MA678 Final Project

Modeling Airbnb Prices using Linear and Multilevel Regression

Andrew Zhang

Abstract

The purpose of this project was to model 2016 Airbnb prices based on survey data such as room ID, host ID, neighborhood, and bedrooms and then test the model on different years of data to see if prices for those years are modeled the same. There were three regression models that were implemented to try to best fit this data: linear regression, log-linear regression, and a multilevel hierarchical model. We found that prices were often proportional to the housing rent of the area. Furthermore, prices seemed to be determined by the number of bedrooms being offered in the listing and how accommodating the listing was for visitors. After applying the 2016 models to the 2014 and 2017 datasets, we were able to conclude that there was potentially a difference in how AirBnb priced 2016 listings compared to those of 2014. In addition, the 2017 listings seemed to be predicted much more accurately potentially due to the fact that 2016 is closer, timewise, to 2017. Thus, we concluded that AirBnb priced rentals differently between 2014 and 2017.

Background

Airbnb, although relatively new, has brought about a huge change in the field of hospitality. Rather than stay in hotels or motels, people now have the option to stay in the comfort of other people's homes for a much cheaper rate. Airbnb allows it's hosts to come up with prices for people who decide to stay in their homes, but Airbnb makes recommendations on prices as well. For this particular study, we want to look at a city that has a sizable number of listings. Therefore, we've selected Los Angeles, as it is a popular hotspot for tourism and hospitality. The goal of this project is to understand how Airbnb prices listings and whether or not we can create a model that can accurately predict different years of listings. We will then apply this model to a 2014 and 2017 AirBnb dataset to see how well the model predicts prices and compare those predictions to the actual prices of the listing.

Data

Variables

Variable Name	Description
room_id	Identification of rooms given by Airbnb
host_id	Identification for hosts given by Airbnb
room_type	Type of room rental being listed(Shared Room, Private Room, Entire Home/Apartment)
borough	A subregion of the city or search area for which the survey was carried out in
neighborhood	A subregion of the city or search area for which the survey was carried out in
reviews	Number of reviews given to a particular listing
overall_satisfaction	Average rating(out of five) that a particular listing has received from past visitors
accommodates	Number of guests the listing can house
bedrooms	Number of bedrooms the listing has
price	The amount of money required to stay per night
minstay	Minimum stay for a visit(by day)

latitude/longitude	Latitude and longitude of the listing posted on the Airbnb website
last_modified	Date and time that the datapoints were read from the Airbnb website

Data Cleaning/Exploratory Data Analysis

Looking at the original dataset, we can see that it has 30671 observations with 13 variables. The first thing we look at is the structure of the data and the number of NA's within each column. Immediately, we can remove the variables *borough*, *minstay* as the columns are filled with NA's. Variables *room_id*, *host_id*, and *last_modified* are also removed because they don't provide much for the analysis. Next, we look at *neighborhood* which is a factor with 202 levels. We only want to look at listings with a significant number of observations and ignore those with only a few listings. For this reason, we filter the dataset for neighborhoods containing more than 200 counts. Finally, we create three separate datasets for each of the room types(Entire Apartment, Private Room, Shared Room).

Next, we look at the importance of each variable.

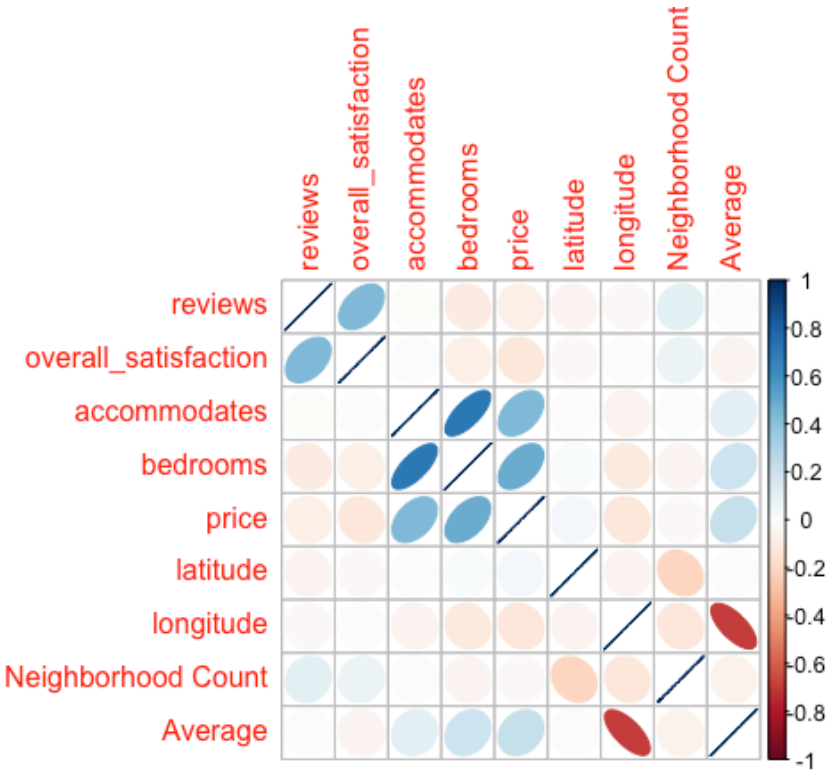


Figure 1

One of the first things we want to look at in terms of variables and their interactions is a correlation plot. We want to focus on the variable *price* as it is the response variable we are interested in. From *Figure 1*, we can see that there is a strong correlation between price, accommodations, and bedrooms. We can also see a strong correlation between bedrooms and accommodations.

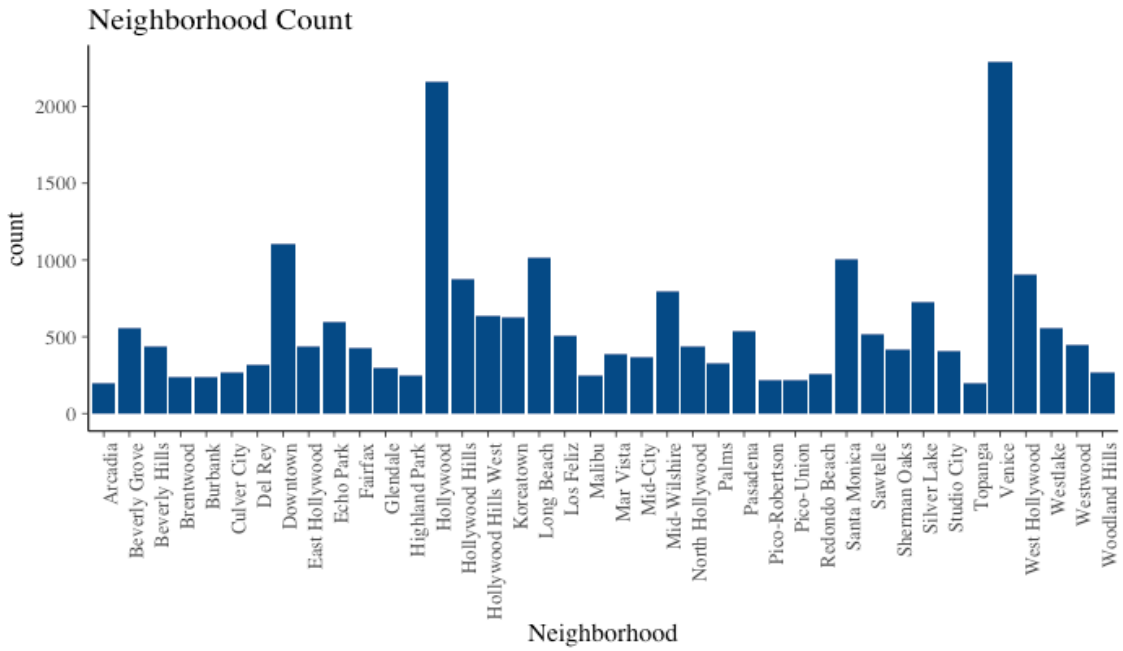


Figure 2

In *Figure 2*, we can see that there are a couple of neighborhoods with significantly more counts than the others. Cities like Venice, Hollywood, Long Beach, Santa Monica are all some of the most popular cities in Los Angeles. Originally, there were 202 neighborhoods, but since we only wanted the top cities, we reduced the number of neighborhoods to 40 based on the counts.

