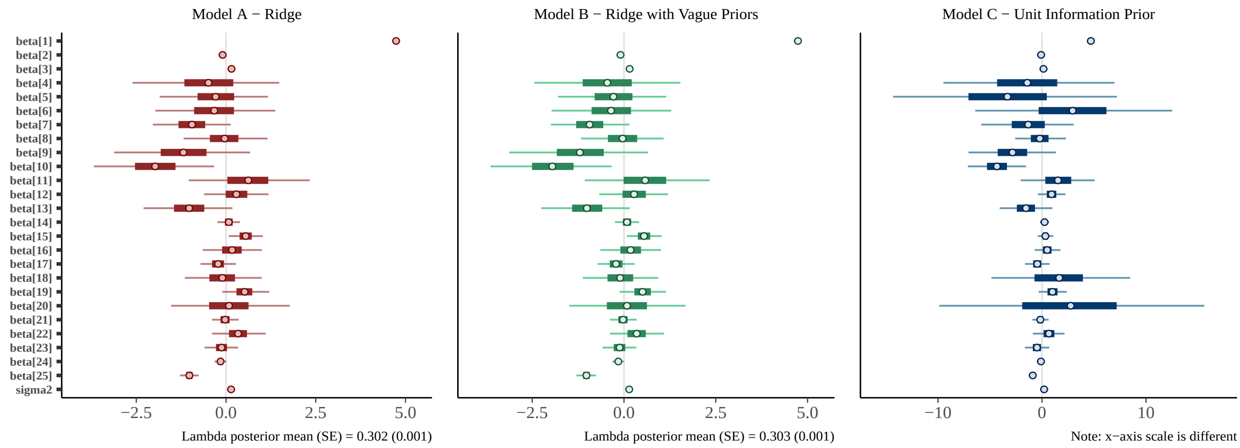

Figure 4: Comparison of Posterior Mean Among Pooled Models with Different Priors (Dark Intervals = 50% Credible Interval & Light Intervals = 95% Credible Interval)

[7]LOOCV essentially isolates one of the observations to treat as a singleton testing set and the rest as the training set. Prediction error is then aggregated by repeating this for all observations (Gelman, Hwang, & Vehtari, 2013)

3 Multilevel Analysis

3.1 Model Framework

Recall from 1.2 that the each observation in the data has multiple groups in 2 levels – every property is listed by a single host. This motivates a model that fully exploits such hierarchical structure of the data. Such a model can be constructed by augmenting the core linear regression model of (1) in the following way: For every property i listed by host j, we have,

$$\log y_{ij} = x_{ij}^\top \beta_j + \varepsilon_{ij} \qquad \varepsilon_{ij} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2) \tag{2}$$

$$\beta_j^{(k)} = z_j^\top \mu_k + \eta_k \qquad \eta_k \overset{i.i.d.}{\sim} \mathcal{N}(0, \omega_k^2) \tag{3}$$

Here, (2) and (3) represent the relationship in the property-level and the host-level, respectively. Precisely, we represent the features in the former level by x_{ij} and the features in the latter level by z_j, for all property i and host j. The main distinction between this framework and (1) is that the coefficient parameters on the property-level features are now different depending on whom their host is.

In the context of our problem, the property-level part of the model will be identical as it was described in 2.1, apart from the distinction that the parameters have different priors depending on its host. For the host-level, i.e. z_i, we will begin by only considering the intercept term. This specification will be named as model D. Additionally, the observation that we made in 1.2 regarding that *host listings count* correlated moderately with many other continuous property-level features. To explore whether those correlations are spurious or not, we will inspect another model, model E, with the predictor *host listings count* for each host. In other words, this alternative model specification aims see if this host-level feature can help explain part of the between-host randomness or not. Mathematically, for all k, in model D, we have $\mu_k = \mu_{1k}$, and in model E, $\mu_k = (\mu_{1k}\ \mu_{2k})^\top$.

