STA302 Final Project Part 3

Ji Hoon Kim Grace Boss, Geon Lim, Dongkeun Jang

2023-12-06

Contribution

Ji Hoon Kim: Results, Formatting R Markdown

Donkeun Jang: Methods

Geon Lim: Discussion

Grace Boss: Introduction and Ethics

Introduction

Our goal is to analyze internal factors of Airbnbs in order to understand how they influence the pricing of listings. Such factors include the number of bedrooms, the number of bathrooms, the rating of cleanliness on a scale from zero to five, whether or not the host was rated 'Superhost' (if 'Superhost', it takes on the value one multiplied by the number of bedrooms, if no takes on value zero), and whether or not the listing was shared or not with a one indicating shared and a zero indicating not shared. These are the factors that we have decided upon for our analysis, as they have come up in many previous analyses as being important factors for pricing as well as being viable for a linear regression analysis, as they complement each other (Chattopadhyay). While there are other important factors aside from those we have decided upon when discussing pricing, we made the decision to analyze only internal factors, despite external factors also having equally important sway, but create a more complicated analysis which is not as viable for a linear regression (Voltes-Dorta). As a result of both lack of data and simplicity, we decided against including such factors, and instead decided to perform a regression that reports solely upon internal factors. It is important to note that there other factors are also present when determining a pricing scheme, and these can take on a theoretical approach as opposed to a purely numeric one, hence we may see inconsistencies in our linear regression as a result of an inability to quantify these theoretical approaches (Kwok).

While many of these sources and many previous analyses tend to focus on a variety of factors that influence pricing schemes and take on more complex modelling techniques to account for these factors, they fail to hone in on one specific area that can influence pricing, and hence we hope to create a more specialized model. This model will contribute to the understanding of those pricing their listings as well as potential customers, as it will allow for them to better understand how the factors that are in the control of the listing provider contribute to the price without requiring potential confounding variables within the estimation. As a result, our analysis aims to answer in a more specific manner how only internal factors can influence the pricing of an Airbnb.

Methods

Initially, our model, comprising four numerical and two categorical variables, violated all four linear regression assumptions and two conditions for multiple linear regression. To address Normality and Linearity issues, we applied the Box-Cox power transformation. However, uncorrelated errors persisted due to data characteristics. The Box-Cox transformation aimed to maintain Normality and Linearity, addressing the building of an effective linear model focused on accurate predictions.

In our modified linear model, we calculated the Confidence Interval (CI) for each slope coefficient, which helped justify the reasonability of these coefficients and enabled hypothesis testing. We then established a 95% Confidence Interval for the mean of the response variables, indicating the likely range of the true mean. Additionally, we computed a 95% Prediction Interval (PI) for actual response values, providing a range of possble value for an actual response.

The next step was regarding the ANOVA (Analysis of Variance) test. The purpose of the ANOVA test was to evaluate whether there is a statistically significant linear relationship for at least one predictor in the model. This was done by calculating the mean squares, specifically the Mean Square Regression (MSreg) and the Mean Square Residual (MSR), to derive the test statistic, F*. We set our confidence level same at 95%, as this is a commonly used benchmark. The test concluded that if the test statistic exceeds the critical value, it indicates a significant linear relationship for at least one predictor.

After conducting the ANOVA test, the next step involved decomposing the model to perform a new hypothesis test. This test was crucial for determining if multiple predictors can be simultaneously removed from the model. Given the ANOVA test's indication of a significant linear relationship with at least one predictor, it was important to identify which predictors have this relationship. By using individual T test, beforehand, we were able to find which variable is significantly related. Realizing that the predictors are significant, here, we used the partial F test. This test assesses whether removing a subset of predictors significantly affects the model's performance. The comparison is made by looking at the Sum of Squares Regression (SSreg) values of the full model and a reduced model. If these values are similar, the Residual Sum of Squares (RSS) will also be comparable, suggesting that a reduced model can also be used. On the other hand, if the RSS of the reduced model is significantly higher than that of the full model, it suggests the full model is more appropriate.

Getting the result acknowledging between the reduced model or full model is better, we further evaluated the goodness of these two data sets. We applied adjusted R-square to see which predictor should be removed or added. When we derive a conclusion stating that one of the model has a higher R-squared and a adjusted R-squared value, it signifies a better model fit, as it suggests that the model explains a greater proportion of the variance and is not overly complex.

After calculating the adjusted R-squared for a particular model and assessing multicollinearity levels, it is essential to address automated model selection. This process aims to determine if there exists a better model than the one we selected through the Partial T-test. Automated selection evaluates model performance using criteria such as the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and adjusted R-squared. Smaller values in these metrics generally indicate a better model fit.

This study adapted three types of automated selection methods:

1. Forward Selection: This method begins with an intercept-only model and then proceed to look at whether we get a better model by adding predictors.

2. Backward Selection: Backward selection starts by beginning with a full model and proceed to see if a smaller model is better by deleted predictors.