Figure 3

Figure 3 above shows a distribution of pricing in relation to the number of reviews left for each listing separated by room type. As we can see from the visual, a large majority of the data points are concentrated around 0, regardless of room type. This indicates that the majority of people do not leave reviews after their stays. However, we can also see that there is a segmentation of pricing amongst the groups, where apartments generally cost more than private rooms which generally cost more than shared rooms. In addition, apartments appear to have more reviews than private rooms or shared rooms.

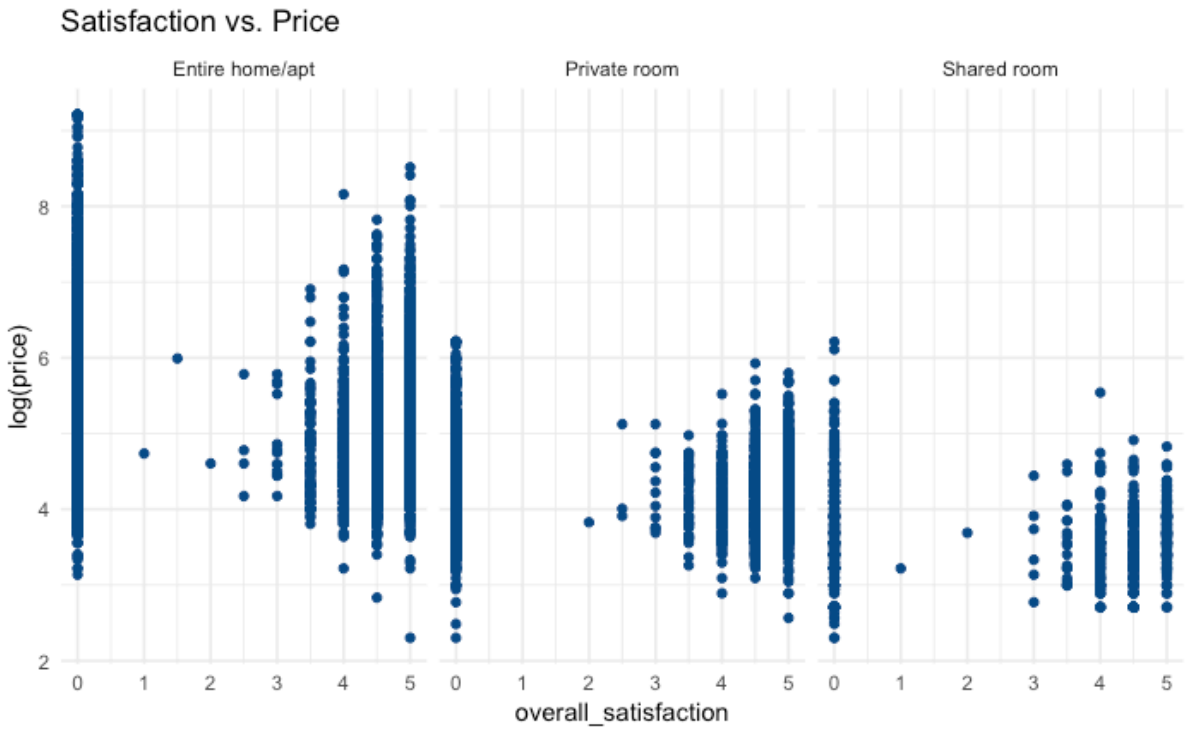


Figure 4

Looking at *Figure 4* above, we can see that each of the types of rentals have a significant number of 0's in overall satisfaction. These 0's represent surveys that were left unanswered, however, prices tend to increase as the satisfaction increases in all types of rentals. We can also see that the prices generally decrease for each of the room types.

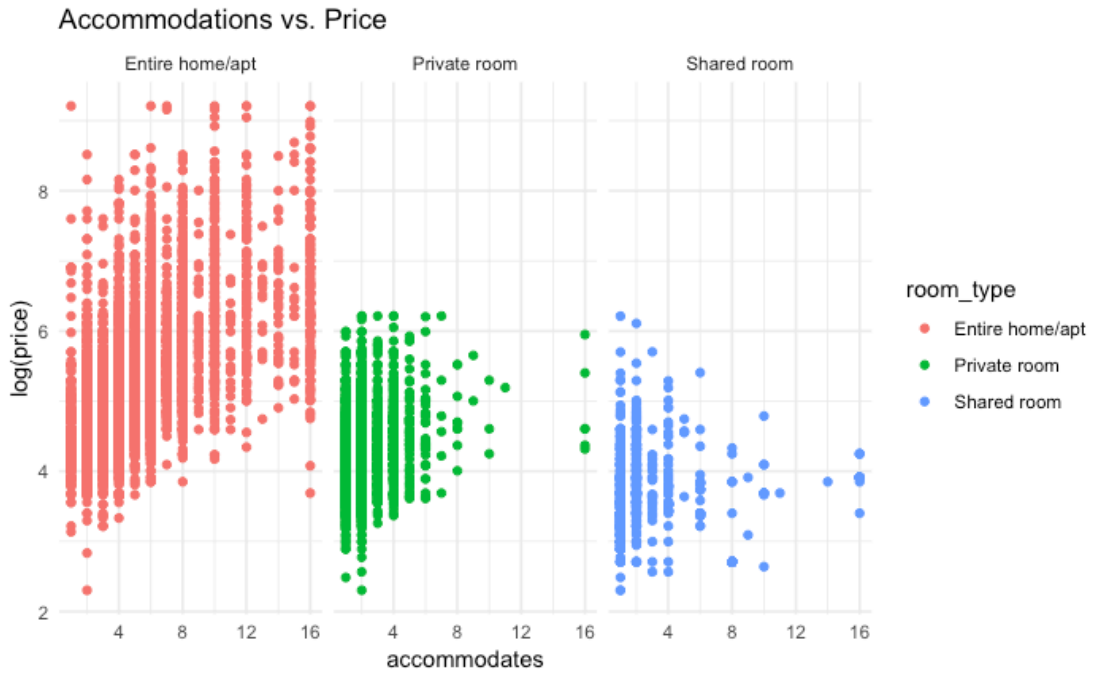


Figure 5

From *Figure 5* above, we can see that there is a positive trend between accommodation and price only in the apartment rentals. This makes sense as private rooms and shared rooms won't be able to house as many people as an entire apartment. Therefore, we can see that accommodations for apartment range between 1 and 16, whereas accommodations for private and shared rooms are clustered between 1 and 6.

