Airbnb Price Model

Andrew Zhang

12/6/2018

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 both the 2014 and 2017 datasets, we were able to conclude that the way Airbnb priced in 2016 was potentially different from the way it priced listings in 2014. However, it seemed to price 2017 listings well, allowing us to conclude 2017 listings priced similarly to 2016 listings.

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. Due to the abundance of data per room rental, we decided to create a model for each of the three room types to understand the characteristics of each rental type. We then applied the "best" models to the 2014 and 2017 AirBnb datasets to see how well the models predicted prices and compared those predictions to the actual prices of the listing. We suspect that the 2017 will be predicted better than 2014 as there is a lower likelihood that Airbnb would have changed their pricing criteria in such a short amount of time.

Variables

Table 1: Variables

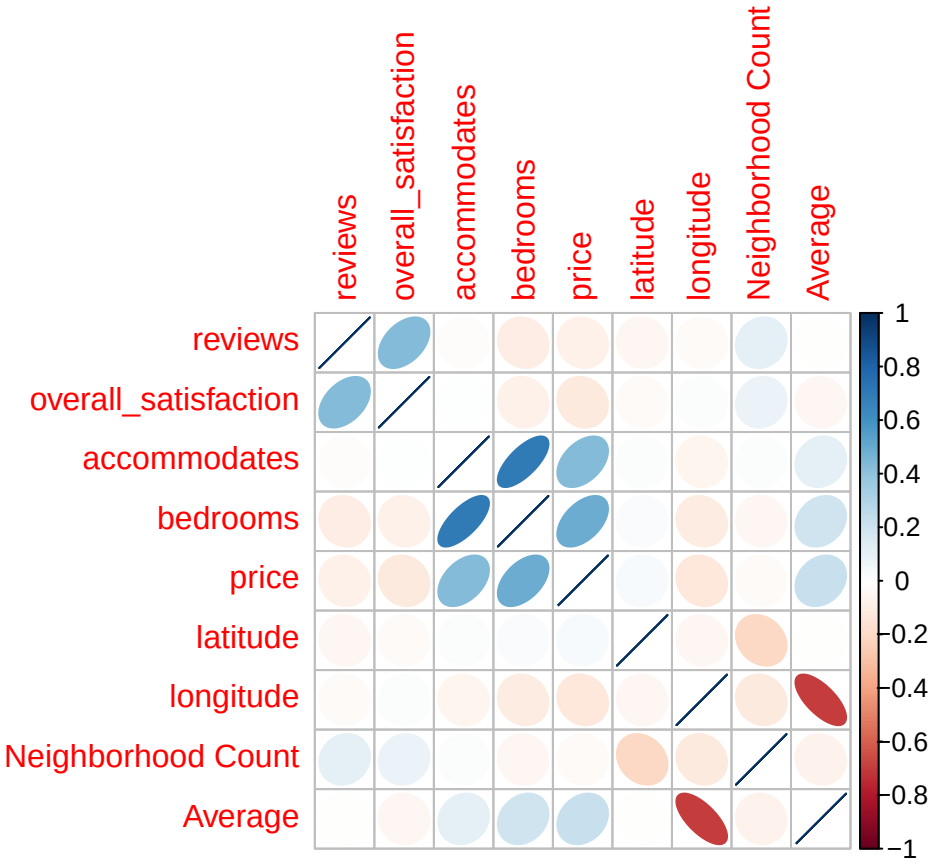
variable_names	descriptions
room_id	Identification of rooms given by Airbnb
host_id	Identification for hosts given by Airbnb
room_type	Type of room rental being listed(Shared Room, Private Room, Entire Home/Apartment)
borough	A subregion of the city or search area for which the survey was carried out 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)

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

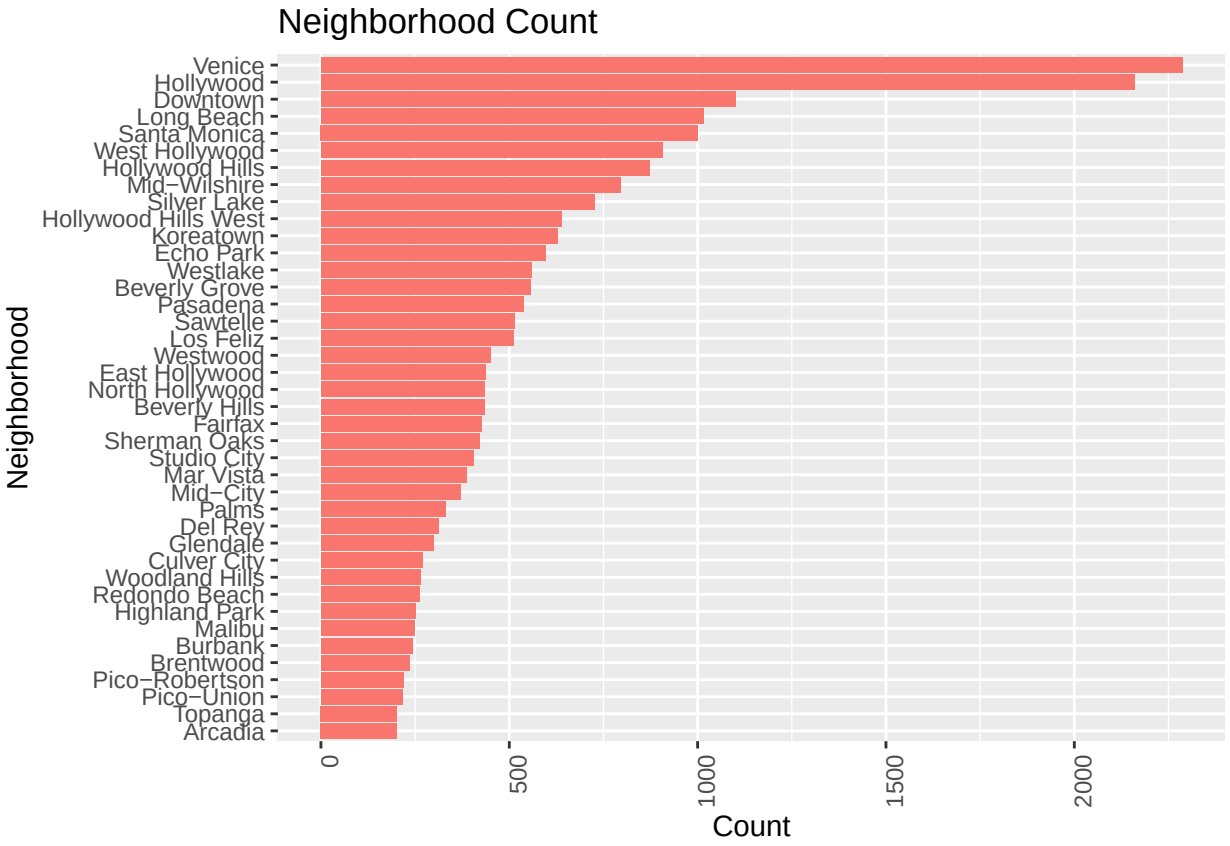
Data Cleaning/Exploratory Data Analysis

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 and 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 predictive power 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 by filtering for each of the room types(Entire Apartment, Private Room, Shared Room).

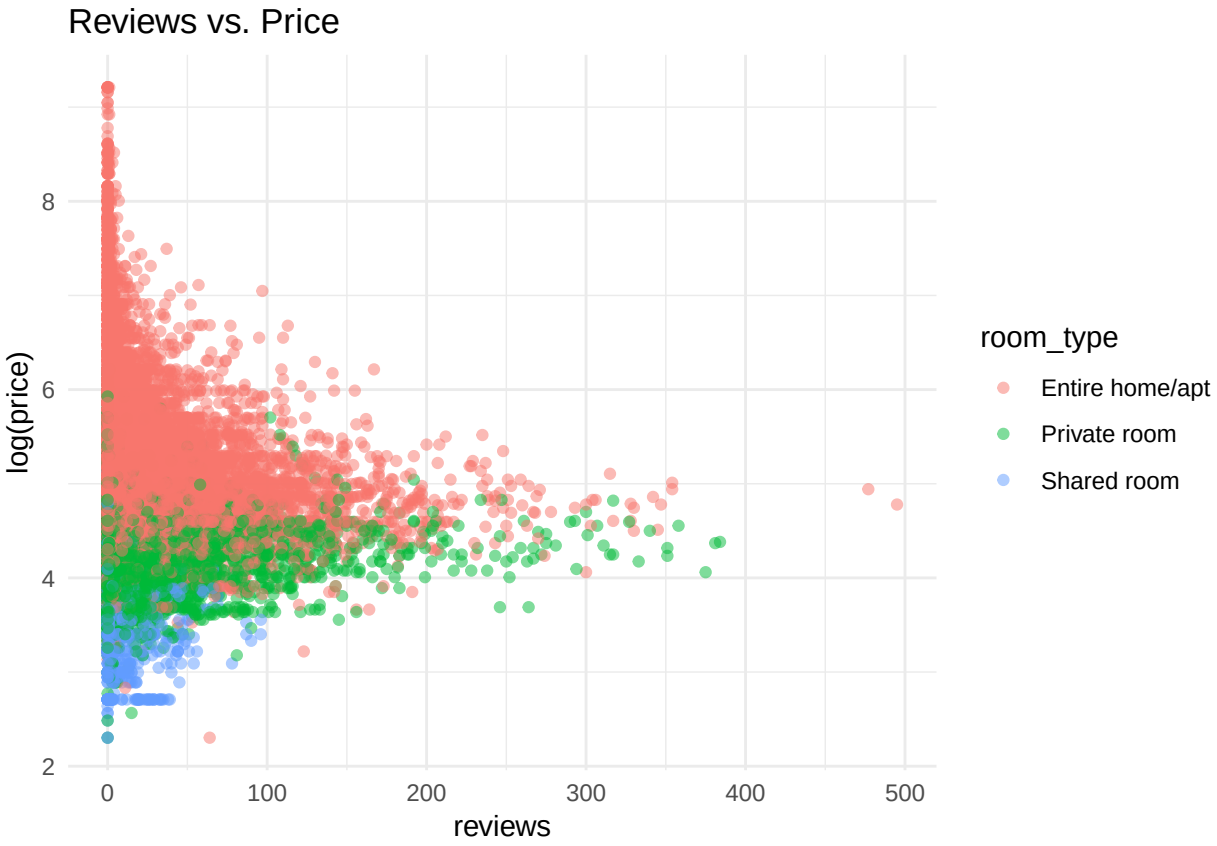
Next, we look at the importance of each variable.