3. Stepwise Selection: This method combines elements of both forward and backward selection, allowing us to both add predictiors as well as delete them

For each method, we calculated AIC and BIC values. The model giving the smallest values for these criteria is considered the best fitting model. By comparing AIC and BIC across different model-building approaches, we aimed to identify the most effective model in terms of predictive accuracy.

Within the process of getting the result, it was crucial to recognize problems related with predictors by assessing multicollinearity. This occurs when two or more predictors in a regression model are highly correlated. This correlation leads to inaccurate coefficient estimates, conflicting significance levels, and inflated variance estimates, which can compromise the model's reliability. Therefore, to measure the level of multicollinearity, it was important to address the value of Variance Inflation Factor (VIF) as it explicitly quantifies the impact that the multicollinearity between predictors has on the variance. We specifically measured the degree to

which the variance has been inflated due to multicollinearity. A VIF value exceeding 5 often indicates a concerning level of multicollinearity.

Additionally, the analysis included assessing problematic observations to ensure overall accuracy in the dataset. This involved identifying observations that disproportionately influence the model. These are three key measures that were used:

1. Leverage Points: These are observations that have a significant impact on the model as a whole.

2. Outliers: Observations that significantly differ from the rest of the data points.

3. Influential Points: These are identified using Cook's D, DFFITS, and DFBETAS, each measuring the influence of an observation on the fitted values and estimated coefficients.

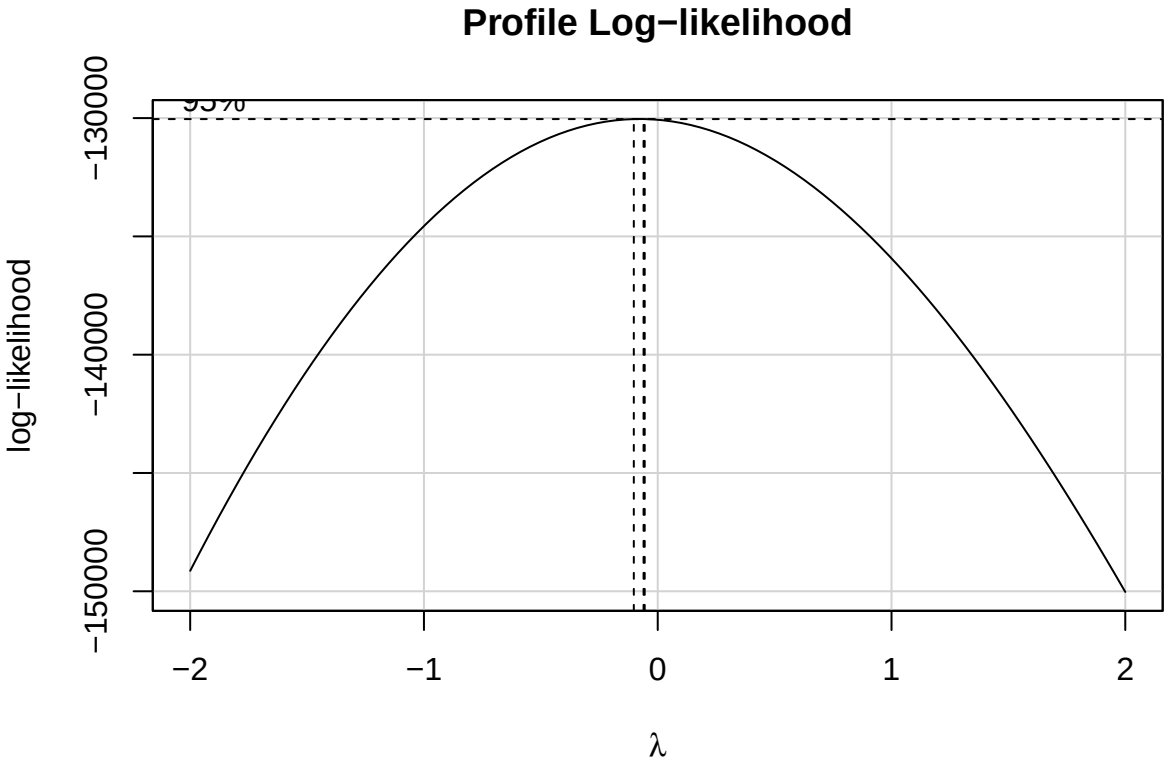
Calculating and examining these measures allowed us to identify potentially problematic observations that could disproportionately influence the overall dataset. This evaluation played a crucial role in ensuring the accuracy and reliability of the final model.

Consequently, in this research, we successfully developed a Multiple Linear Regression model closely aligned with our research question, employing a systematic and stastistical approach of following ANOVA test to Partial T-test. Important section to our methodology was addressing multicollinearity among predictors, crucial for the model's accuracy and interpretability. Through the use of Adjusted R-squared, Variance Inflation Factor (VIF), and automated selection methods, we refined our model, ensuring it was both statistically credible and relevant to our research question. We also managed influential data points like leverage points and outliers to enhance the model's reliability. Therefore, our efforts resulted in to find a best MLR model that found specific predictors that contributed to the overall price of Airbnb listings in Canadian Dollars.