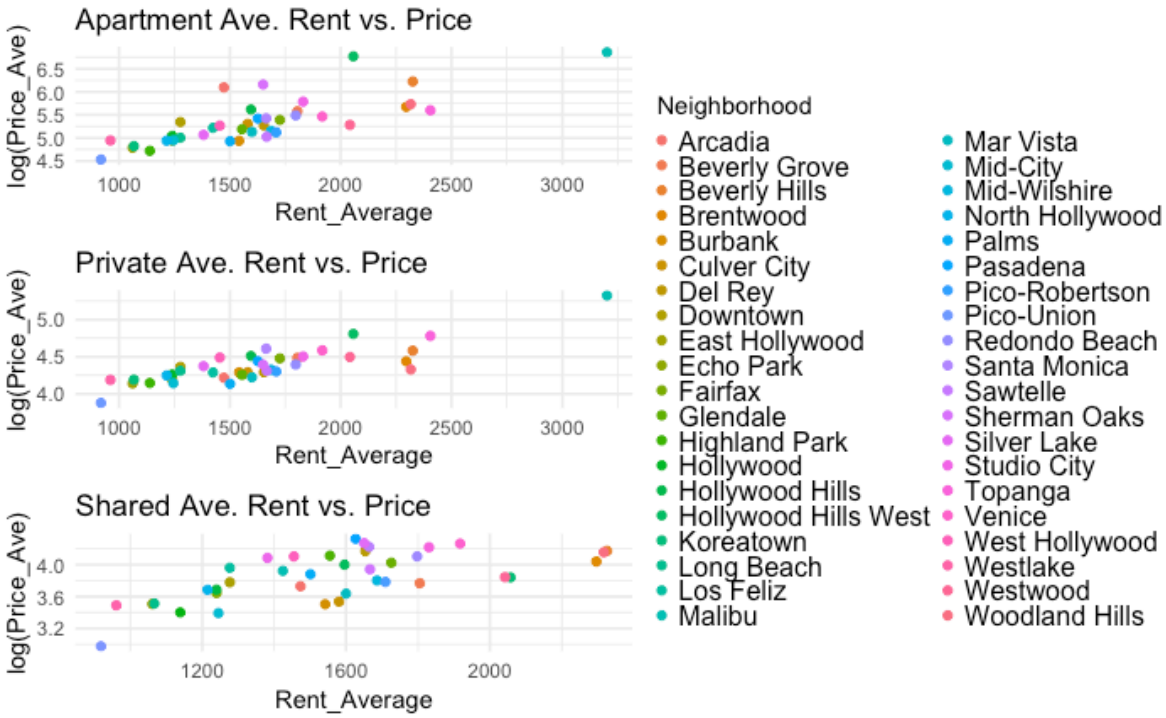


Figure 6

One of the interactions we want to look at when referring to hospitality prices is the respective rent for specific areas. Therefore, given there are multiple observations for the same neighborhoods, we took the average of the prices listed within these neighborhoods and aggregated all the observations into one observation. Then we compared them to the respective average rent for the area. From *Figure 6*, we can see that there is a positive trend between the rent and the price of the listing. This is true for apartment, private rooms, and shared rooms. Logically, this makes sense as the more luxurious the area, the more expensive a listing with be.

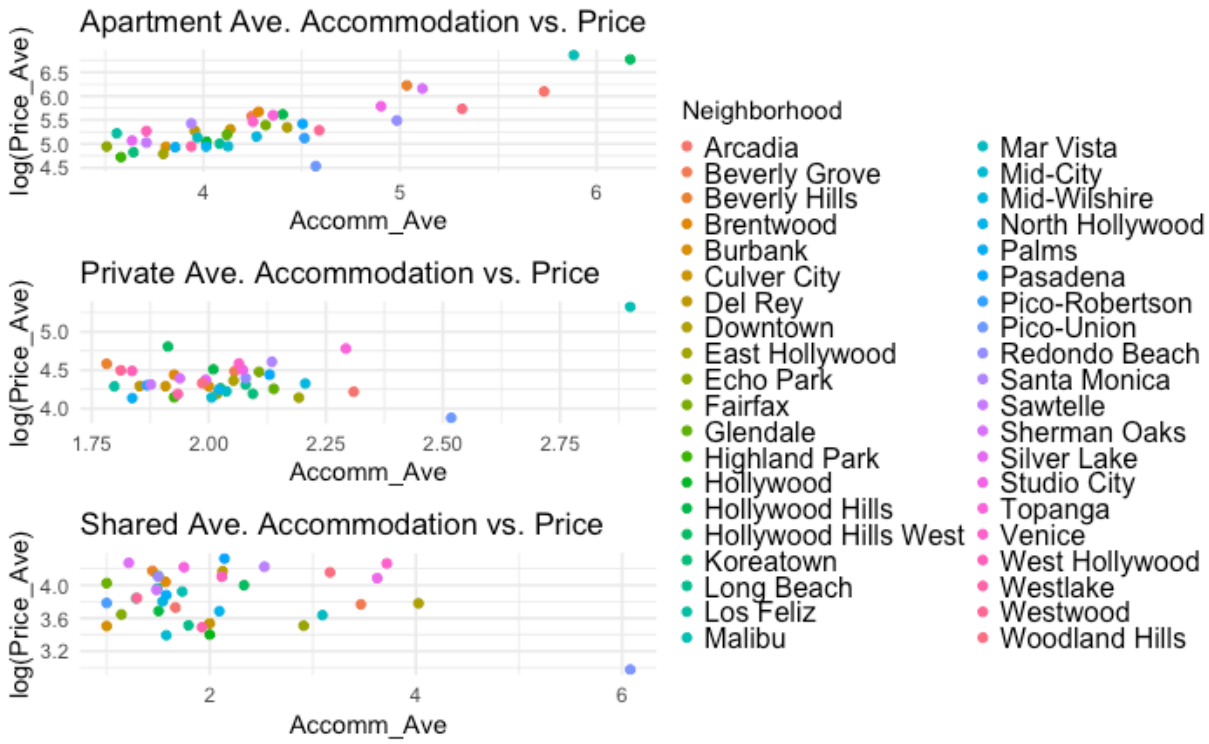


Figure 7

According to the correlation plot created earlier, the next variable we want to look at is accommodations. In *Figure 7*, we can see that for apartments, there is a clear positive trend between the number of accommodations and the prices for the listing. This is reflective of the trend in *Figure 5*.

Methods

This data was taken from Tom Slee on his website where he compiled and collected the data from Airbnb's website.

Linear Regression

We start off modeling applying a linear regression to see whether the data follows a linear normal distribution. To a certain degree it does, however, there were a number of things that could be done to help improve model performance.

Log Linear Regression

If we look at the linear regression model, we can see there are extreme values being plotted. This is a combination of multiple variables, but we look at the *price* variable. We can also see that for certain neighborhoods, there are extreme values. In order to combat this, we apply a log transformation to the prices. We can see that the model performs significantly better in the model check section.

Log Linear Regression with Interaction

If we go back to the correlation plot in *Figure 1*, we can see that there is a strong correlation between bedrooms and accommodations. For that reason, it was worth looking into a potential interaction between the two using a model.

Multilevel Model

We ran a multilevel model because there were multiple observations for certain variables such as host_ID and neighborhoods. Therefore, we created linear models taking into account the nested factor levels in neighborhoods.

Results

Given that we ran three different linear regression models, it seems like the best model for all three room types is the log-linear regression with interactions. *Table 1* is a chart of the R-squared values for the raw data.