3.2 Prior and Hyper-prior Distributions

The prior for $\{\beta_j^{(k)}\}$ is highlighted by the construction of (3). Similar to the argument made in 2.2, the lack of external information about the context and any of the parameters prevents us from adopting any priors carry substantial information. Therefore, for the hyper-parameters $\{\mu_k\}$, which are the parameters on the host-level predictors, we will stick with vague normal distribution priors – normal distributions centered at 0 offer some shrinkage power to avoid excessive over-fitting. For the remaining hyper/parameters, i.e. σ, $\{\omega_k\}$, they can be classified as scale hyper/parameters in some part of the entire model. Under a multilevel model setting, it has been suggested that a good candidate for non-informative or weakly-informative hyper/priors is the Half-Cauchy distribution[8] (Gelman, 2006). Hence, we will follow such advice in our analysis.

3.3 Results

The model that was ran (call model D) imposed the hyper/prior for $\sigma \sim$ Half-Cauchy $(0, 1)$, and for all k, $\omega_k \sim$ Half-Cauchy $(0, 1)$, $\mu_{1k} \sim \mathcal{N}(0, 100)$. From running the MCMC simulation for the same number of iterations as we did in 2.3 through *Rstan*, we did not yield results that converged in every iteration. 15 iterations out of the 4000 post-warmup iterations diverged. Nonetheless, the chains are mixed and examples are shown in Fig. 5 for some of the parameters. Similar to what we did for the pooled model, we can also consider the rough geometric sampling of the predictive distribution for the log of price in Fig. 3

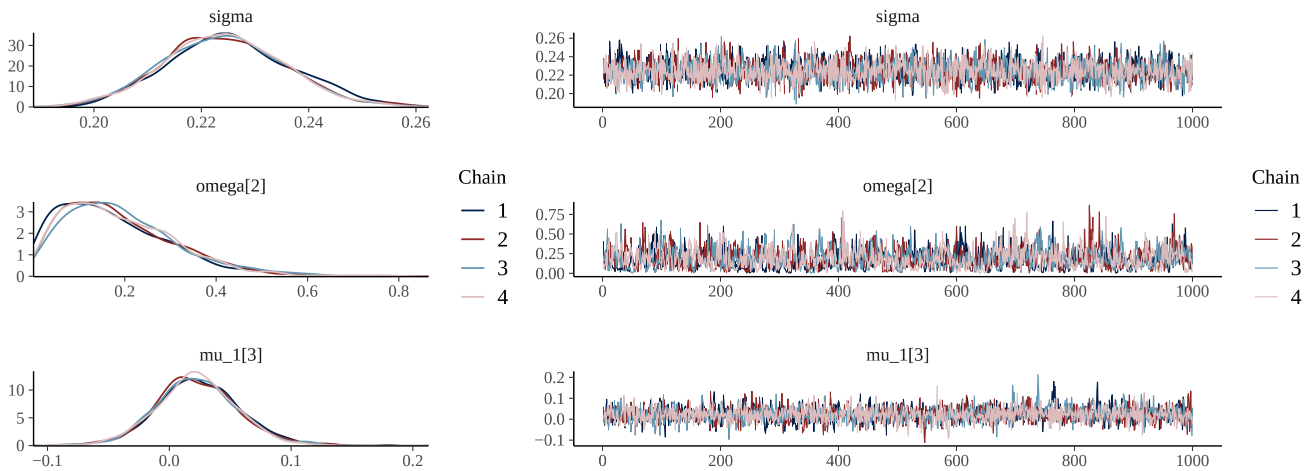

Figure 5: Selection of MCMC Sampled Density & Trace Plots (Model D)

[8]Despite potentially achieving conjugacy, Inverse Gamma does not qualify as a weakly/non-informative prior from Gelman's view, as the posterior of scale parameters depend heavily on the arbitrary choice of the distribution parameters