Results

MODULE 4

In part 1 of this assignment, we saw that all 4 linear assumptions were violated as well as the 2 MLR conditions; 1. Conditional Mean Response and Predictor vs. Predictor relationship. Our first limitation occurs here because there is no way to address uncorrelated errors because that is the nature of the data. However, we can address Normality and Linearity assumption through applying Box-cox transformation on the predictors and the response variable. This transformation will likely correct constant variance assumption as well, if not, we may apply variance stabilizing transformation on the response variable to address constant variance assumption.



We see that lambda is very close to 0, implying that we should use the natural log function on the response variable.

Due to the result above, I created a new column called "logPrice" which is the log value of the price of the AirBnB per night.

```
## [1] "Estimate power"
```

```
##           Bedrooms          Bathrooms Cleanliness_Rating
##          -3.039405          -1.943660          10.030199
```

We can only apply boxcox transformation on numerical predictors and these values must be strictly positive otherwise certain functions like log() cannot be applied for transformation. Price is strictly positive thus there were no issues above, but because there are some properties with 0 bathrooms and some have cleanliness rating of 0/5, these values were removed before applying the transformation. During this process 38 data points were removed which is negligible since less than 0.3% of the dataset. As shown above, the value of the power for the transformation was approximately -3 for Bedrooms, -2 for Bathrooms and 10 for Cleanliness rating

```
##
## Call:
## lm(formula = logPrice ~ I(1/Bedrooms^3) + I(1/Bathrooms^2) +
##     Cleanliness_Rating^10 + Entire_Place + Superhost:Bedrooms,
```

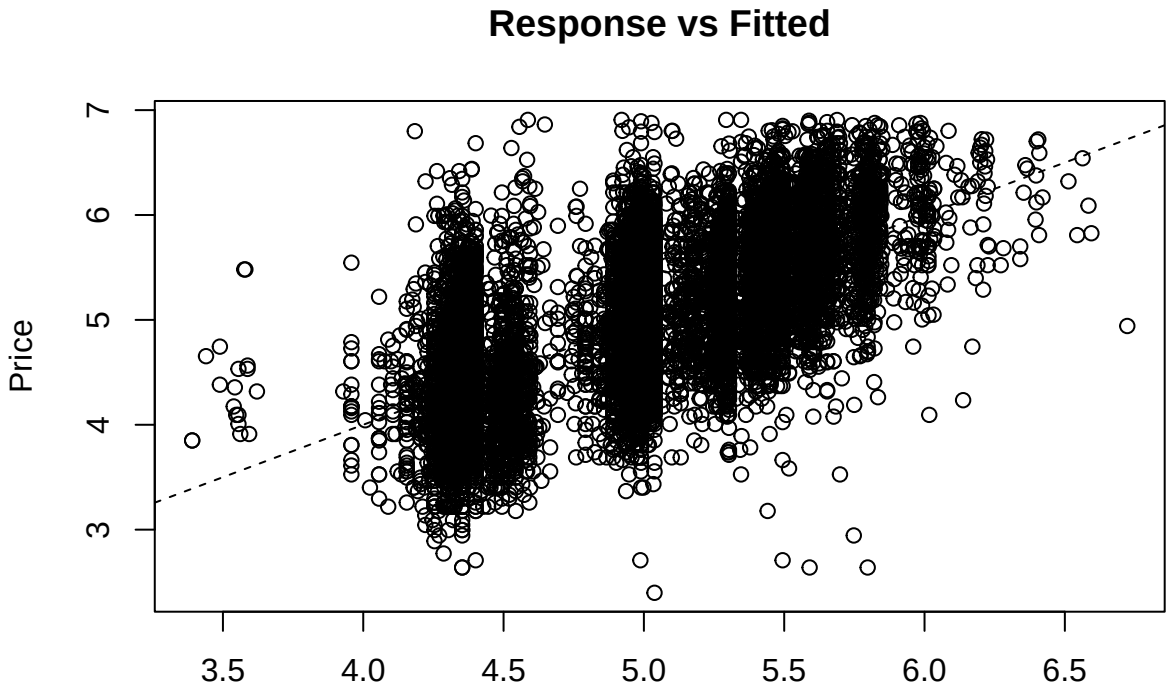
```
##     data = final_data1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.15836 -0.38454 -0.03676  0.36595  2.61531
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           4.805952   0.058352  82.362   <2e-16 ***
## I(1/Bedrooms^3)      -0.196419   0.022615  -8.685   <2e-16 ***
## I(1/Bathrooms^2)     -0.254893   0.015904 -16.027   <2e-16 ***
## Cleanliness_Rating    0.098712   0.009886   9.985   <2e-16 ***
## Entire_PlaceShared   -0.637169   0.010677 -59.678   <2e-16 ***
## Superhostf:Bedrooms   0.141515   0.011444  12.366   <2e-16 ***
## Superhostt:Bedrooms   0.189635   0.012072  15.709   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.538 on 14245 degrees of freedom
## Multiple R-squared:  0.4336, Adjusted R-squared:  0.4333
## F-statistic:  1817 on 6 and 14245 DF,  p-value: < 2.2e-16
```