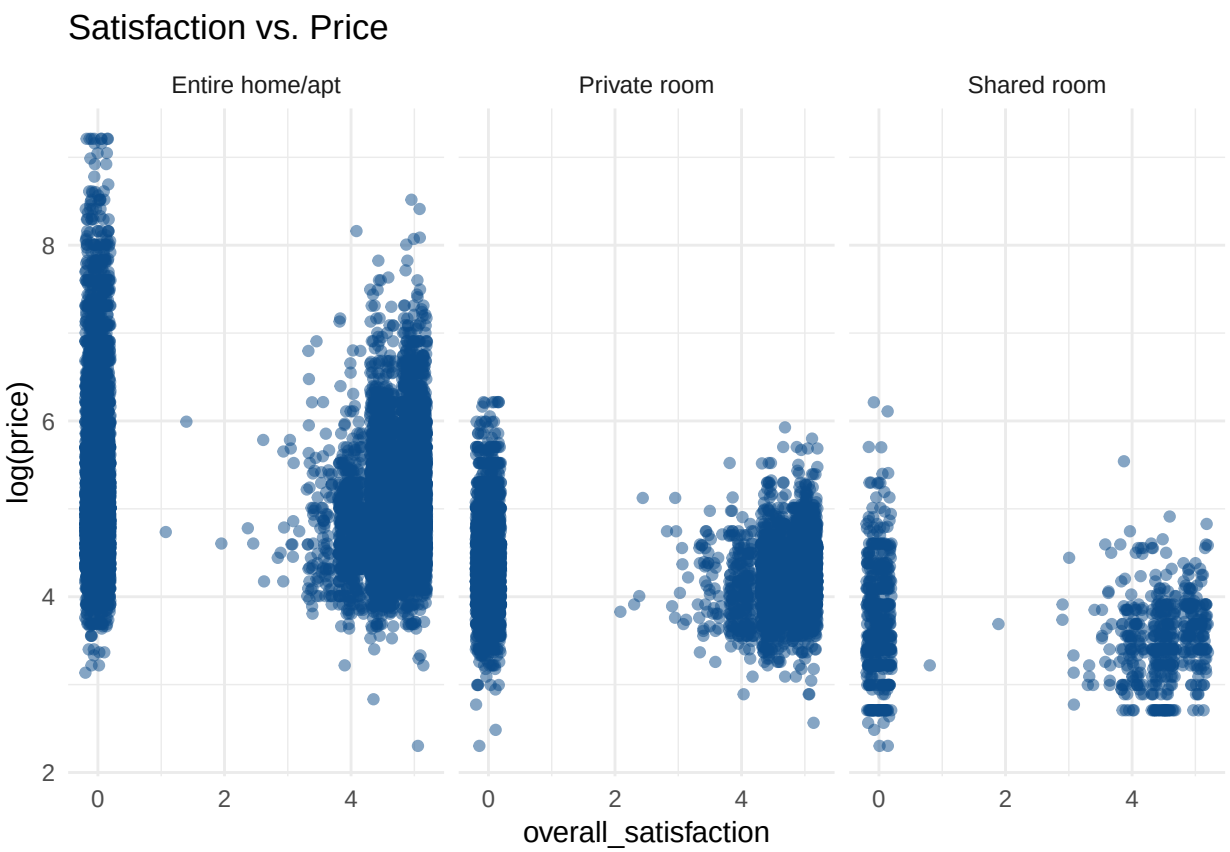
One of the first things we want to look at in terms of variables and their interactions is a correlation plot. We want to focus on the price as it is the response variable we are interested in. From the figure above, we can see that there is a strong correlation between price, accommodations, and bedrooms. We can also see a correlation between bedrooms and accommodations, however, we suspect that these things are all trivially related. Nonetheless, we will consider them in the models.



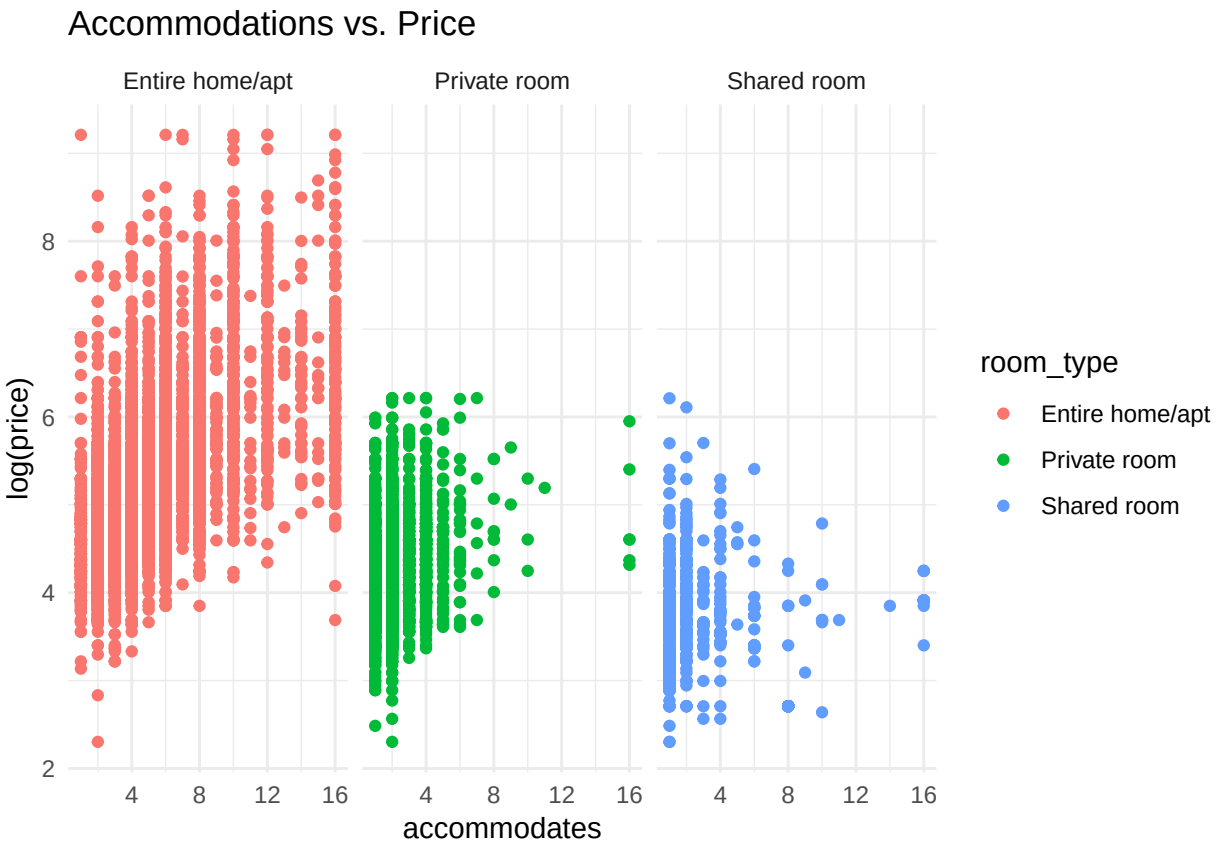
In the figure above, we can see that there are a couple of neighborhoods with significantly more counts than the others. Cities like Venice, Hollywood, Long Beach, and Santa Monica are all some of the most popular cities in Los Angeles. Originally, there were 202 neighborhoods, but since we only wanted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



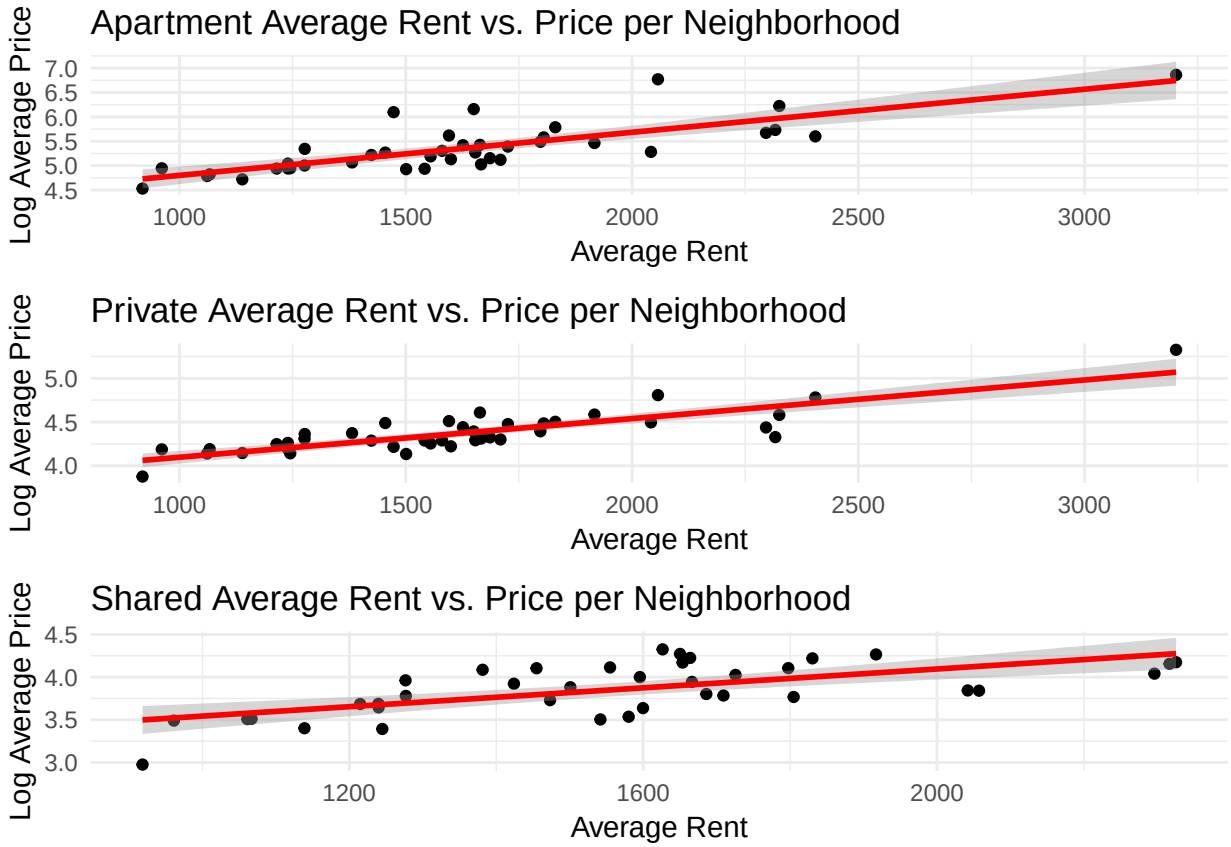
The figure above shows a distribution of pricing, separated by room types, in relation to the number of reviews left for each listing. As we can see from the visual, a large majority of the data points are concentrated around 0, regardless of room type. This indicates that the majority of people did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



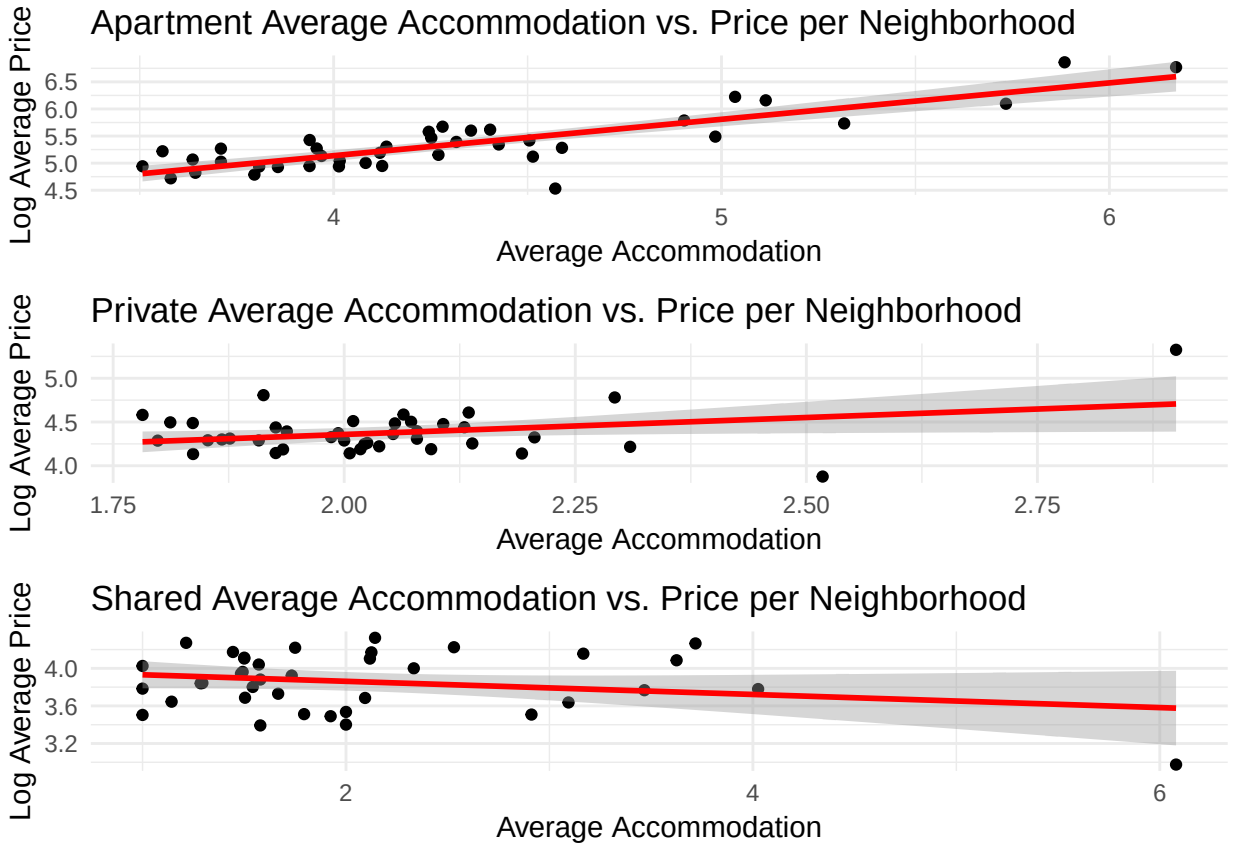
Looking at the figure above, we can see that each of the types of rentals have a significant number of 0's in overall satisfaction. These 0's represent surveys that were left unanswered, however, prices tend to increase as the satisfaction increases in all types of rentals. We can also see that the prices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