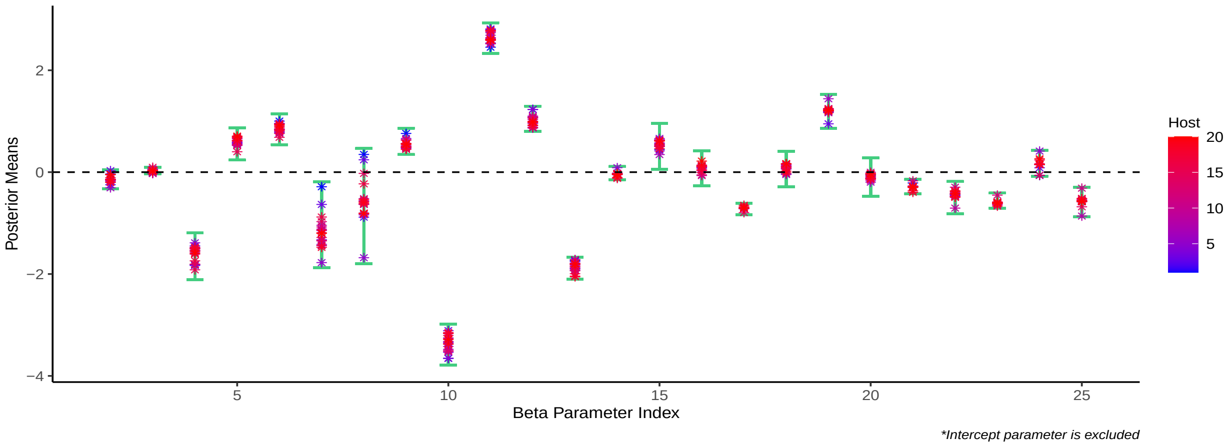

Figure 6: Posterior Means of β Parameters for Every Host

In terms of identifying which predictors carry the strongest predictive power towards the log of price, it is no longer a simple task to interpret due to the additional grouping of parameter distributions in the host-level. A crude approach is to consider only the posterior means of the $\{\beta_j^{(k)}\}$ parameters for all hosts. A scatter plot of such is displayed in Fig. 6[9]. The interval for each parameter using the maximum of the standard error the 20 posterior means, and extending it through the maximum posterior mean and minimum mean[10]. It gives a conservative representation to the posterior means. Among all the parameters, we note that it is the similar set of predictors that influence price the most as the set mentioned in 2.3, i.e. location-related predictors, *room type* is private room dummy, etc.

Despite making such observations and claims, we cannot conclude definitively about them. Notably, the two plots on the left of Fig. 7 show all posterior means for two randomly chosen host. It is clear that many of the posterior distributions have massive variability, and this is a common theme among the results for all hosts. Indeed it is true that the model can internalise between-host variation and reflect it in the results, as it is evident in the right plot of the figure, where the densities for the posteriors of the intercept term is not aligned among hosts. However, the fact that some hosts only own as little as 3 properties may cause limited within-host variation, and subsequently drive the degree of uncertainty and standard errors higher.