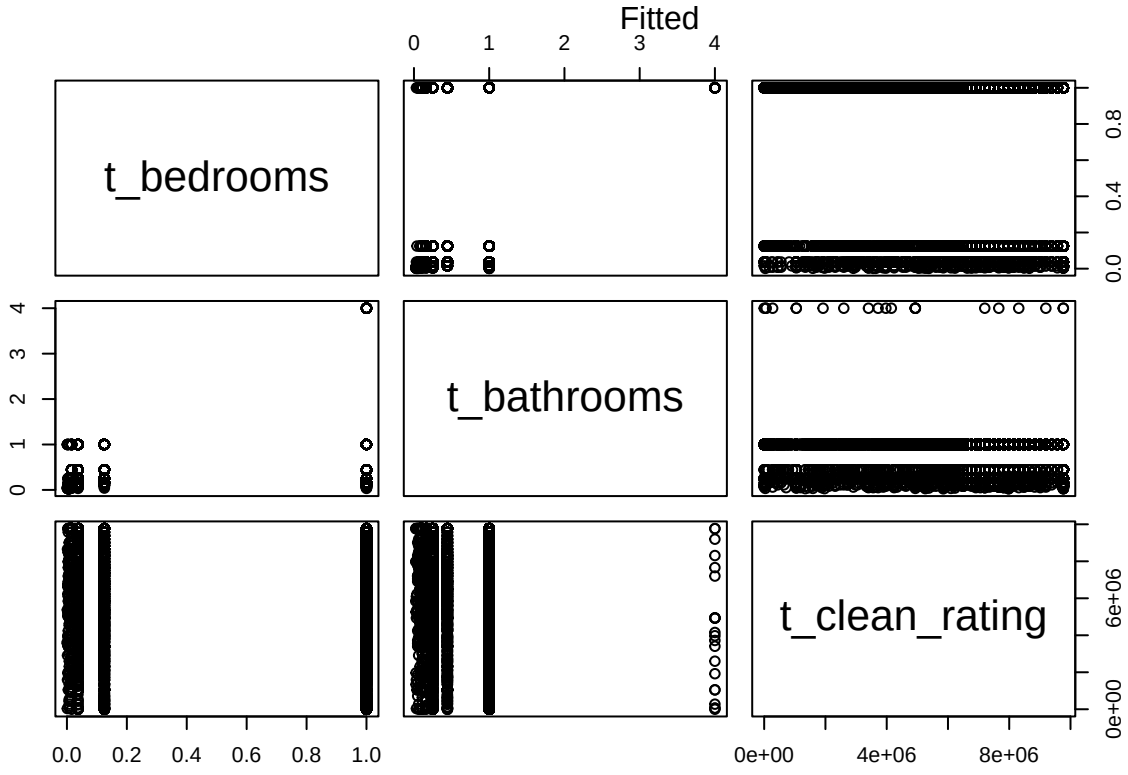
Here this is the fixed model of the original model with the power transformation applied.