From the figure above, we can see that there is a positive trend between accommodation and price only in the home and apartment rentals. This makes sense as private rooms and shared rooms won't be able to house as many people as an entire home or apartment. Therefore, we can see that accommodations for home and apartment range between 1 and 16, whereas accommodations for private and shared rooms are clustered between 1 and 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 averages of the prices listed for each neighborhood and aggregated all these observations into one observation. Then we compared them to the respective average rent for that particular area. From the figure above, we can see that there is a positive trend between the rent and the price of the listing. The data is relatively homogeneous for all three rental types which reflects the positive trend. Logically, this makes sense as the more luxurious an area, the more expensive a listing will be.



Next, we want to look at is accommodations as it is closely related to pricing. In the figure above, we can see that for apartments/homes, there is a clear positive trend between the number of accommodations and the prices for the listing. However, looking at the Private Rooms and Shared Rooms, the data is more heterogenous, making it more difficult to draw any relationship between the prices and accommodations. These patterns are reflective of the trends featured in the accommodations vs. price exploratory figure described earlier in the report.

Methods

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

Outlier Detection

Given that we run the model twice, we will see that the second set of models perform signficantly better than the first set in both R-squared values and AIC values. This can be attributed to the use of Cook's Distance to detect outliers. By determining the influential points in each dataset, we were able to remove them and create a model more representative of the general population of points.

Linear Regression

We start off modeling 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 were done to help improve model performance.

Log Linear Regression

If we look at the linear regression model, we can see there are outliers within the data. In fact, this can be seen in some of the exploratory plots as well. This is a combination of multiple variables, but if we look at the price variable, we can see that for certain neighborhoods, there are extremely large values and in order to combat this, we apply a log transformation to the prices. Afterwards, we can see that the model performs significantly better by evaluating both the residual plots and the QQ plots.

Log Linear Regression with Interaction

If we go back to the correlation plot, 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 within the 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 for three separate room type datasets, the models we selected as the most optimal were:

- **Apartment/Home: Log Linear with Interaction**
- **Private Room: Log Linear Model**
- **Shared Room: Log Linear with Interaction**

Table 2 is a chart of the R-squared values for the raw data, without removing outliers.

Table 2: Raw Data R-Squared Values

	Linear Model	Log Linear Model	Log Linear with Interaction
Apartment/Home	0.3039	0.6017	0.6033
Private	0.2190	0.2215	0.2218
Shared	0.1781	0.3069	0.3069

In the model checking section, we will see there are significant outliers in the data. By removing these outliers and plotting the results, we can see that our models improve significantly. Table 3 is the chart of R-squared values for the data with outliers removed.

Table 3: No Outlier R-Squared Values

	Linear Model	Log Linear Model	Log Linear with Interaction
Apartment/Home	0.5326	0.6700	0.5620
Private	0.2920	0.3211	0.8058
Shared	0.5620	0.3198	0.8058

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/homes 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

Table 4: Linear Regression AIC

	Linear Model	Log Linear Model	Log Linear with Interaction
Apartment/Home	184545.0	11078.00	11027.00
Private	55840.0	1081.30	1130.30
Shared	5223.9	-288.55	-288.55

Next we look at the multilevel AIC values

Multilevel

Table 5: Mutlilevel AIC

	Linear Model	Log Linear Model	Log Linear with Interaction
Apartment/Home	184545.0	11078.00	11027.00
Private	55840.0	1081.30	1130.30
Shared	5223.9	-288.55	-288.55

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. Using the AIC and R-squared values, 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/Home

- Intercept: The average price for an apartment, when all other variables are 0, is exp(-75.21) or 2.17 x 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 x 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

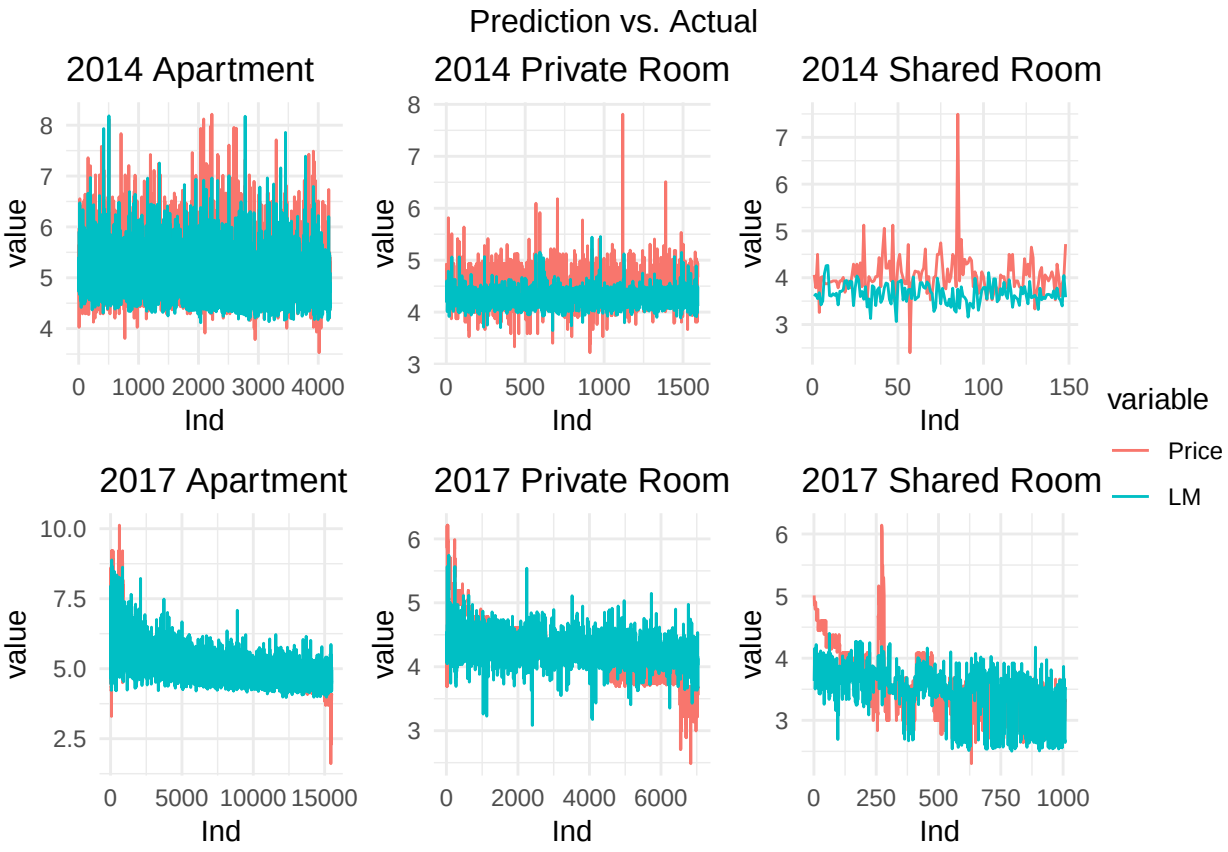
We look at 4 models for each of the 3 different room types: apartment, private room, and shared room. Please refer to the *Appendix* for the models.

First, we will take a look at the residuals vs. fitted plots. From the figures, we can see that the linear models had several outliers that deviated away from the center of 0. However, applying log 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.

Despite these few outliers in these "cleaned" datasets, looking at the Normal Q-Q plots, we can see that the data emulates something close to a normal distribution. Before we cleaned and removed these outliers using Cook's Distance, we could see a stronger tail at each end of the Normal Q-Q plots.

Discussion

Implication



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 apartment/homes rental and shared rooms rental types, while the log linear model was better suited for the private rooms. Although we included an interaction term, there are very slight differences between the log-linear and the interaction models. For the most part, they both yield the same results.

If we look at figure 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 2017 datasets, which are given in red. Each of the best models were utilized for their respective room type(log linear with interaction for apartments/homes and shared rooms, log linear for private rooms). Although it makes no sense from a business standpoint to predict 2014 prices using a 2016 model, we utilize it in order to provide a reference point for the differences between the 2016 and 2017 predictsion. Looking at the plots, we can see a difference between the plots for 2014 and 2017. On one hand, the predictions for 2017 are much more in line with the actual prices from the original dataset. There are clear outliers, however, the predictions seem to fit the generalized population of data well as we don't see as many red lines above the blue lines. On the other hand, if we look at the 2014 predictions, we can see that the model doesn't fit as well, as the majority of the blue predictions sit below the actual listing prices in 2014. This could be an indication that the way AirBnb priced 2016 listings was still the same in 2017, whereas it had a different pricing criteria in 2014. Logically, 2017 is closer to 2016, so the way listings are priced should be similar, whereas, compared to 2014, there is more probability that Airbnb could've propose new asepcts as to how they wanted to change listing prices.