	Linear Model	Log Linear Model	Log Linear with Interaction
Apartment	0.3039	0.6017	0.6033
Private	0.219	0.2215	0.2218
Shared	0.1781	0.3069	0.3069

Table 1

However, as we will see in the model checking, there are significant outliers in the data. By removing those and plotting the results, we can see that our models improve significantly. *Table 2* is the chart of R-squared values for the cleaned data.

	Linear Model	Log Linear Model	Log Linear with Interaction
Apartment	0.5326	0.67	0.6604
Private	0.2929	0.3211	0.3198
Shared	0.562	0.8058	0.8058

Table 2

We can still see that the log-linear regression with interactions suits the data from private and shared rooms the best, but the log-linear model fits the data collected from apartments better. Next we look at the AIC values for both the linear regression models and the multilevel

hierarchical model nested by neighborhoods. AIC will help us compare all of the models as multilevel models do not have R-squared values.

Linear Regression

	Linear Model	Log Linear Model	Log Linear with Interaction
Apartment	184545	11078	11027
Private	55840	1081.3	1130.3
Shared	5223.9	-288.55	-288.55

Table 3

Multilevel

	Linear Model	Log Linear Model	Log Linear with Interaction
Apartment	184674	11277.1	11239.8
Private	55964	1251.1	1301.1
Shared	5279.1	-175	-175

Table 4

Now that we have compared all of the models, we can see that the linear regression models perform better than the multilevel models just slightly. Therefore, we can conclude that the log-linear model is best for the private room listings and the log-linear with interaction is best for apartment and shared room listings.

Interpretation

Apartment

- Intercept: The average price for an apartment, when all other variables are 0, is exp(-75.21) or 2.17×10^{-33}
- Reviews: For every one unit increase in reviews, while all other variables are held constant, the price of the listing increases by exp(-0.0005079) or 0.99949
- Overall_Satisfaction: For every one unit increase in overall satisfaction, while all other variables are held constant, the price of the listing increases by exp(-0.009568) or 0.9905
- Accommodates: For every one unit increase in accommodations, while all other variables are held constant, the price of the listing increases by exp(0.05305) or 1.0545
- Bedrooms: For every one unit increase in bedrooms, while all other variables are held constant, the price of the listing increases by exp(0.2876) or 1.333
- Latitude: For every one unit increase in latitude, while all other variables are held constant, the price of the listing increases by exp(-1.875) or 0.15333
- Longitude: For every one unit increase in longitude, while all other variables are held constant, the price of the listing increases by exp(-1.216) or 0.2964

- Accommodates*Bedrooms: For every one unit increase in, while all other variables are held constant, the price of the listing increases by exp(0.001711) or 1.002
- Neighborhoods: The differences in pricing amongst the various neighborhoods ranges from -56.3% to 4.133%

Private

- Intercept: The average price for an apartment, when all other variables are 0, is exp(-116.5) or 2.53×10^{-51}
- Reviews: For every one unit increase in, while all other variables are held constant, the price of the listing increases by exp(-0.0001529) or 1.000
- Overall_Satisfaction: For every one unit increase in, while all other variables are held constant, the price of the listing increases by exp(0.02316) or 0.9771
- Accommodates: For every one unit increase in, while all other variables are held constant, the price of the listing increases by exp(0.08402) or 1.0877
- Bedrooms: For every one unit increase in, while all other variables are held constant, the price of the listing increases by exp(-0.3198) or 0.7263
- Latitude: For every one unit increase in, while all other variables are held constant, the price of the listing increases by exp(-1.073) or 0.3419
- Longitude: For every one unit increase in, while all other variables are held constant, the price of the listing increases by exp(-1.333) or 0.26369
- Neighborhoods: The differences in pricing amongst the various neighborhoods ranges from -52.7% to 9.39%

Shared

- Intercept: The average price for an apartment, when all other variables are 0, is exp(147.4)
- Reviews: For every one unit increase in, while all other variables are held constant, the price of the listing increases by exp(-0.0002226) or 0.99978
- Overall_Satisfaction: For every one unit increase in, while all other variables are held constant, the price of the listing increases by exp(0.02545) or 1.026
- Accommodates: For every one unit increase in, while all other variables are held constant, the price of the listing increases by exp(0.0182) or 1.01836
- Bedrooms: NA since the number of bedrooms is only 1
- Latitude: For every one unit increase in, while all other variables are held constant, the price of the listing increases by exp(-3.094) or 0.04532
- Longitude: For every one unit increase in, while all other variables are held constant, the price of the listing increases by exp(1.3844) or 0.3253
- Neighborhoods: The differences in pricing amongst the various neighborhoods ranges from -4.7% to 147.2%

Model Checking

We look at 4 models for each of the 3 different room types: apartment, private room, and shared room. First, we will take a look at the residuals vs. fitted plots.