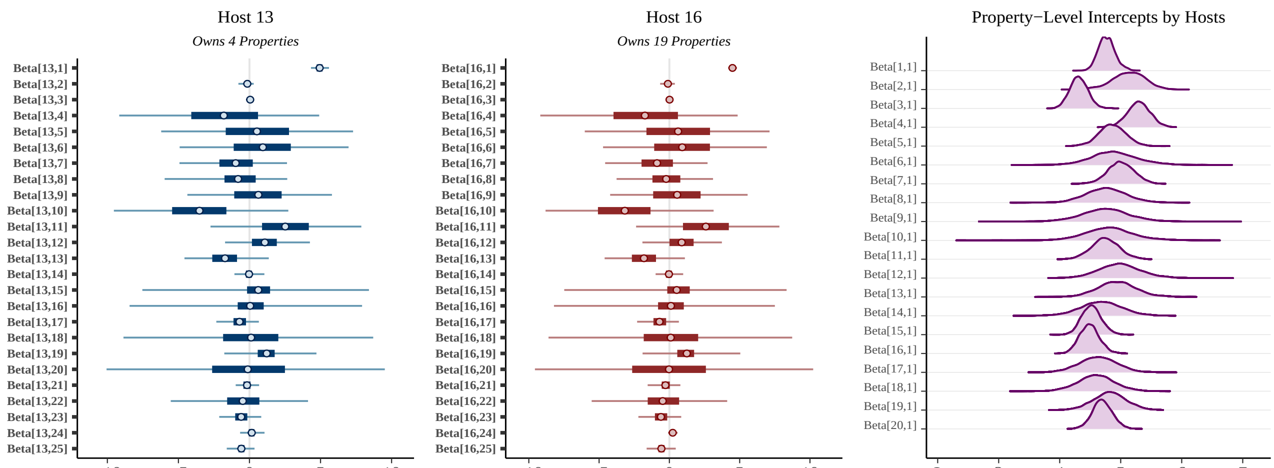

Figure 7: Property-Level Coefficient Parameters for 2 Random Hosts & Densities of Property-Level Intercepts

3.4 Alternative Host-Level Specification

As we have introduced model E[11] in 3.1, we hope to compare it to model D, where we had essentially set $\mu_2 = 0$, to see whether it is statistically justified to include the *host listings count* as a predictor on the host-level. To reiterate, we want to see whether *host listings count* can capture part of the variation in the randomness in the coefficient parameters among the groups.

[9] intercept parameter is exclude as it is represented later in Fig. 7

[10] Given the max and min posterior means $\bar{\beta}_{max}$, $\bar{\beta}_{min}$, we extend the interval to $[\bar{\beta}_{min} - 1.96 s_{max}, \bar{\beta}_{max} - 1.96 s_{max}]$, we create an interval that is continuous over the posterior means, resembling an extended 95% interval

[11] We used the same hyper/priors, and for μ_{2k} we chose hyper-prior $\mu_{2k} \sim \mathcal{N}(0, 100)$, for all k

By LOOCV[12], we find that model D and E have 0 difference, which has a standard error of 3.2. From a frequentist's perspective, we cannot reject the possibility that the difference is 0, under 5% significance level. This is evidence in favour of the claim that *host listings count* does not provide greater predictive power to the model and model D is not mis-specified. Since a simpler model is prioritised when selecting between two models that perform similar in predicting during training, we can proceed with model D.

3.5 Sensitivity Analysis

Again we want to consider how influential the choice of hyper/priors are on the results. Under the same logic as reasoned in 2.4, we will only consider the alternative which has priors that are even less informative. Here, we will call this new model, model F. In this model, the prior for $\{\beta_j^{(k)}\}$ remains the same as it was constructed in (3). The other hyper/priors for the rest of the hyper/parameters are: $\sigma \sim$ Half-Cauchy $(0, 5)$, and for all k, $\omega_k \sim$ Half-Cauchy $(0, 5)$, $\mu_{1k} \sim \mathcal{N}(0, 400)$, $\mu_{2k} \sim \mathcal{N}(0, 400)$.

LOOCV tells us that model D has higher ELPD than model F does, by a difference of 2.4. Notably, the standard error for this difference is 3.1. Hence, from a frequentist's perspective, we cannot reject the possibility that the difference is 0, under 5% significance level. Nevertheless, this gives us some confidence that the choice of hyper/priors is not overly pivotal.