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 were 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

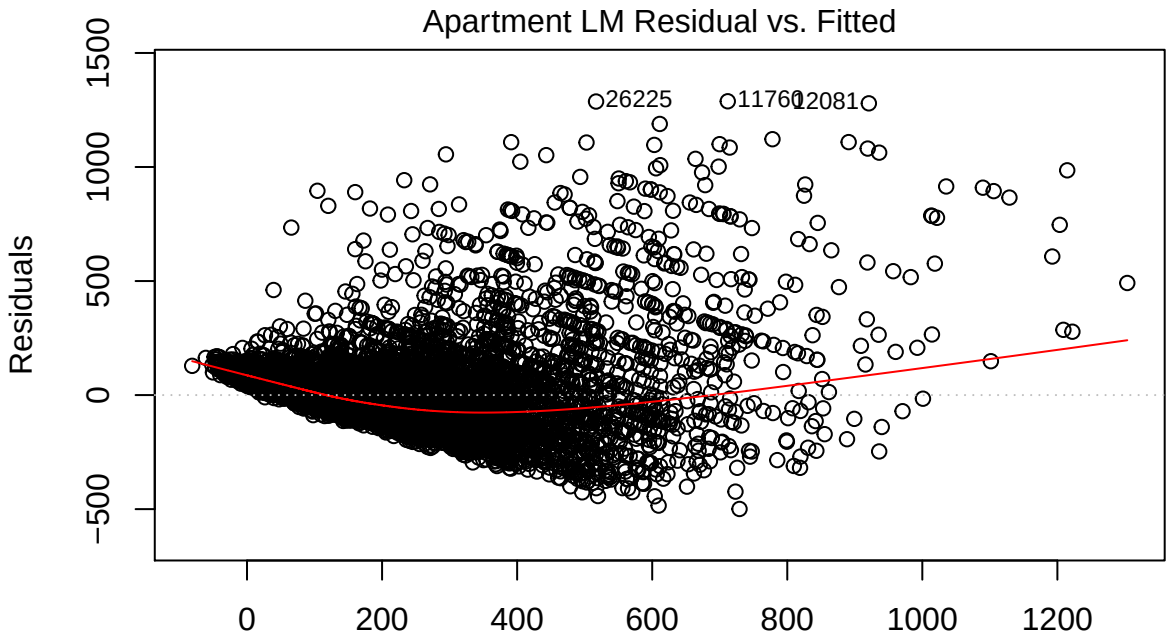
[2] Professor Yajima's RStudio Tutorials

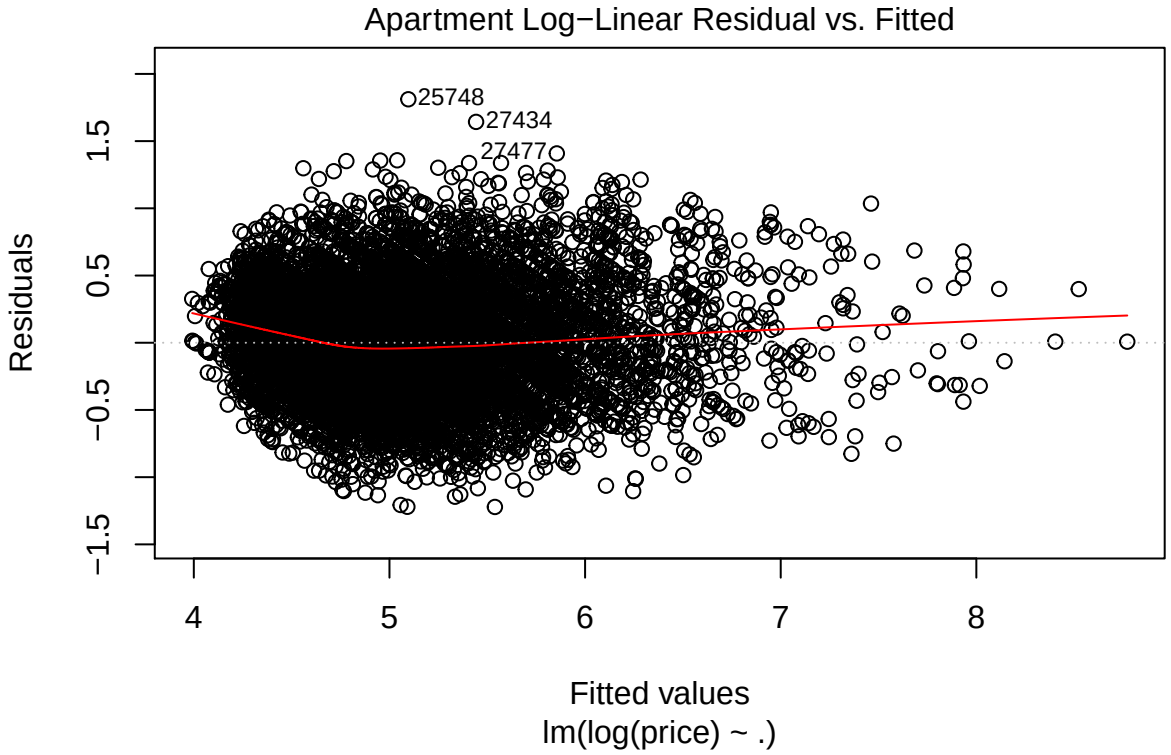
Appendix

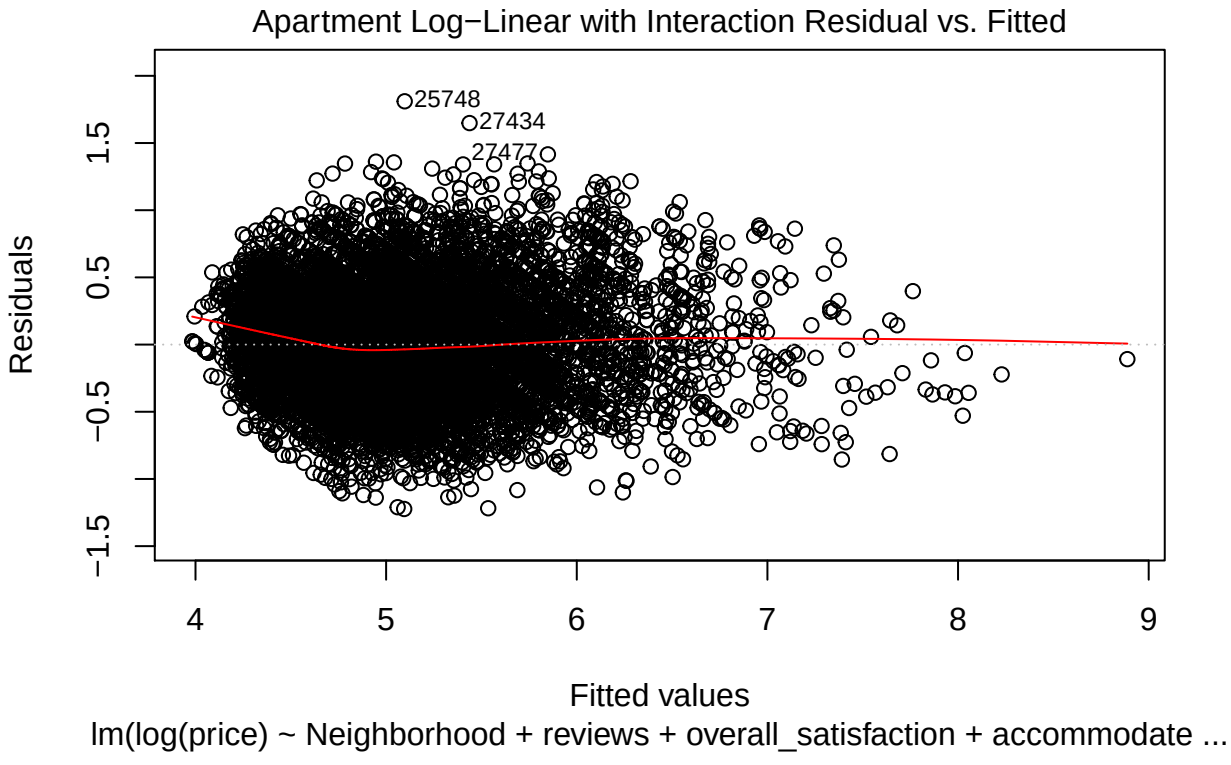
Model Check Figures:

Residual vs. Fitted Plot

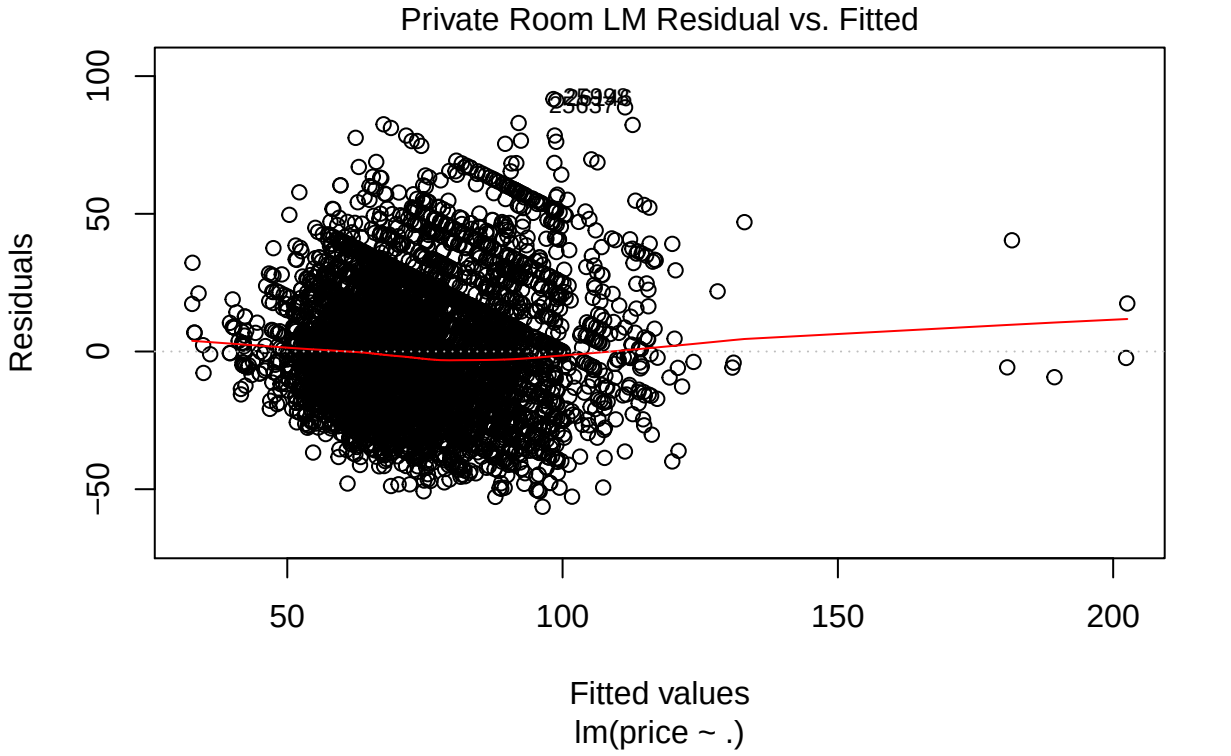
Apartment/Home

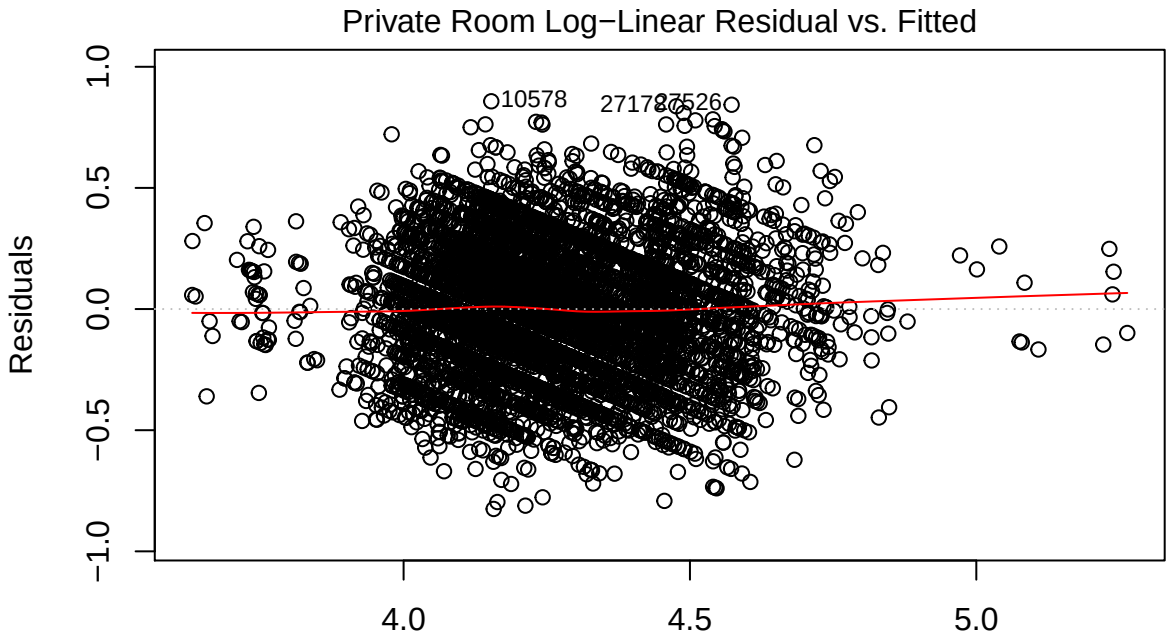

Apartment LM Residual vs. Fitted
Residuals
1500
1000
500
0
−500
26225
11760
12081
0
200
400
600
800
1000
1200
Fitted values
lm(price ~ .)