Apartments

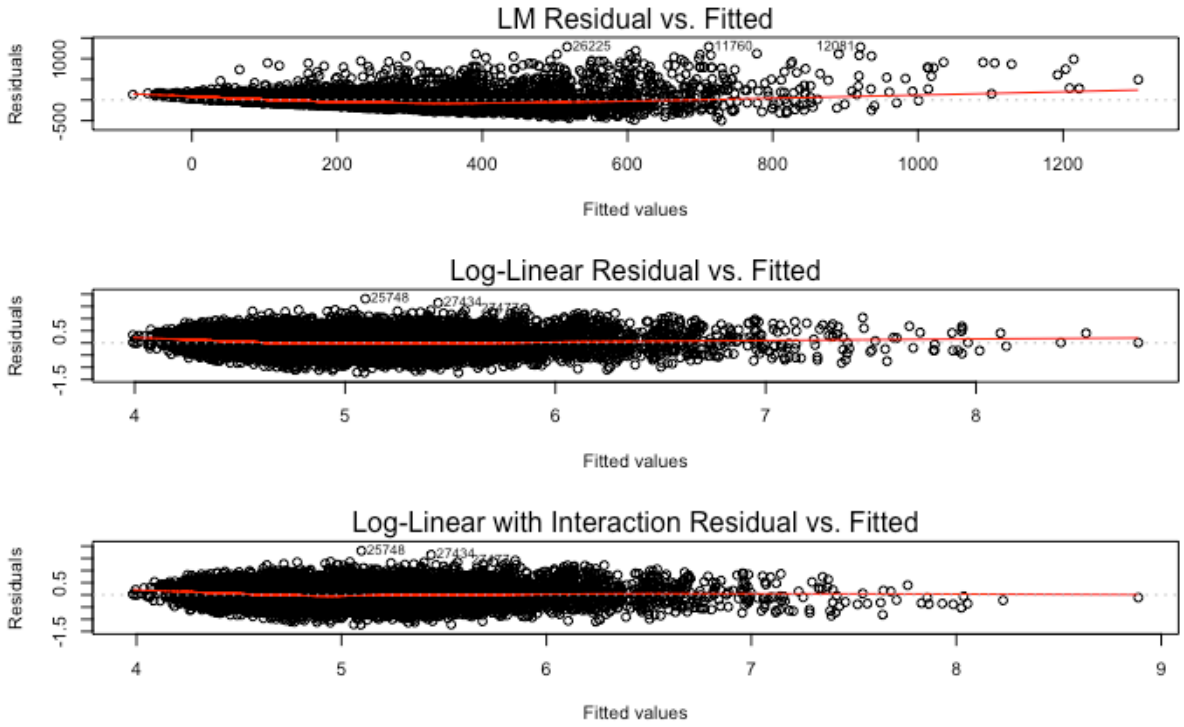


Figure 8

Private

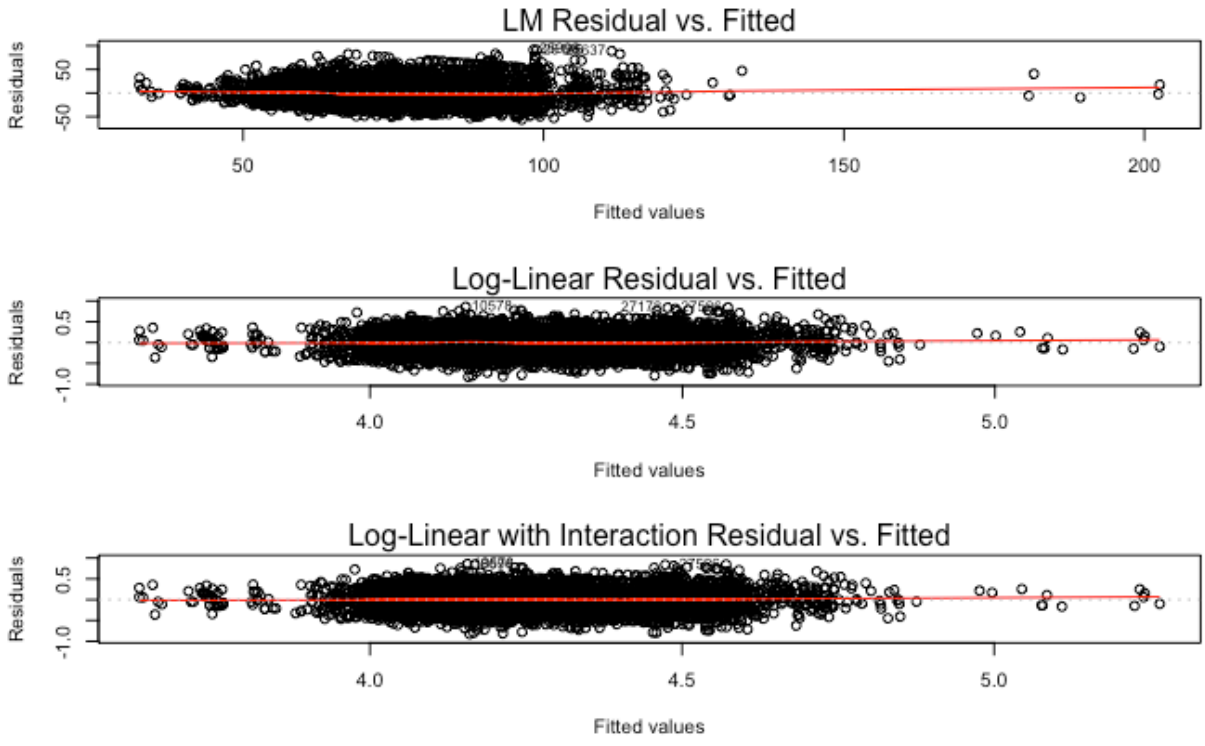


Figure 9

Shared

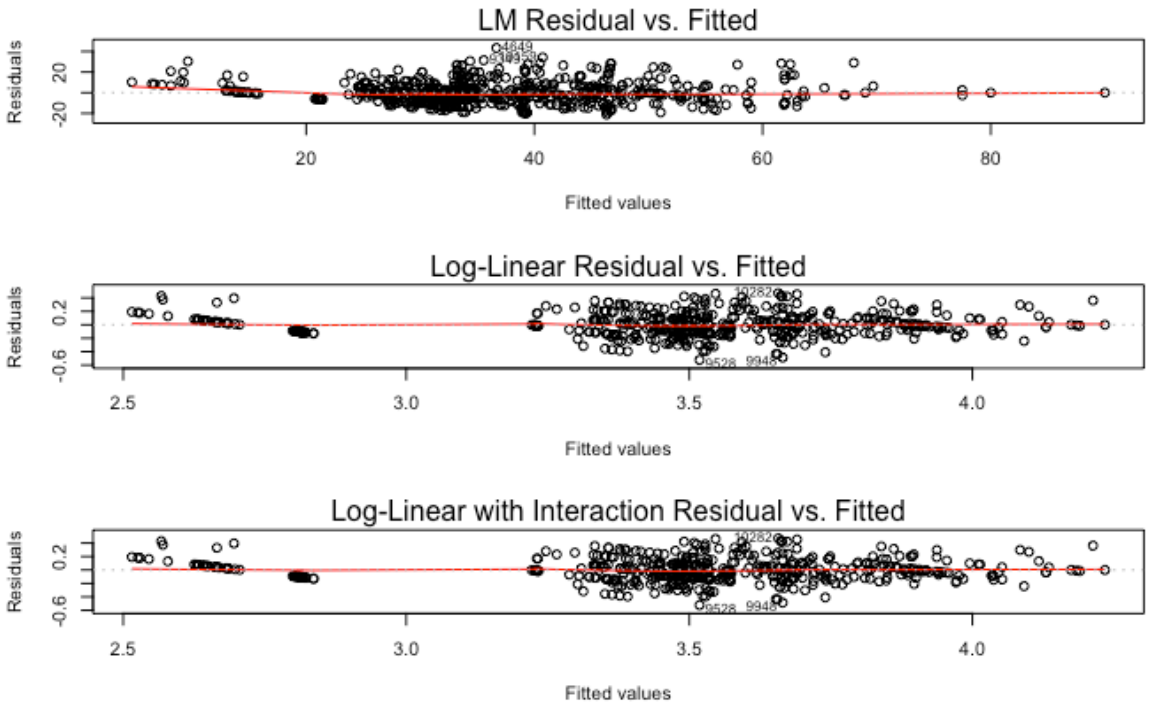
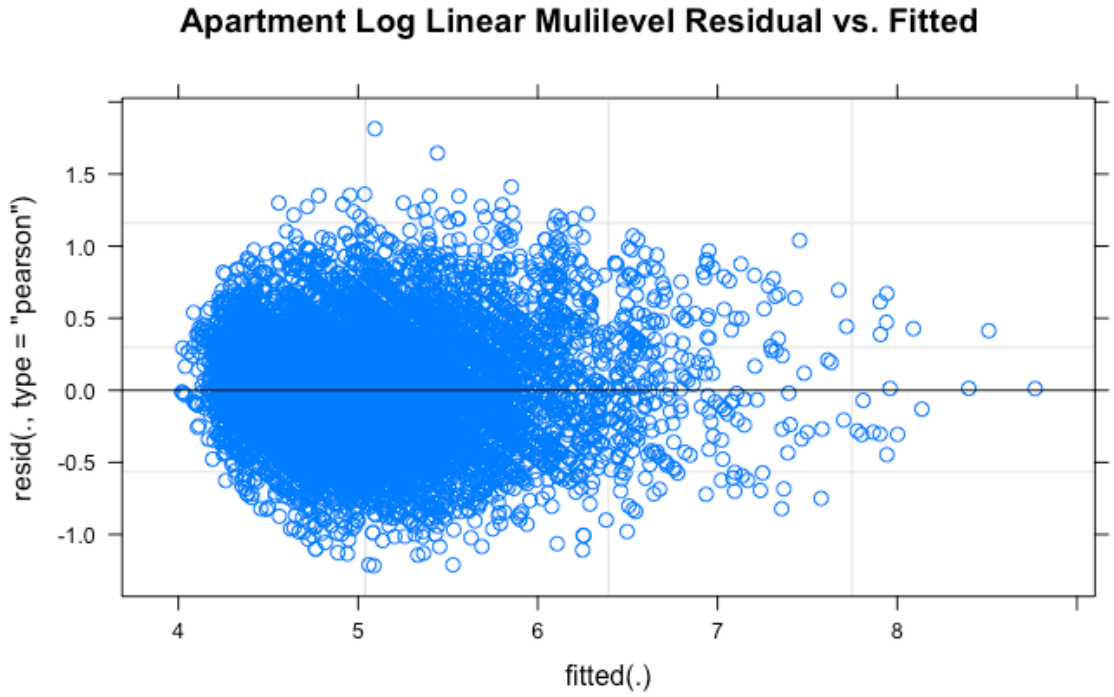