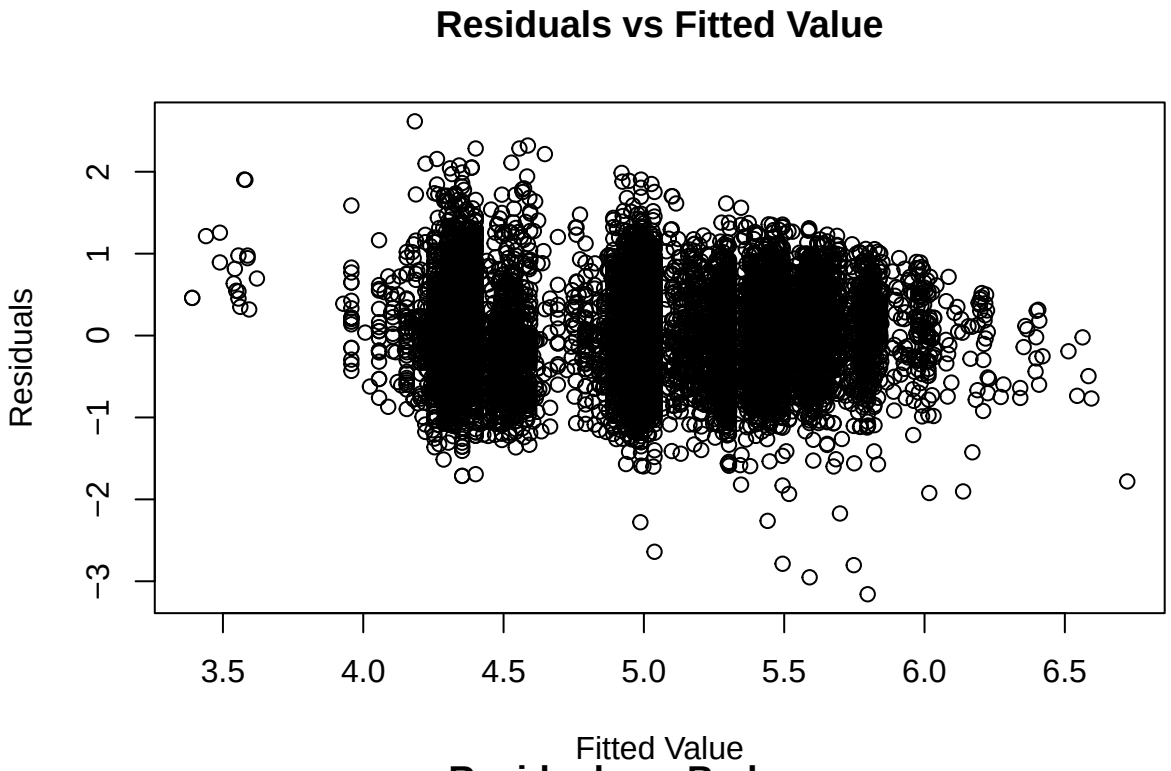
$$\text{logPrice } (Y) = 4.805952 - 0.196419(\text{Number of Bedrooms})^{-3} - 0.254893(\text{Number of Bathrooms})^{-2}$$

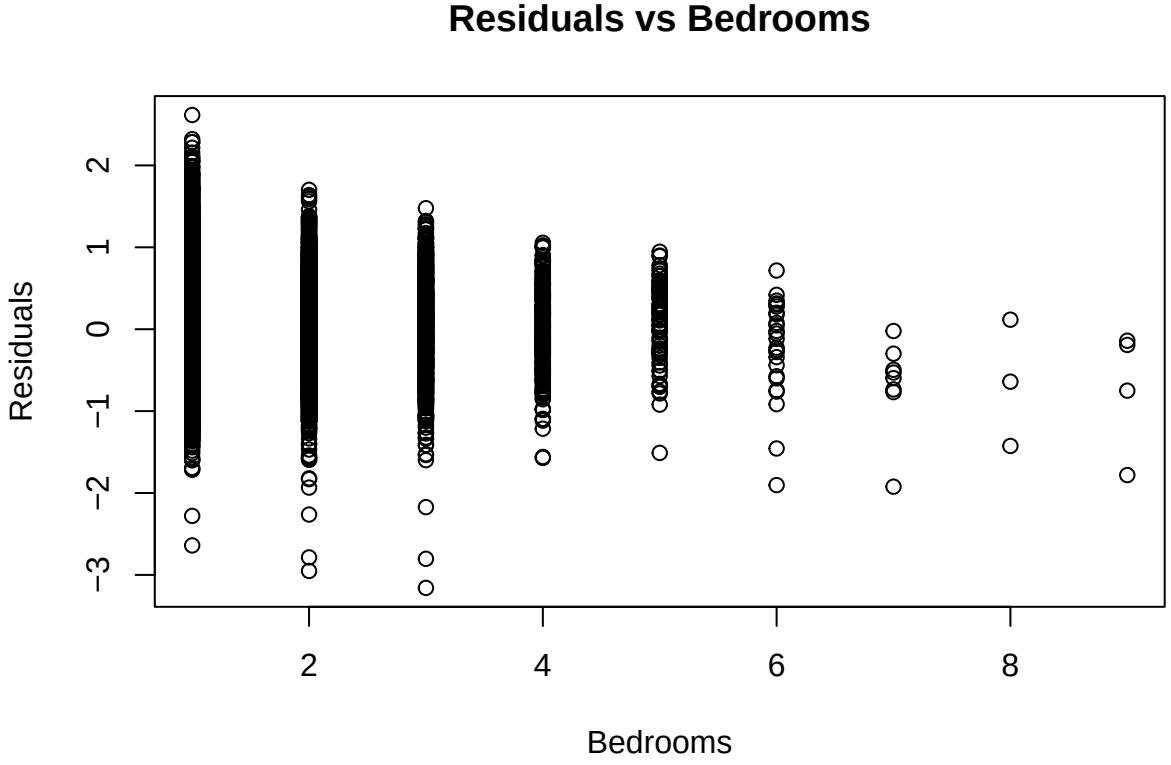
$$+0.098712(\text{Cleanliness Rating})^{10} - 0.637169\ \mathbf{I}(\text{Entire Place} = \text{Shared}) + 0.189635(\text{Superhost}=\text{`t'} : \text{Bedrooms}) + \hat{\epsilon}$$