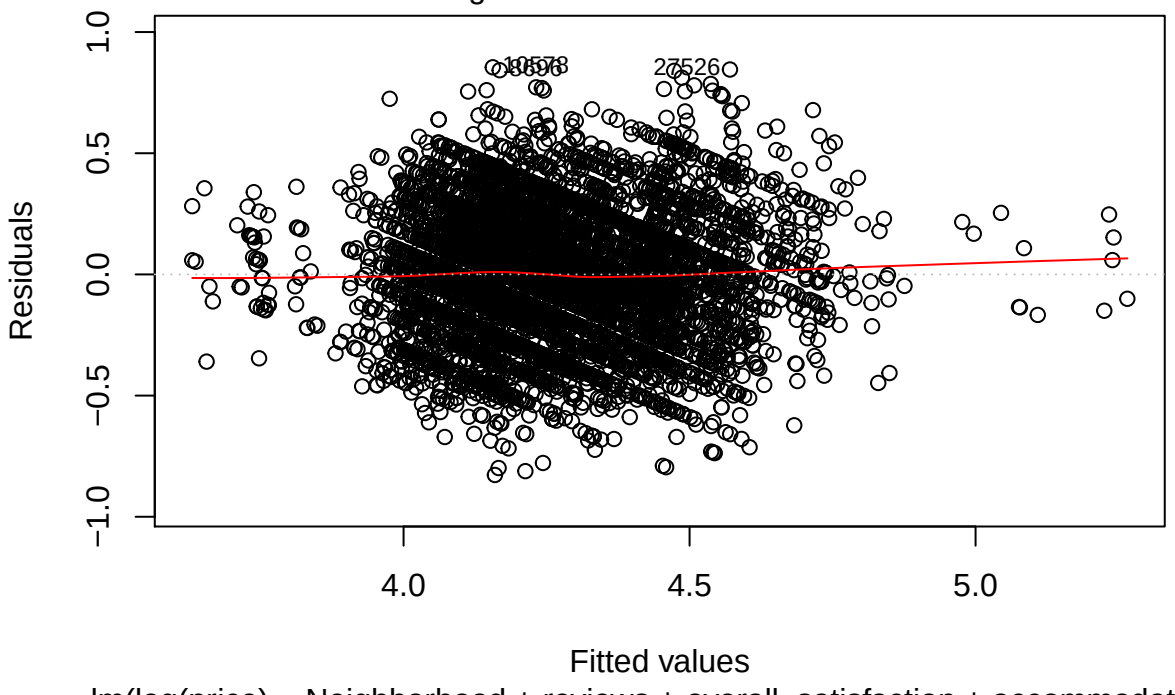

Apartment Log−Linear Residual vs. Fitted
Residuals
1.5
0.5
−0.5
−1.5
25748
27434
27477
4
5
6
7
8
Fitted values
lm(log(price) ~ .)



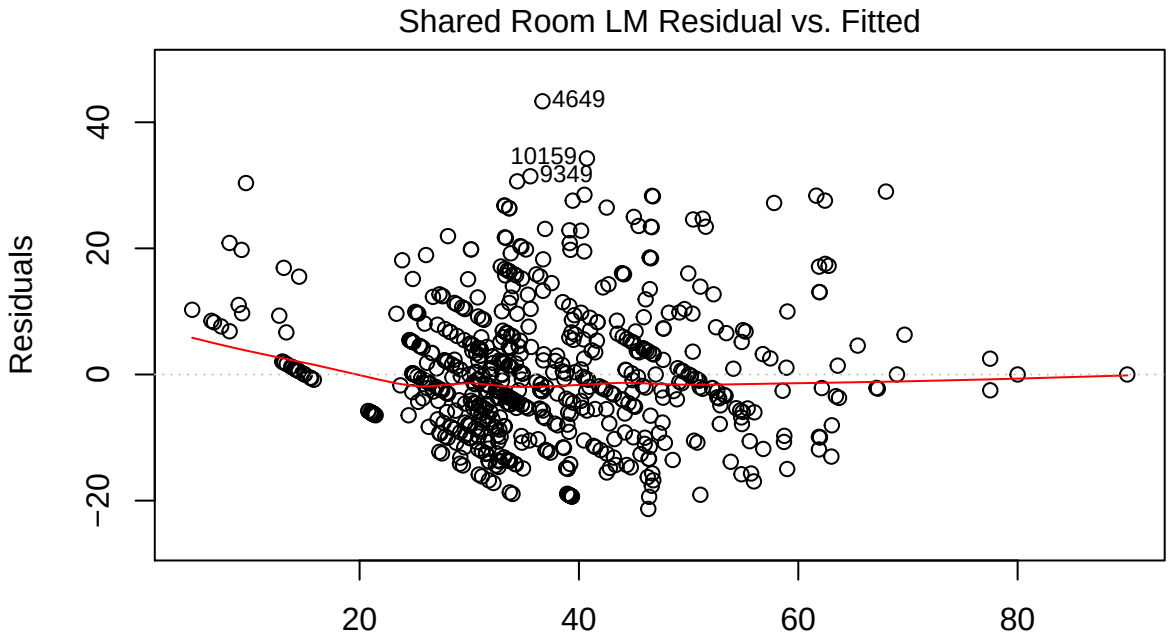
Private Room

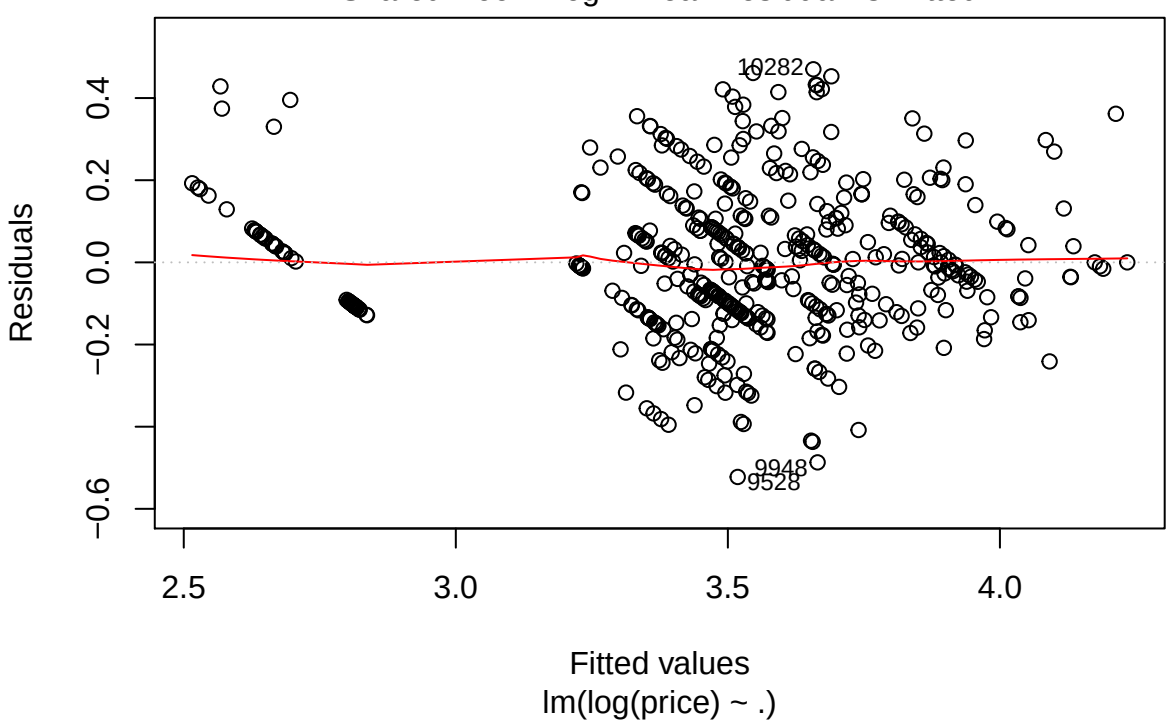



Private Room Log−Linear Residual vs. Fitted
Residuals
1.0
0.5
0.0
−0.5
−1.0
10578
27526
4.0
4.5
5.0
Fitted values
lm(log(price) ~ .)


Private Room Log−Linear with Interaction Residual vs. Fitted
Residuals
1.0
0.5
0.0
−0.5
−1.0
27526
4.0
4.5
5.0
Fitted values
lm(log(price) ~ Neighborhood + reviews + overall_satisfaction + accommodate ...

Shared Room

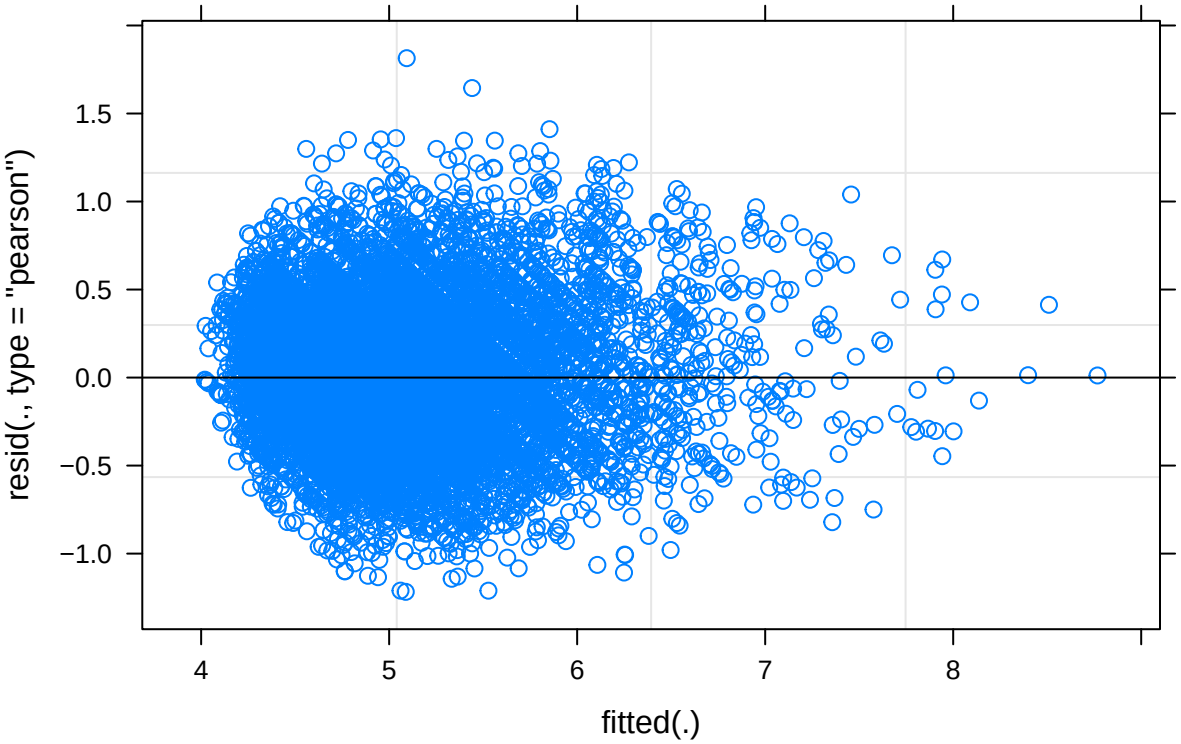

Shared Room LM Residual vs. Fitted
Residuals
Fitted values
lm(price ~ .)
4649
10159
9349
40
20
0
−20
20
40
60
80