4 Model Comparison

4.1 Prediction for Seen Data

In this section we want to focus on comparing between model A and model D. Informally, we can compare the plots in Fig. 3. Although this does not reveal much about the model's tendency to over-fit, model D does produce a shape for the predictive distribution that resembles the probability density of the data better than model A does. To be better compare the two, we should consider the host-level information of the property listings. Incorporating such information into the model is what differentiates a multilevel model from a pooled model. The plot on the top of Fig. 8 compares how well the two models fit for properties that are listed by 4 selected hosts. For the hosts that own the most properties, we can see that model D does significantly better than its counterpart at predicting against seen data. However, the two models perform similarly poor when considering hosts that own small number of properties; Either models cannot predict reliably for hosts with limited within-host variation because of small observation size.

We can further quantify the prediction error against the data via histogram on the bottom. The histograms show that for the hosts 3 & 4, the predictive errors associated to model D are much more centered at 0 than the predictive errors associated to model A are. In other words, the prediction made by model D generates more error that is negligible, thus fits better to the seen data.

4.2 Prediction for Unseen Data

The comparison made in 4.1 is limited to assessing the models' ability to predict data that they were developed with. Hence, we still have to compare how well the models can predict prices accurately when considering new observations, i.e. validation. LOOCV is one of the methods that can aid us on this matter. Fig. 9 compares between the true log of price and the predicted values by LOOCV for 50 randomly chosen observations[13]. Based on visually interpreting the figure, we cannot tentatively discern notable difference in the predictive power between the two models – both models can predict accurately sometimes but fail at times. The only vague observation that we can make is that the standard errors of the predicted values in model D are considerably smaller than they are in model A. Indeed, if the predictive performance of the two models is similar but the precision of the model D predictions is better, model D would also be preferred.

Nonetheless, we can rigourously evaluate the models through considering numerical results of LOOCV and Widely Applicable Information Criterion (WAIC)[14]. In table 2, we compare the overall LOOCV and WAIC results for the two models. The ELPD differences are statistically different from 0, with 5% sig. level. This is evidence that model D has superior predictive power for unseen observations that are not excessively dissimilar to the observations used in the training the model[15]

[12]It is possible to evaluate these two model specifications from a hypothesis testing perspective where we assess the Bayes factor w.r.t. them through packages like *bridgesampling*. However, it requires substantial samples and massive computation power to calculate the marginal likelihood for complex models.

[13]The predicted values for each observation are essentially predicted values by a model that was generated by data minus such observation

[14]WAIC effectively estimates for ELPD by computing the log pointwise predictive density, and penalising it by the degree of model complexity. In other words, it captures potential over-fitting if too many predictors are included in a model (Gelman et al., 2013)

[15]The models are limited to extrapolating only for observations that have categorical features that are common with the categorical groups seen in this analysis, e.g. we can only extrapolate for a portion of Vancouver *Airbnb* listings