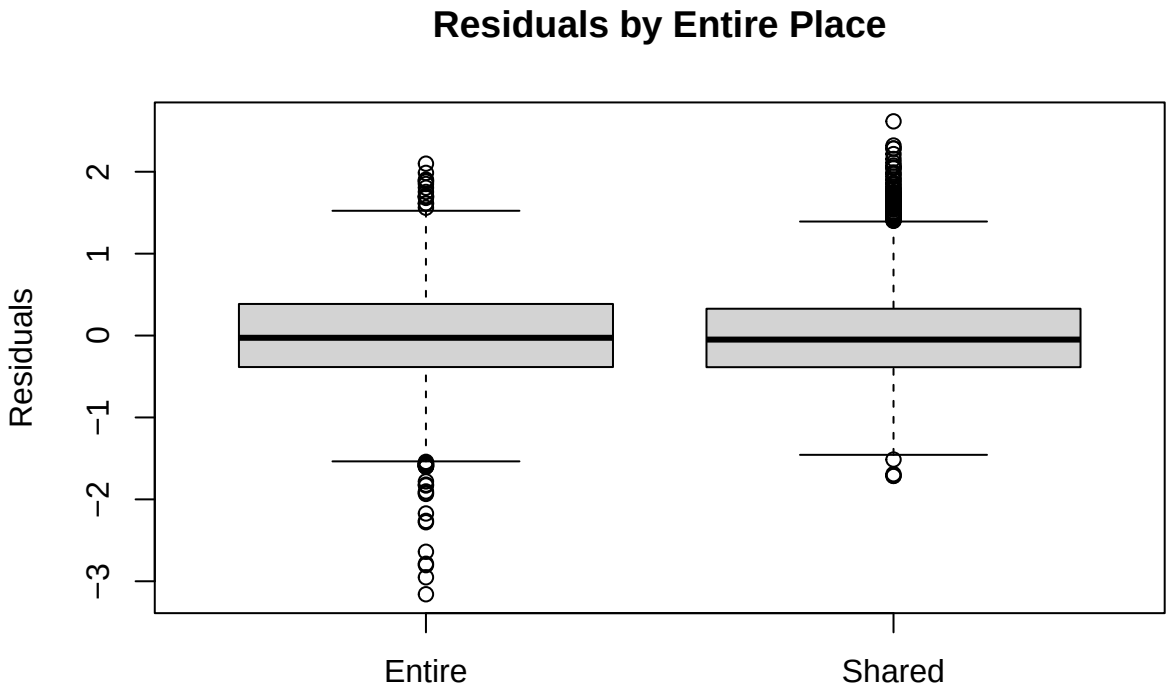
Rechecking Assumption

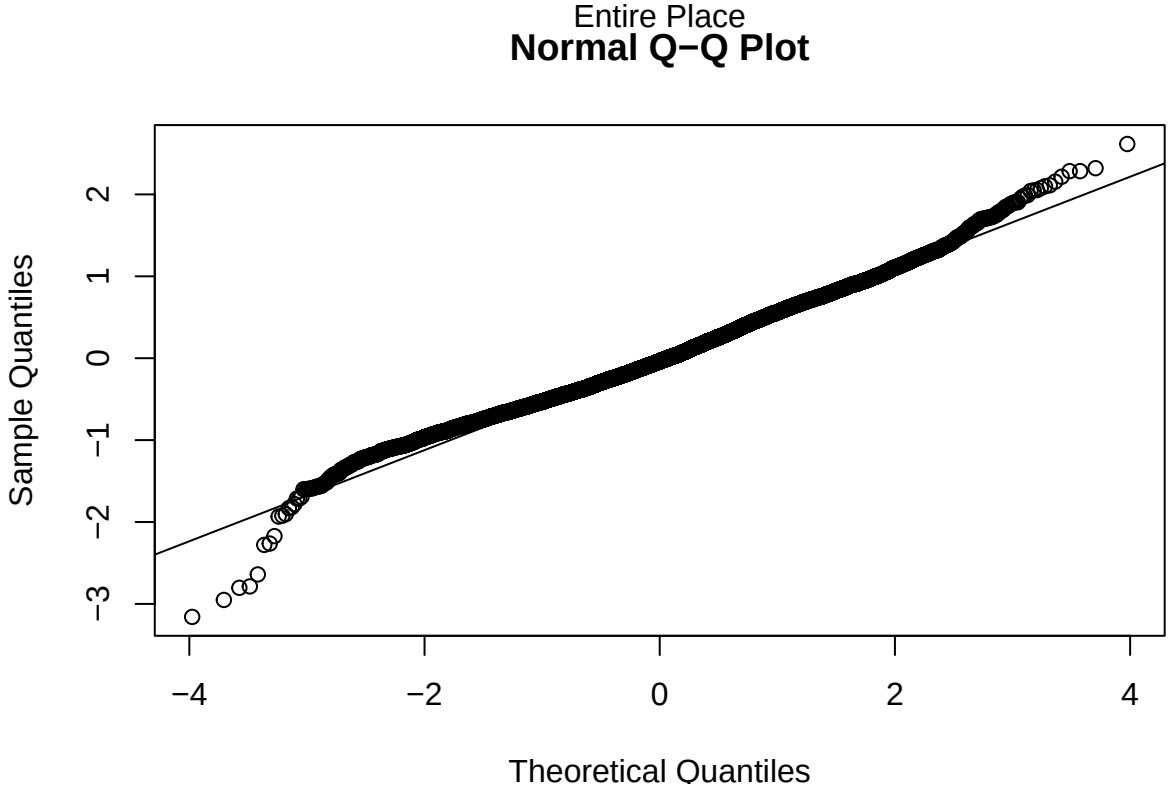





Residuals vs Fitted Value
Residuals
Fitted Value
3.5
4.0
4.5
5.0
5.5
6.0
6.5
−3
−2
−1
0
1
2


Residuals vs Bedrooms
Residuals
Bedrooms
2
4
6
8
−3
−2
−1
0
1
2





Firstly, looking at the MLR Condition 1: Conditional Mean Response, There is a random scatter around the diagonal and no simple function seems to be present, therefore we can safely conclude that the Condition 1 holds. For MLR Condition 2: Predictor vs Predictor, There are absolutely no signs of curves or non linear patterns implying that the predictors are at most linearly related. Therefore we also conclude that Condition 2 holds. Second, looking at the residual vs Fitted values, we see that there is a slight fanning pattern, which implies that there may be a violation of constant variance. Otherwise, the plot appears to have random

scatter, not violating uncorrelated errors or linearity. Similarly, when observing each numeric predictor versus residuals, a similar fanning pattern is a present, hence supporting the violation of constant variance, while the graphs otherwise have random scatter which does not violate uncorrelated errors or linearity. The Residuals vs Entire place plot appears to have some significant clustering of outliers, which could point towards to the uncorrelated errors assumptions. This violation is further supported by the Residuals vs Superhost plot, as there appears to be some clustering of outliers in this plot as well. Lastly, the QQNorm plot only has slight deviations on each end from the straight line, and hence there is no evident violation of normality. Only MLR Condition 1,2 and Residual vs Fitted, Residual vs Bedrooms, Boxplot for entire_place, and QQplot is shown above respectively. Rest in the appendix.

MODULE 5

```
## [1] "confidence interval for coeffecients"
```

```
##                            2.5 %     97.5 %
## (Intercept)          4.69157515  4.9203297
## I(1/Bedrooms^3)     -0.24074731 -0.1520915
## I(1/Bathrooms^2)    -0.28606702 -0.2237184
## Cleanliness_Rating   0.07933312  0.1180905
## Entire_PlaceShared  -0.65809669 -0.6162413
## Superhostf:Bedrooms  0.11908368  0.1639467
## Superhostt:Bedrooms  0.16597281  0.2132980
```

```
## [1] "confidence interval for mean response"
```

```
##        fit      lwr      upr
## 1 4.988479 4.971173 5.005785
```