Shared Room Log−Linear Residual vs. Fitted
Residuals
Fitted values
lm(log(price) ~ .)
10282
9948
9528
0.4
0.2
0.0
−0.2
−0.6
2.5
3.0
3.5
4.0

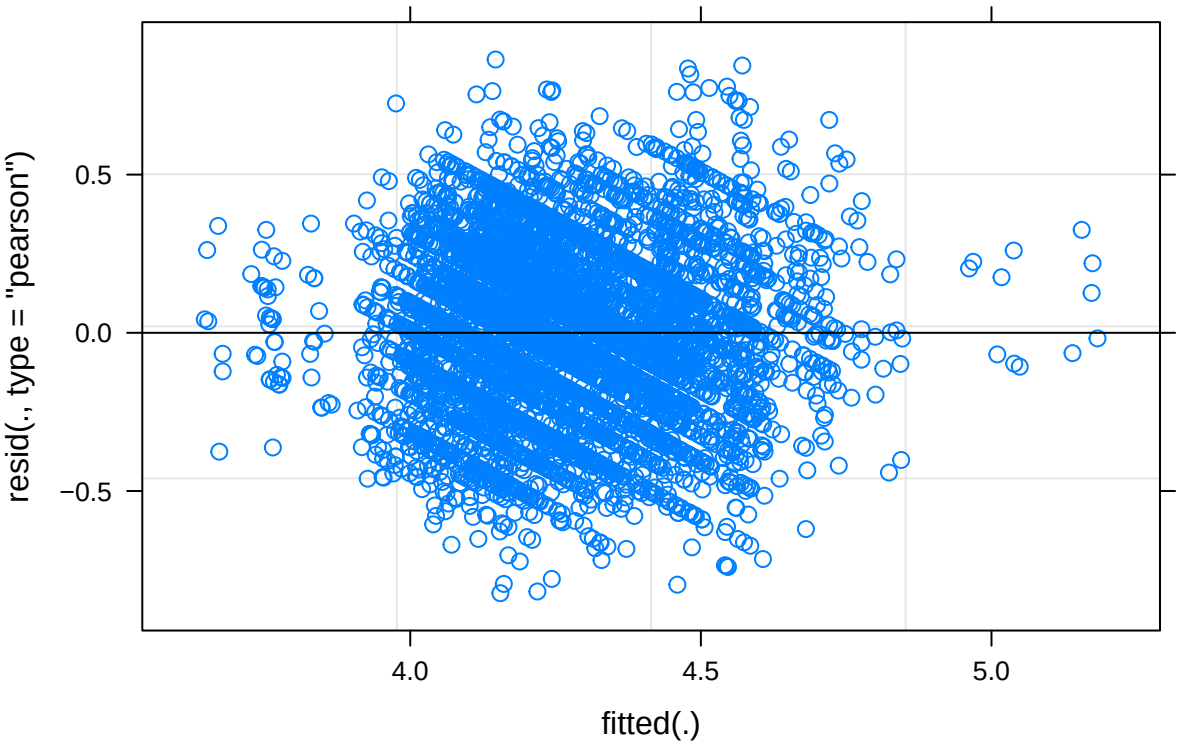
Multilevel Model

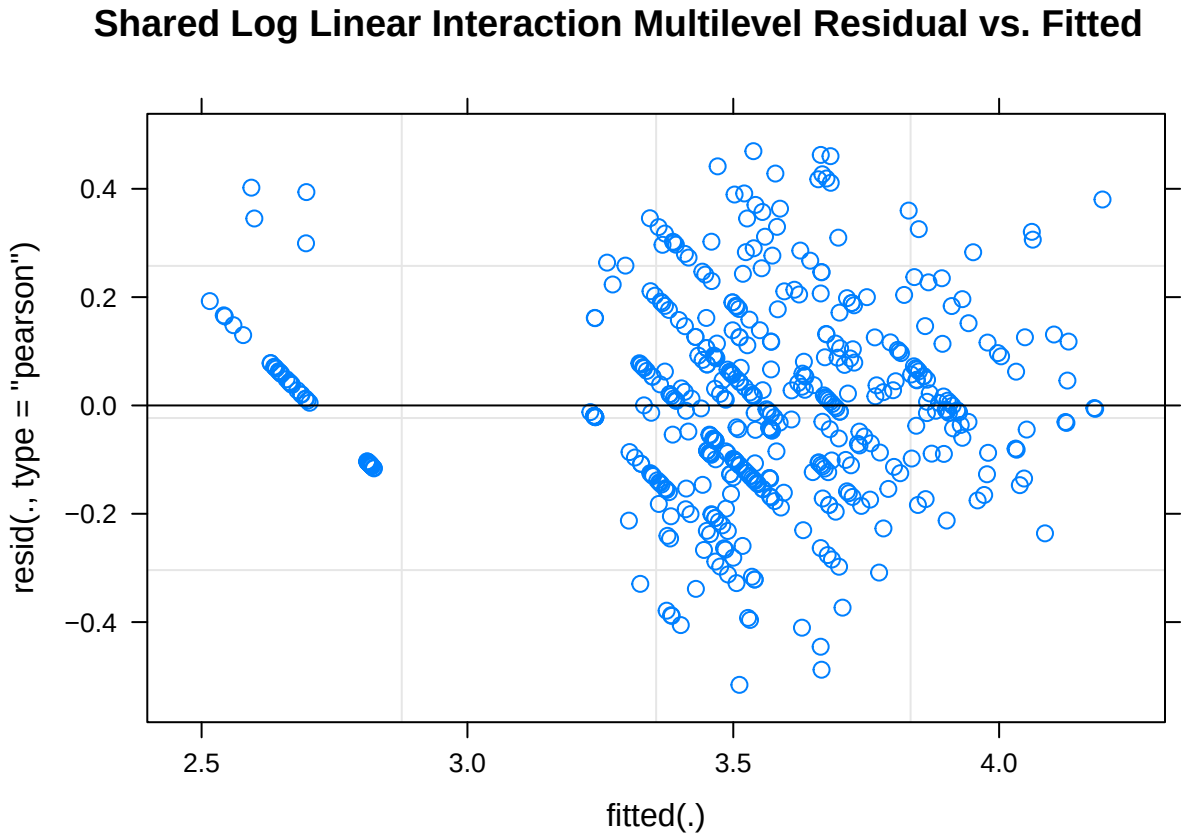
Apartment/Home

Apartment Log Linear Mulilevel Residual vs. Fitted



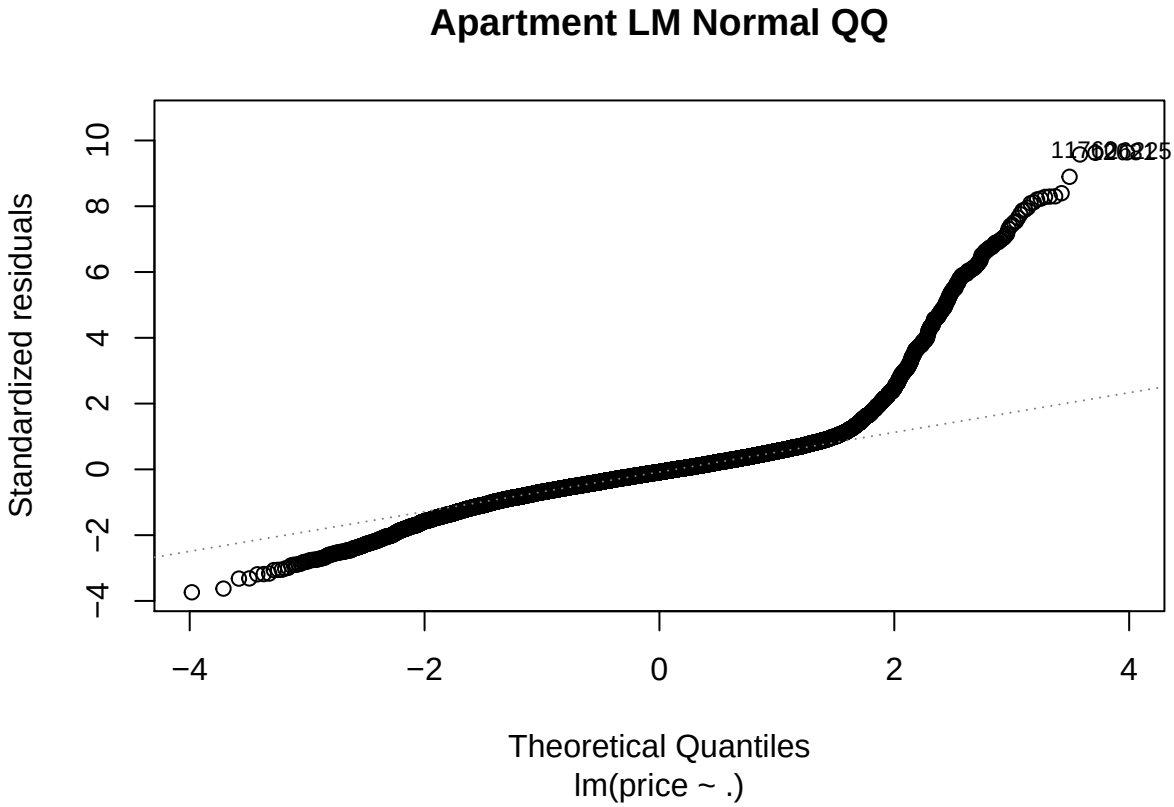
Private Log Linear Multilevel Residual vs. Fitted