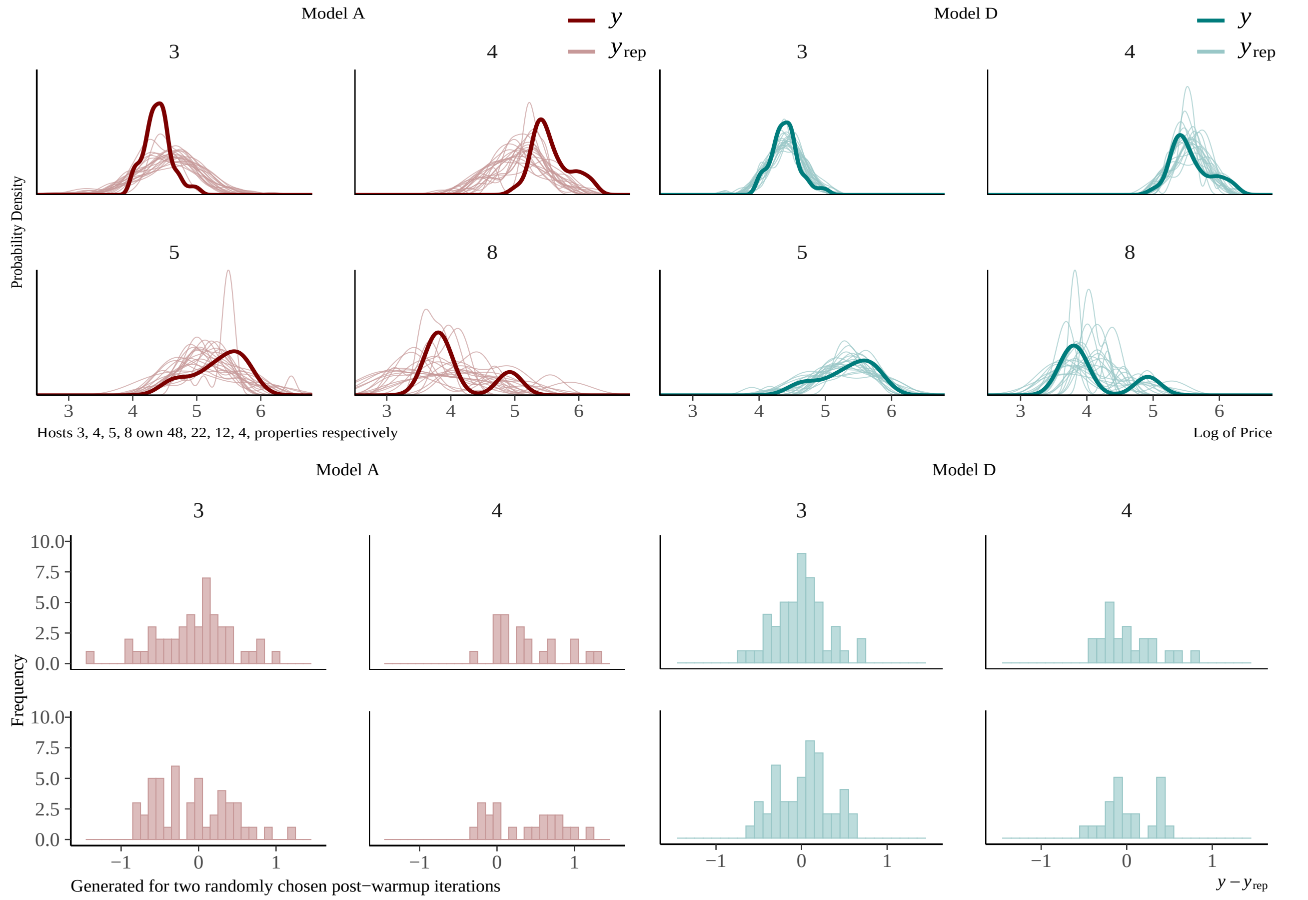

Figure 8: Comparison Between Actual Log of Price & MCMC Sampled Values by 4 Hosts (Top) and Histogram of the Predictive Errors for 2 of the the Hosts (Bottom)

Method:	LOOCV	WAIC
ELPD Difference:	-117.6	-122.7
SE of Difference:	15.1	14.9

Table 2: ELPD Differences (Model A - Model D) via LOOCV & WAIC

4.3 Limitations

Despite the perceived superiority that the multilevel model has over the pooled model in making accurate predictions on price, it is improbable that the former is flawless. To begin, we should note the fact that the MCMC sampling for the multilevel models encountered some divergent transitions. Although those transitions did not occupy a large proportion of samples, where the maximum was 0.4% of total iterations, it may already signify that the results are, to some extent, undependable. This is the need to examine diagnostics for the MCMC sampling to better understand any fundamental issue that may arise in future model design.

Furthermore, if we consider Fig. 10, we see that the probability integral transformation (PIT) of the LOOCV results for the both model A and D do not remotely present themselves as uniform probability densities[16]. In particular, the shapes of the LOO-PIT for both models are shaped similarly. This indicates some underlying bias that is not accounted for in either model (Gabry, Simpson, Vehtari, Betancourt, & Gelman, 2019). That is, the models are mis-specified and the mis-specification likely originates from the property-level, since the bias is common among the two models. To a large extent, however, this bias perhaps should have been expected.