```
## [1] "prediction interval for actual response"
```

```
##        fit      lwr      upr
## 1 4.988479 3.933827 6.043131
```

This is the 95% confidence interval of the estimated coefficients.

Additionally, we see that our 95% confidence interval for the logPrice of the airbnb (Entire Property) per night with 1 Bedroom. 1 Bathroom and 4.0/5.0 Cleanliness Rating is [4.971173,5.005785]

Finally, our prediction interval is [3.933827,6.043131] which represents the 95% mode likely logPrice of the airbnb (Entire Property) per night with 1 Bedroom. 1 Bathroom and 4.0/5.0 Cleanliness Rating.

MODULE 6

```
## [1] "The F-Statistic of the fixed model is:"
```

```
##    value
## 1817.172
```

```
## [1] "with our F* value being:"
```

```
## [1] 2.099231
```

Our F-statistic of the model is 1817.172 as shown in the summary above, since its significantly larger than the F* (1817 $>$ 2.09) we fail to reject the null hypothesis and conclude that significant linear relationship exists for at least one predictor.

```
## Analysis of Variance Table
##
## Model 1: logPrice ~ Bedrooms + Cleanliness_Rating + Superhost:Bedrooms
## Model 2: logPrice ~ I(1/Bedrooms^3) + I(1/Bathrooms^2) + Cleanliness_Rating^10 +
##     Entire_Place + Superhost:Bedrooms
```

```
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1  14248 5385.9
## 2  14245 4122.8  3    1263.1 1454.7 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Because there were no insignificant predictors in fact all predictors are very significant, it is not appropriate to do the partial F test. However, when conducting a partial F test with the simpler reduced model (logPrice = Bedrooom + Cleanliness_Rating + Superhost:Bedrooms) as an example, the R squared value decreases significantly (25.9859389 %) relative to the original model (43.33%) and it is no longer a reliable model. Even though the partial F test tells us that reduced model is better, it is most likely due to all predictors being significant in the first place and it automatically suggests simpler model. Therefore we conclude that our fixed_model is still the best model.

MODULE 7

```
## [1] "R^2 and the Adj. R^2 value of fixed_model"
```

```
## [1] 0.4335541
```

```
## [1] 0.4333155
```

As shown in the summary above, about 43% of the variation is explained by the model. Below, we check for many multicolinearity.