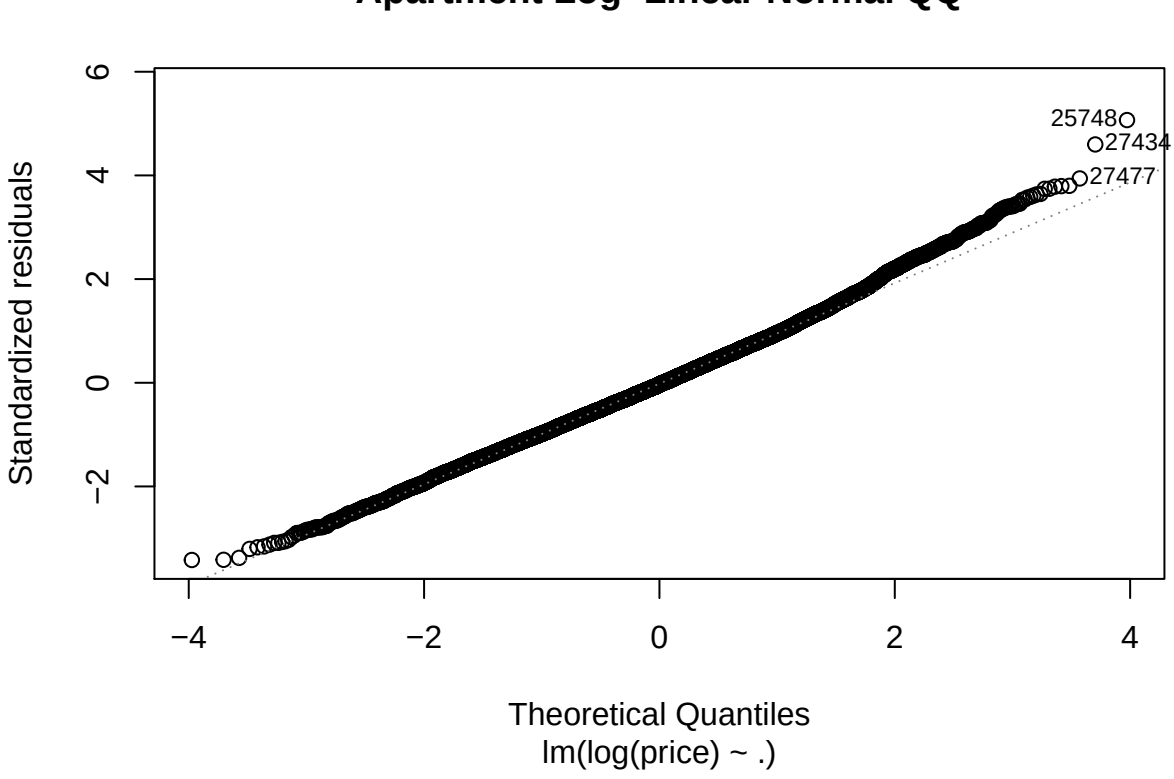


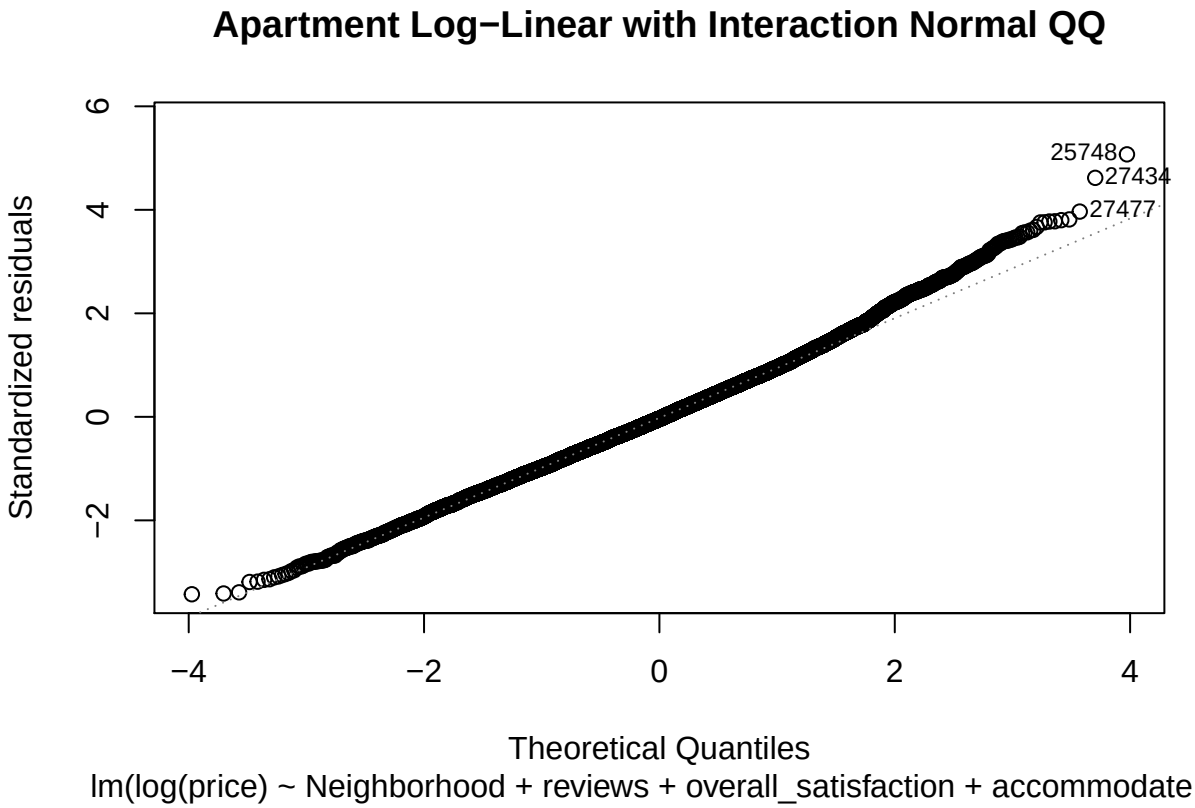


Normal QQ-Plot

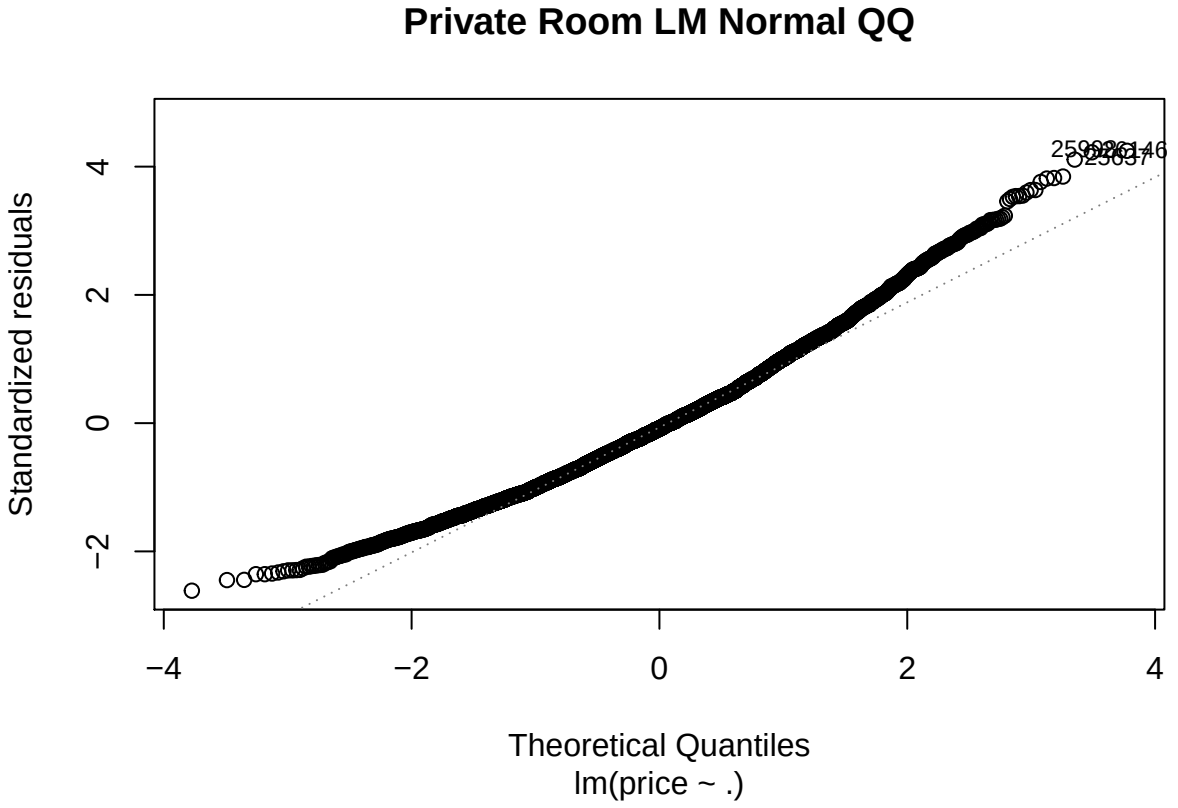
Apartment/Home

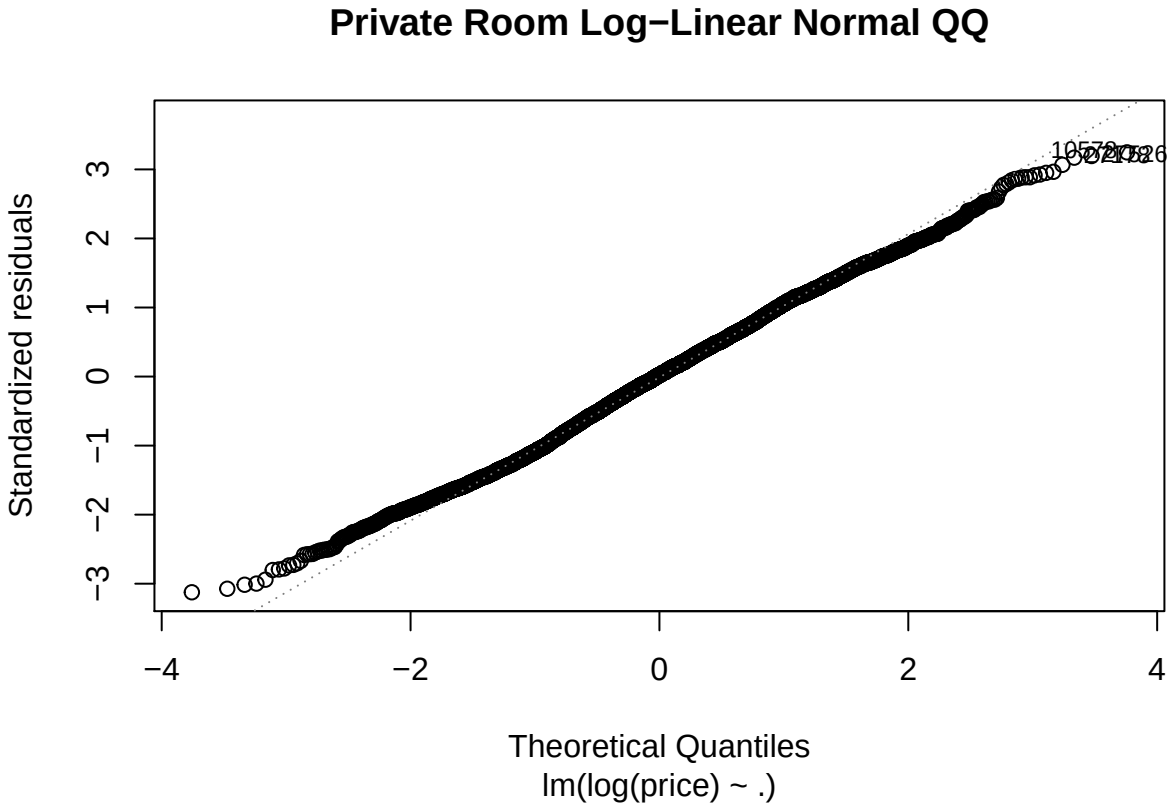

Apartment LM Normal QQ
Standardized residuals
Theoretical Quantiles
lm(price ~ .)