Firstly, the model were not specified optimally in the analysis. This specifically relates to our attempt to model the *latitude* and *longitude* variables. Although we have already included 5th order polynomial for them, it remains plausible that the two variables do interact with each other. Yet, no interaction term was included. In fact, this relates to how the models were set-up and how the priors were chosen. Under the model frameworks, apart from using the unit information prior, Bayesian models constructed in this article worked under a strong premise – all the $\{\beta\}$ parameters were all independently distributed. The inclusion of interactions would likely violate it. Thus, before extending the model, prior distributions will need to be more flexible to allow for correlations among the parameters. Possibly, we could design a model based around the basis of a Wishart prior for the co-variance matrix for β, as Chung et al. (2015) suggested .

[16] By the properties of PIT, for a continuous outcome (e.g. predicted log of price) an estimate of its CDF (CDF of the posterior predictive distribution), given an unseen observation (singleton log of price partitioned by LOO) should broadly resemble a uniform density (Abril, 2019)

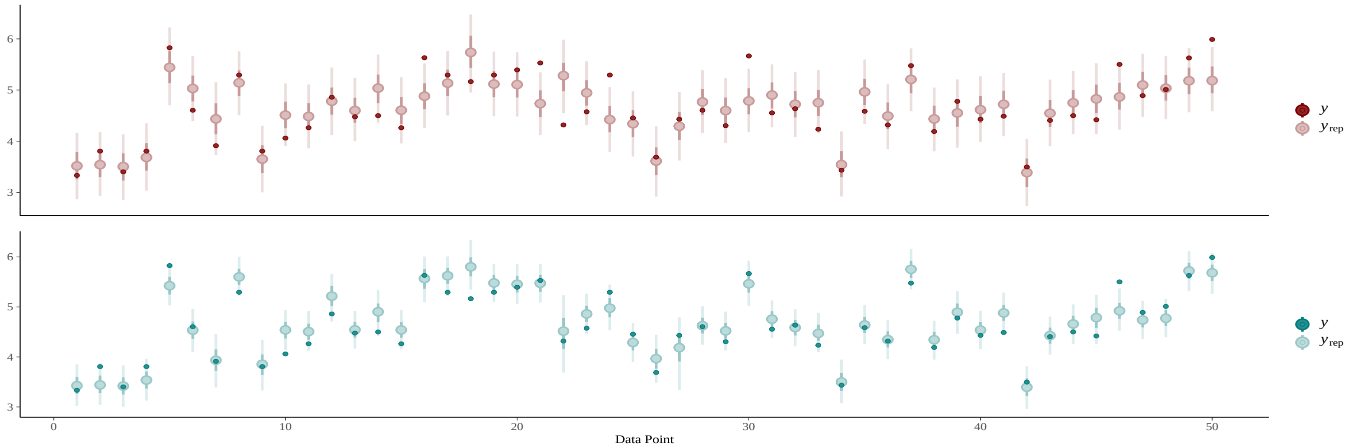

Figure 9: Comparison Between Actual Log of Price & LOOCV Predicted Values for 50 Randomly Selected Points of Data for Model A (Top) & Model D (Bottom) (50% CI and 90% CI are shown)