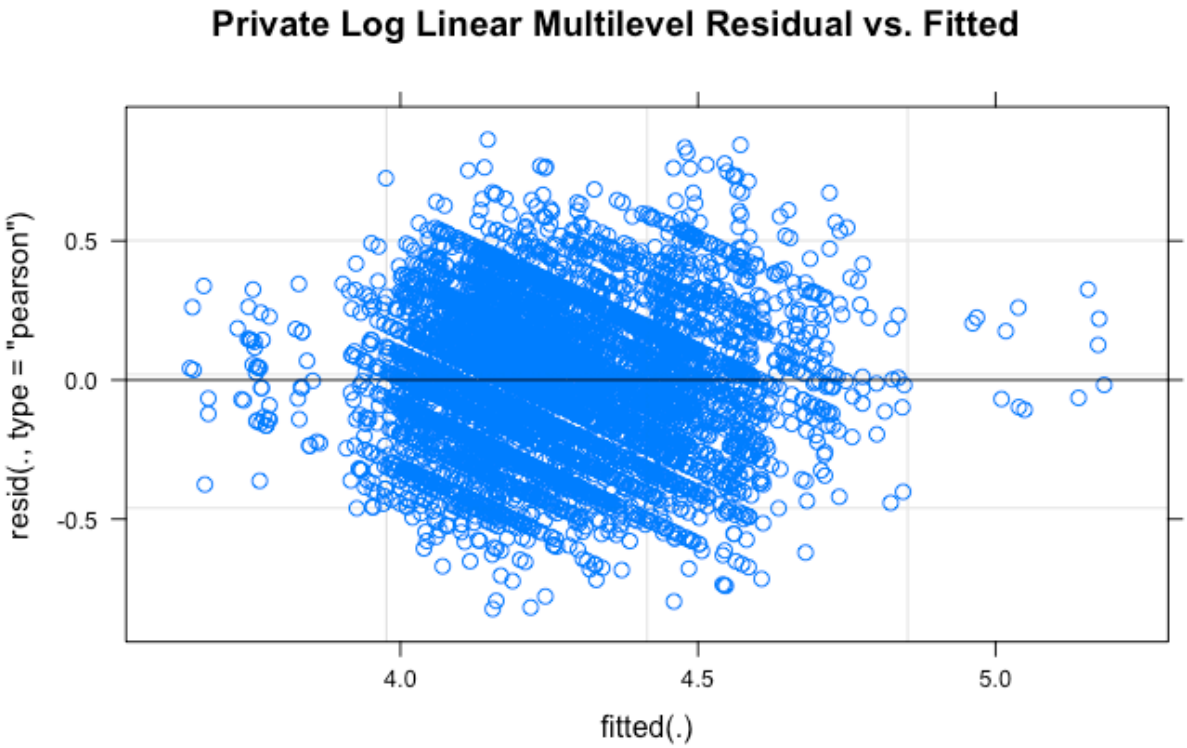
Figure 10

Now we look at the model diagnostics for the multilevel model.