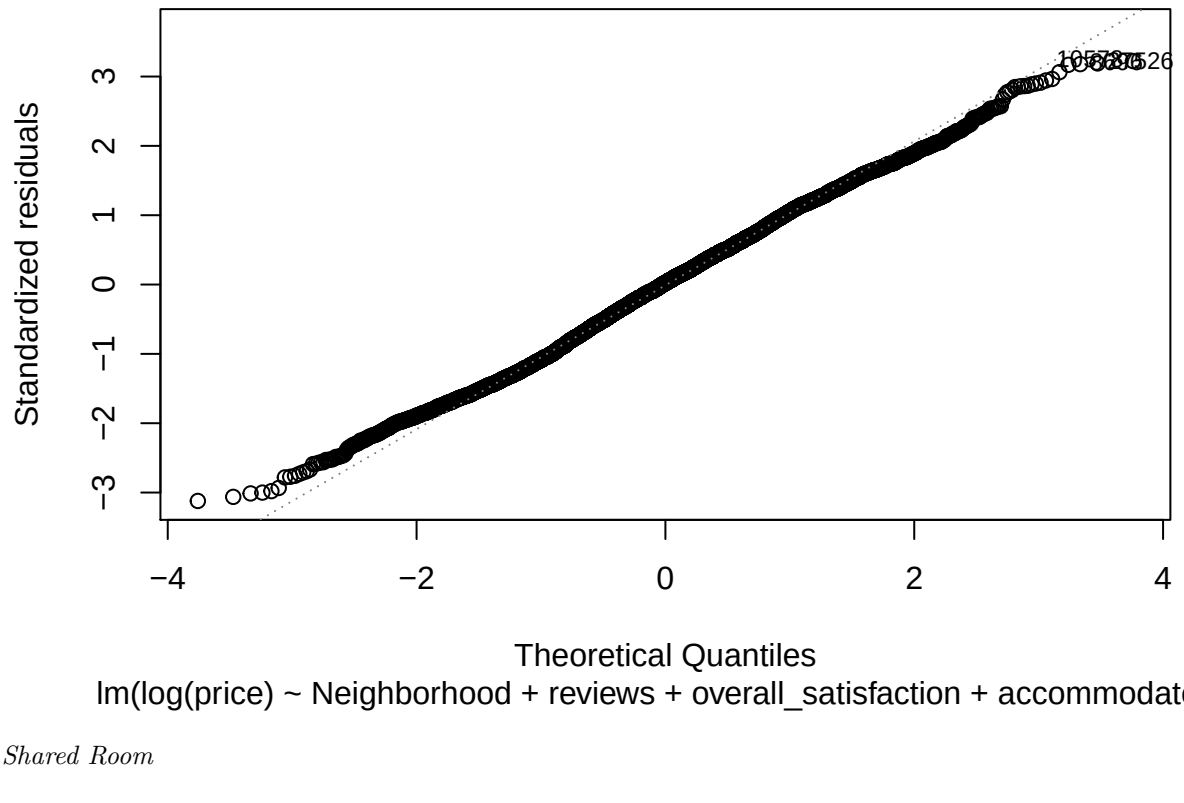

Apartment Log−Linear Normal QQ
Standardized residuals
25748
27434
27477
Theoretical Quantiles
lm(log(price) ~ .)



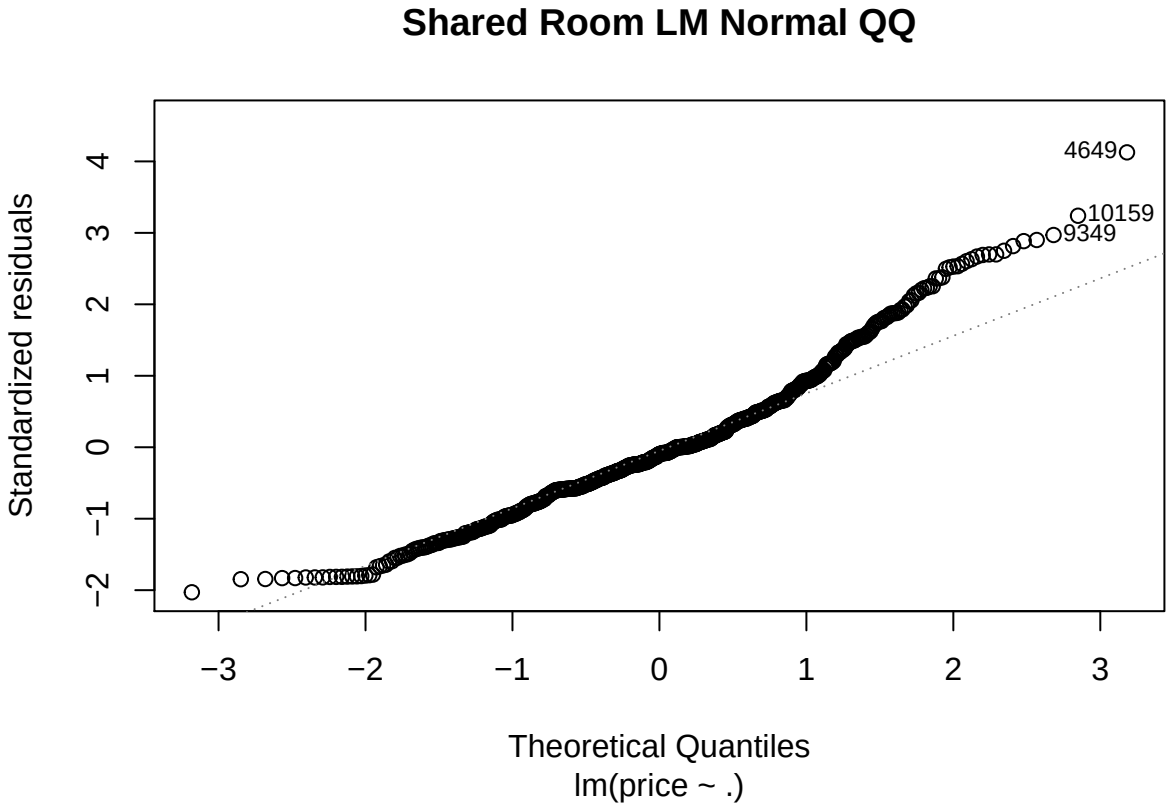
Private Room

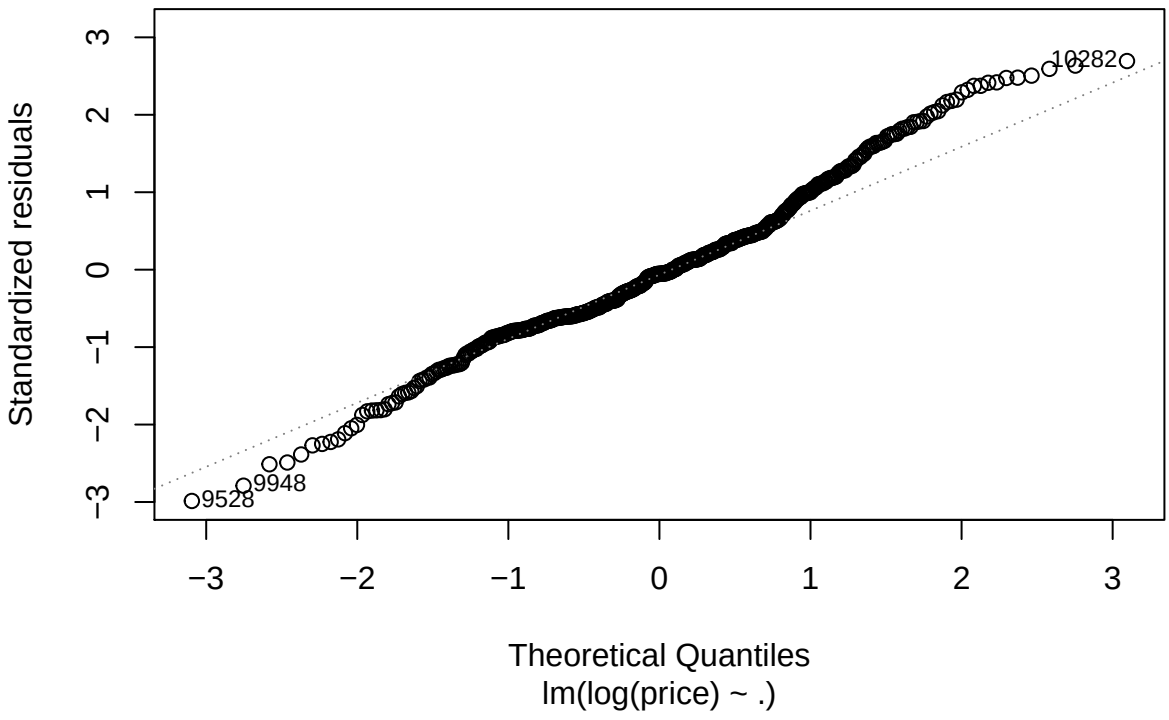


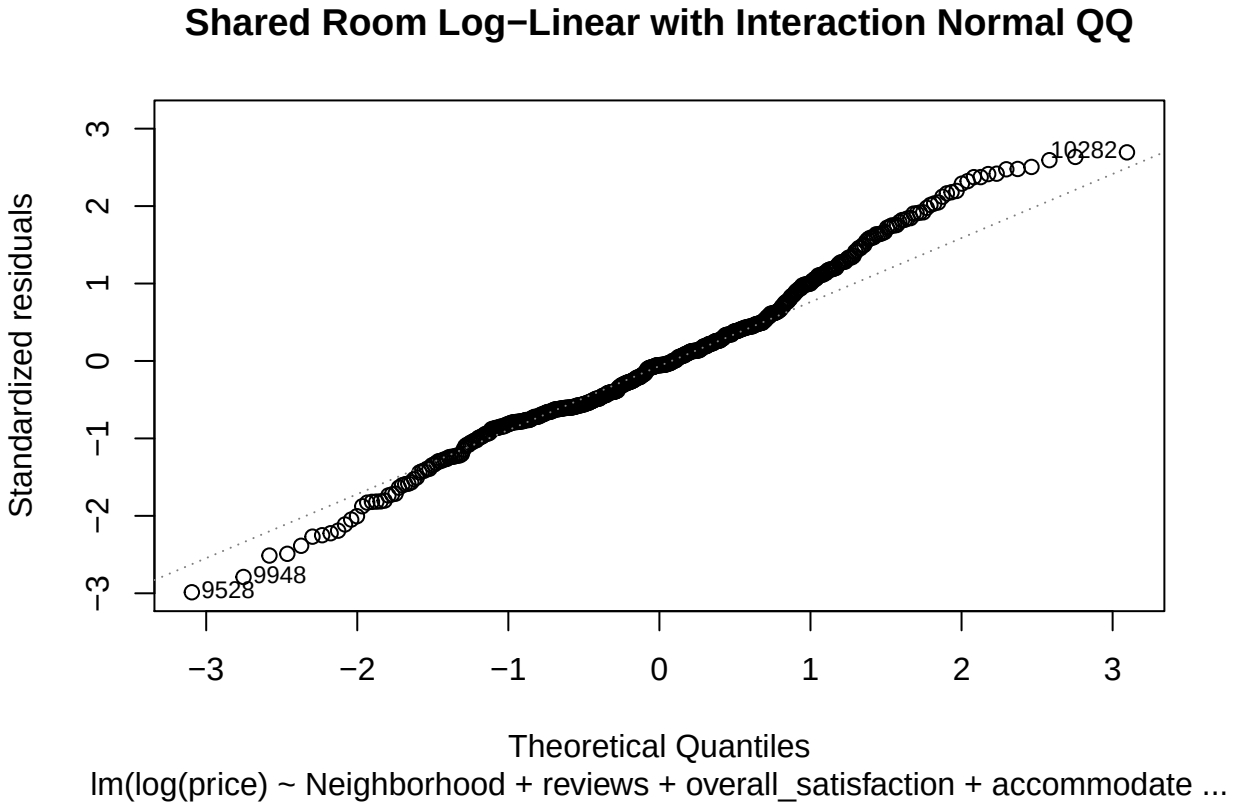

Private Room Log−Linear Normal QQ
Standardized residuals
−3
−2
−1
0
1
2
3
−4
−2
0
2
4
Theoretical Quantiles
lm(log(price) ~ .)


Private Room Log−Linear with Interaction Normal QQ
Standardized residuals
−3
−2
−1
0
1
2
3
−4
−2
0
2
4
Theoretical Quantiles
lm(log(price) ~ Neighborhood + reviews + overall_satisfaction + accommodate ...

Shared Room


Shared Room LM Normal QQ
Standardized residuals
Theoretical Quantiles
lm(price ~ .)
4649
10159
9349


Shared Room Log−Linear Normal QQ
Standardized residuals
Theoretical Quantiles
lm(log(price) ~ .)
10282
9948
9528


Shared Room Log−Linear with Interaction Normal QQ
Standardized residuals
3
2
1
0
−1
−2
−3
10282
9948
9528
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(log(price) ~ Neighborhood + reviews + overall_satisfaction + accommodate ...