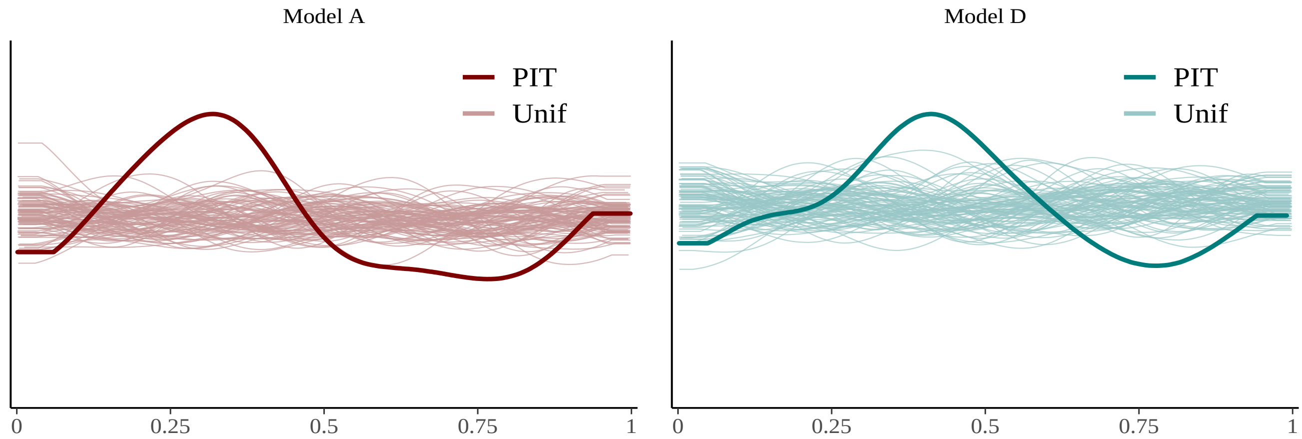

Figure 10: Comparisons of LOO-PIT to Uniform Distribution for Models A & D

Secondly, the variables in the data intuitively do not contain much information about each property that may truly determine price. Information on a property's size, interior design, furniture, would explain more of the variation in its price than the variables are available in the data. Unfortunately, information on previous tenants' review were not complete in the data and thus was not included to avoid misinterpretations and possibly more bias. Besides, the lack of computational power prevented us from including too many predictors or fitting highly-complex models.

5 Conclusion

In conclusion, the models that were constructed throughout this article have attempted to predict the (log of) prices of Vancouver properties that are listed on *Airbnb*. Given the hierarchical relationship between listed properties and their hosts, a multilevel model has unsurprisingly delivered the most accurate predictions for the property prices. The pooled model results suggested that predictors like *latitude*, *longitude*, *room type* carry the greatest strength to predict for the price. Although results from the multilevel model did not suggest otherwise, they did not support such claims as confidently as the pooled model results did. This was primarily due to the more convoluted structure of the model and that variance of the posteriors tended to be larger. Nevertheless, the attempts made throughout this article do carry some merit in tackling the price prediction problem.

Given the profound implications of accurate price prediction, there is no doubt that this conundrum will be investigated by many. Should one pursue better models for the same topic, exploration of the complete data offered by the *Inside Airbnb* database is encouraged. Without any constraint on computational capacity, one can develop sophisticated spatial models using information of property listings in numerous dimensions, across several cities. Other multilevel models can potentially be constructed with levels like years or regions, assuming exchangeability is satisfied for such them.

References

Abril, O. (2019, Jul). *Loo-pit tutorial.* Retrieved from https://oriolabril.github.io/oriol_unraveled/python/arviz/pymc3/2019/07/31/loo-pit-tutorial.html

Cox, M. (2021, Apr). *Get the data.* Retrieved from http://insideairbnb.com/get-the-data.html

Fetter, F. A. (1912). The definition of price. *The American Economic Review*, *2*(4), 783–813. Retrieved from http://www.jstor.org/stable/1828191

Gabry, J., Simpson, D., Vehtari, A., Betancourt, M., & Gelman, A. (2019, Jan). Visualization in bayesian workflow. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *182*(2), 389–402. Retrieved from http://dx.doi.org/10.1111/rssa.12378 doi: 10.1111/rssa.12378

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian analysis*, *1*(3), 515–534.

Gelman, A., Hwang, J., & Vehtari, A. (2013). *Understanding predictive information criteria for bayesian models.*