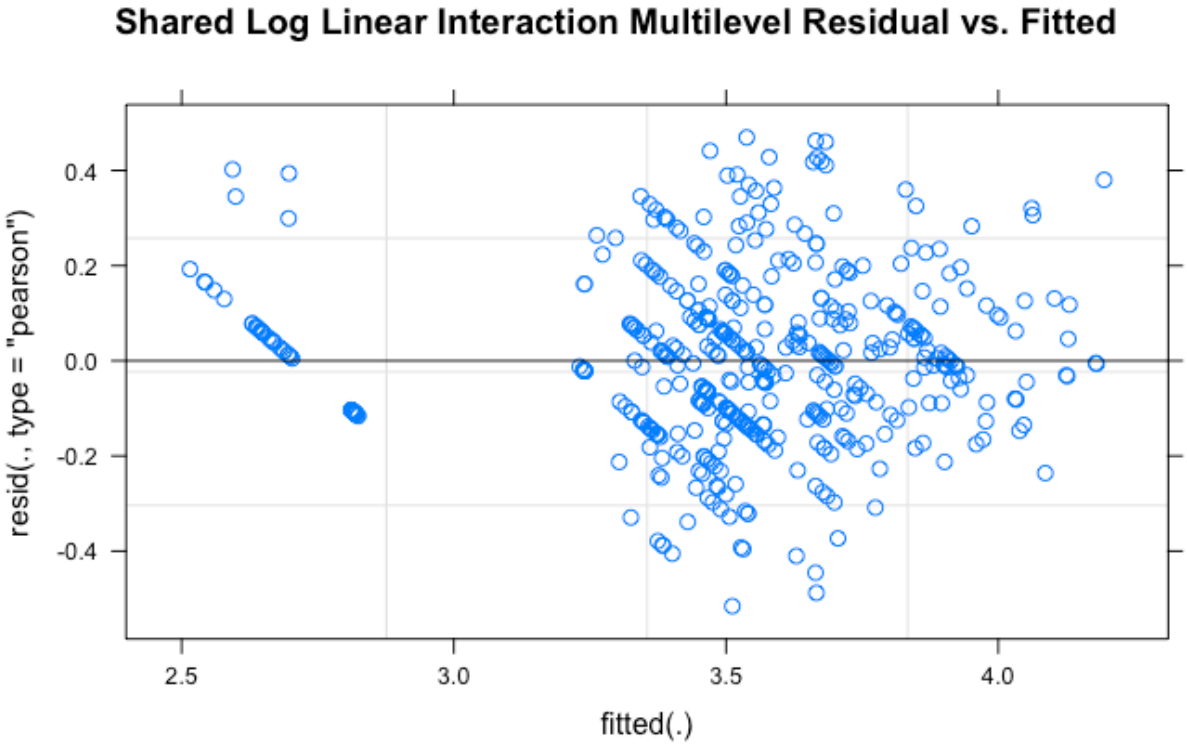
Apartment



Private



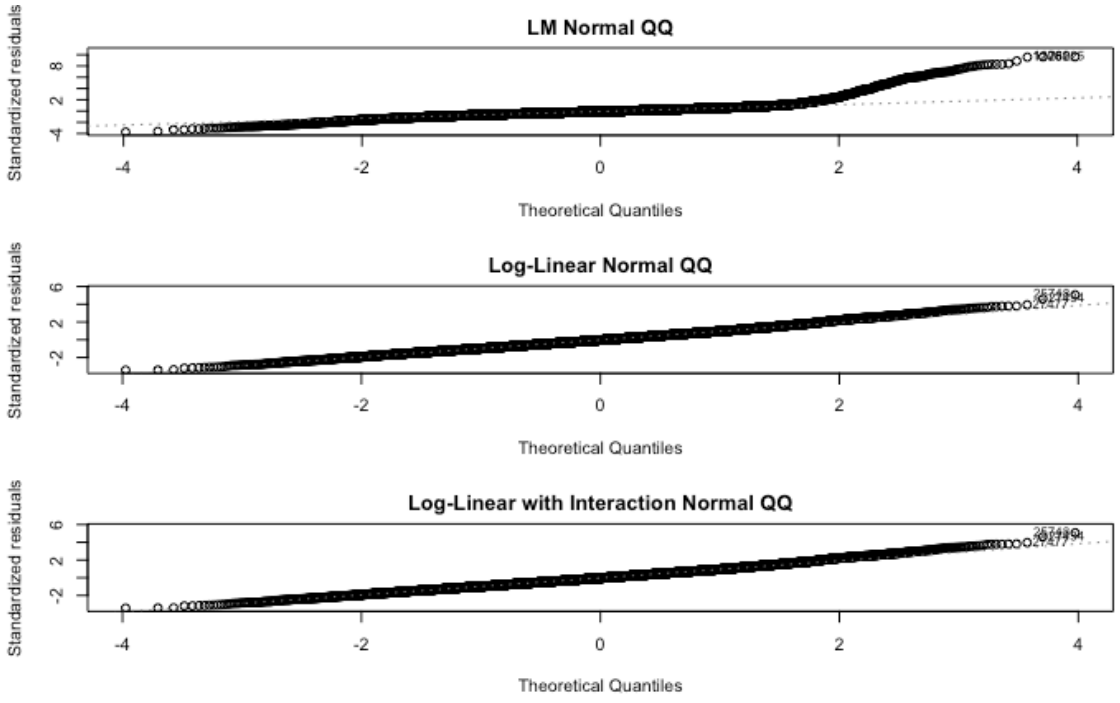
Shared



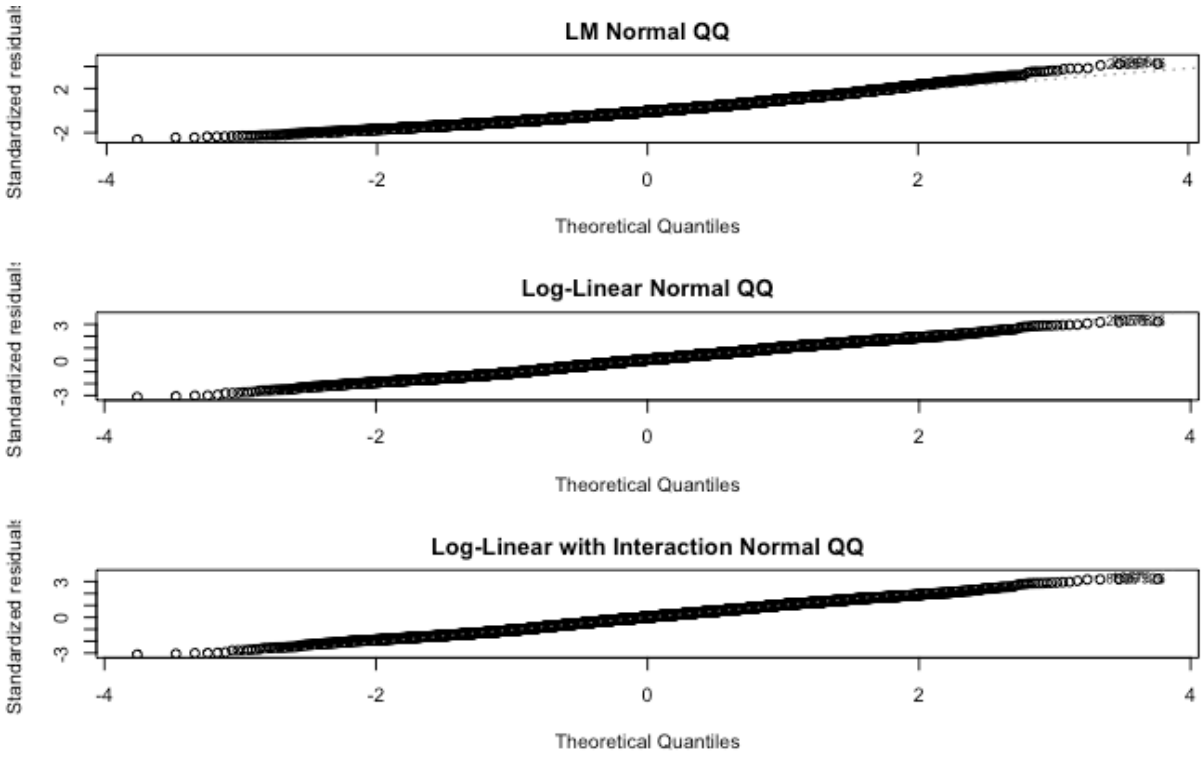
From each of the combined plots, we can see that the original linear model had several outliers that deviated away from the center of 0. However, applying log and interactions helped to center most of the data points closer to 0. We can see that there are still a couple of small outliers in both the log-linear and log-linear with interaction residual plots. For the most part, these two plots are identical. Then considering the multilevel model, we can see that similar to the diagnostics for the log-linear models, there are a couple of points slightly further away from the 0 line.

Next, we look at the normal Q-Q plots for each of the room types.

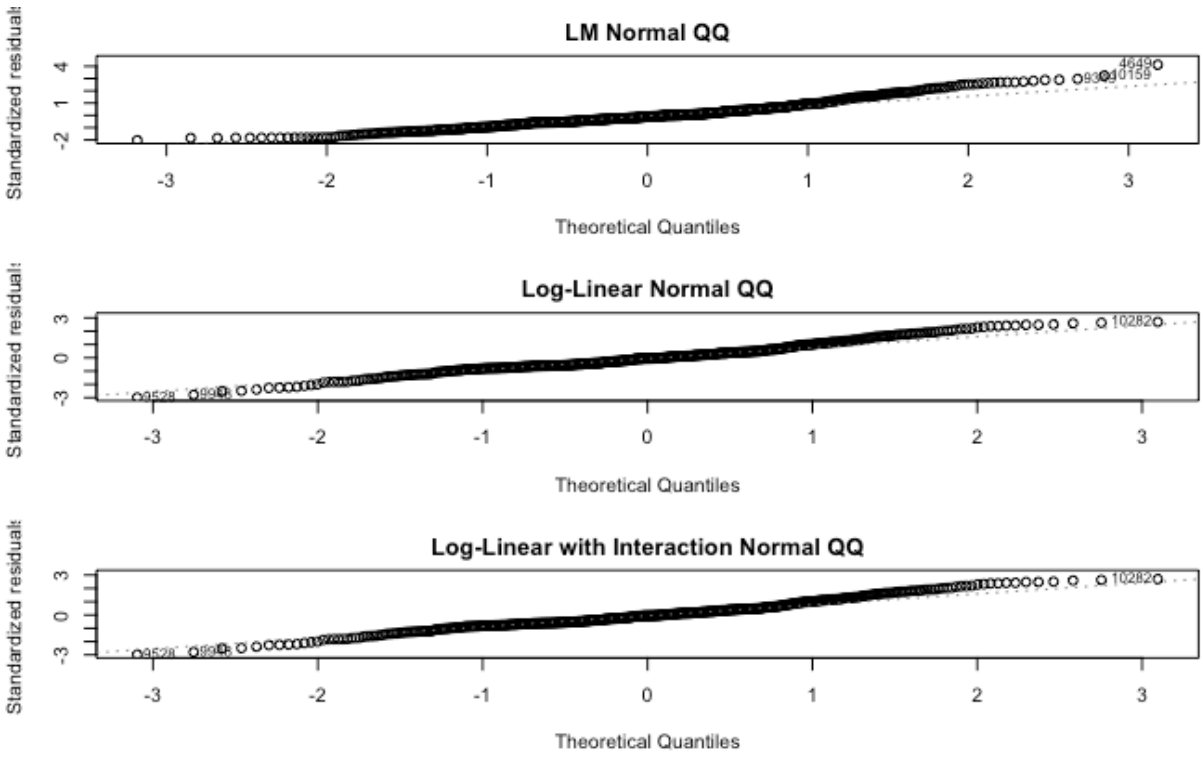
Apartment



Private


LM Normal QQ
Standardized residuals
Theoretical Quantiles
Log-Linear Normal QQ
Standardized residuals
Theoretical Quantiles
Log-Linear with Interaction Normal QQ
Standardized residuals
Theoretical Quantiles

Shared


LM Normal QQ
Standardized residuals
Theoretical Quantiles
Log-Linear Normal QQ
Standardized residuals
Theoretical Quantiles
Log-Linear with Interaction Normal QQ
Standardized residuals
Theoretical Quantiles

Discussion

Implication

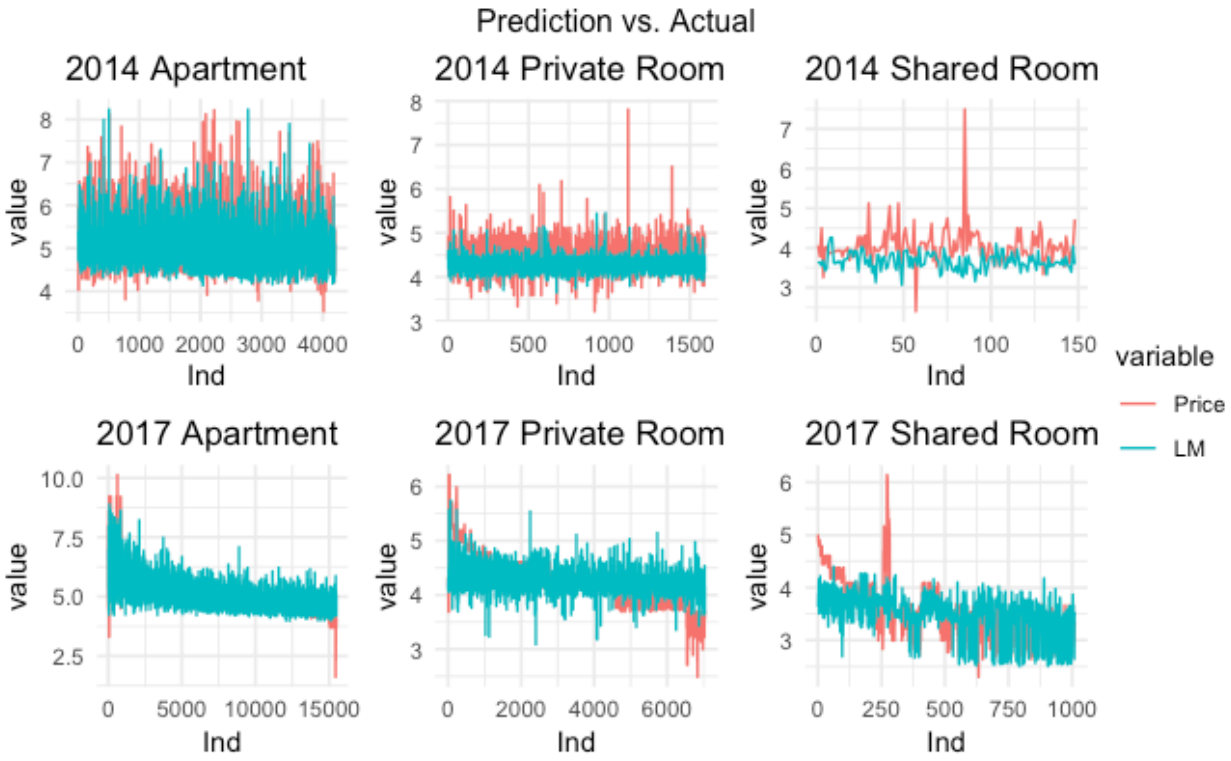


Figure 11

Given that Airbnb is at the forefront of hospitality in 2018, it is important to understand what factors play a role in influencing listing prices. Although there are other variables that can be taken into consideration, we already discovered that the most important variables are number of bedrooms and accommodations while also taking into consideration rental prices of the area. In addition, it seems as though the log-linear with interaction is the best model for the different room types. There are slight differences between the log-linear and the interaction model, but for the most part they both yield almost the same results.

If we look at *Figure 11* above, we plot the logarithmic price predictions for the 2014 and 2017 datasets in blue and compare them with the log of the actual prices given in the 2014 and 2017 datasets, which are given in red. Looking at the plots, we can see that the predictions for the 2014 dataset are generally lower than the actual prices as we can see significantly more red above the blue lines. However, when looking at the 2017 plots, we can see the predictions are much more in line with the actual predictions. There are clear outliers, however, the predictions seem to fit the data better. This could be an indication that the way AirBnb priced 2016 listings different from the way they priced 2014 listings. Logically, since 2017 is closer to 2016, the way listings are priced should be similar, and we can see that in the predictions vs. actual plots for 2017 in *Figure 11*.

Limitation/Future Work

Even though we have 30000+ listings for Los Angeles, many of the columns such as reviews were filled with 0, making it difficult to accumulate a large enough sample size to make predictions from. With that being said, for future use, maybe we can either collect more data or employ more variables for the model. In terms of variables, the data was rather simple and we weren't able to dive deeper into interactions and interactions nested within variables such as host_ID or room_ID.

Another aspect that deserves to be looked at is the longitude and latitude. There are multiple observations per neighborhood but the longitude and latitudes within the neighborhoods are not the same. When looking up rent for these specific areas, it was very difficult to match the latitudes and longitudes to get specific rent prices for these subareas. Therefore, in the future, maybe we can convert these longitudes and latitudes into zip codes, making it significantly easier to match and find rent prices for specific areas.

Finally for the multilevel model, one aspect that was potentially interesting was the nesting of rooms or hosts within the data. Given there are duplicate observations per host, we could theoretically create a model accounting for that. Unfortunately, there are too many unique hosts to create a model. In addition, the room_ID's are all unique, meaning that the same room rented out by a host in one instance will have a different room ID if its rented out in another instance. Therefore, we cannot track rooms and nest the model accounting for the rooms. However, in the future, if more data was collected and the formatting of these host IDs and room IDs were changed, we could create multiple multilevel models accounting for host ID, room ID, and neighborhood.

Acknowledgement

Thank you to Professor Yajima for supplying the tutorials and necessary understanding to complete this project. I would also like to thank Mr. Slee for providing the dataset for all Los Angeles listings.

References

[1] http://tomslee.net/airbnb-data-collection-get-the-data

[2] Professor Yajima's RStudio Tutorials