```
## GVIFs computed for predictors
```

```
##                         GVIF Df GVIF^(1/(2*Df))      Interacts With
## Bedrooms            1.811762  3        1.104121 Superhost, Bedrooms
## Bathrooms          11.194683  0             Inf                 --
## Cleanliness_Rating  1.030146  1        1.014961                 --
## Entire_Place        1.262298  1        1.123520                 --
## Superhost           1.811762  3        1.104121            Bedrooms
##                                                                 Other Predictors
## Bedrooms                                 Bathrooms, Cleanliness_Rating, Entire_Place
## Bathrooms          Bedrooms, Bathrooms, Cleanliness_Rating, Entire_Place, Superhost
## Cleanliness_Rating                     Bedrooms, Bathrooms, Entire_Place, Superhost
## Entire_Place                    Bedrooms, Bathrooms, Cleanliness_Rating, Superhost
## Superhost                                Bathrooms, Cleanliness_Rating, Entire_Place
```

```
## [1] "corelation Bedrooms and Bathrooms"
```

```
## [1] 0.6465326
```

```
## [1] "corelation Bedrooms and Cleanliness_Rating"
```

```
## [1] 0.03157365
```

```
## [1] "corelation Bathrooms and Cleanliness_Rating"
```

```
## [1] 0.01364935
```

Checking for multicolinearity, we see that there is a huge multicollinearity for the bathrooms predictor. To investigate this further, we computed the correlation between each predictors it show thats Bedrooms and Bathrooms have a very high corelation 0.6465326 relative to other predictors. This makes sense because higher priced properties generally tend to have more bedrooms and bathrooms and the amount of each room increases together. It's very rare for a house to have 5 bedrooms but just 1 bathroom.

```
##
## Call:
## lm(formula = logPrice ~ I(1/Bedrooms^3) + Cleanliness_Rating^10 +
```

```
##     Entire_Place + Superhost:Bedrooms, data = final_data1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.10474 -0.38755 -0.03844  0.36508  2.59197
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           4.552178   0.056664   80.34   <2e-16 ***
## I(1/Bedrooms^3)      -0.242776   0.022630  -10.73   <2e-16 ***
## Cleanliness_Rating    0.101137   0.009974   10.14   <2e-16 ***
## Entire_PlaceShared   -0.618172   0.010706  -57.74   <2e-16 ***
## Superhostf:Bedrooms   0.183111   0.011245   16.28   <2e-16 ***
## Superhostt:Bedrooms   0.231640   0.011889   19.48   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5428 on 14246 degrees of freedom
## Multiple R-squared:  0.4233, Adjusted R-squared:  0.4231
## F-statistic:  2092 on 5 and 14246 DF,  p-value: < 2.2e-16
```

Therefore we created another model called fixed_model2 which is the same model as fixed_model, just without the bathrooms predictors to address multicollinearity. The model is given by the following:

$$\text{logPrice } (Y) = 4.552178 - 0.242776(\text{Number of Bedrooms})^{-3}$$

$$+0.101137(\text{Cleanliness Rating})^{10} - 0.618172\ \mathbf{I}(\text{Entire Place = Shared}) + 0.231640(\text{Superhost='t' : Bedrooms}) + \hat{\epsilon}$$

Rechecking Assumption

From these plots (in the appendix), The results are very similar with the fixed_model. Therefore we must investigate deeper.

MODULE 8

```
## [1] "leverage points"
## [1] 856
## [1] "outliers (large dataset n>=50)"
## [1] 14
## [1] "Cooks distance for influence points"
## [1] 0
## [1] "DFFITS"
## [1] 625
## [1] "Beta 0"
## [1] 583
## [1] "Beta 1"
## [1] 835
## [1] "Beta 2"
## [1] 865
## [1] "Beta 3"
## [1] 474
## [1] "Beta 4"
## [1] 744
## [1] "Beta 5"
## [1] 547
```

```
## [1] "Beta 6"
## [1] 593
```

```
## [1] "leverage points"
## [1] 1065
## [1] "outliers (large dataset n>=50)"
## [1] 14
## [1] "Cooks distance for influence points"
## [1] 0
## [1] "DFFITS"
## [1] 583
## [1] "Beta 0"
## [1] 552
## [1] "Beta 1"
## [1] 904
## [1] "Beta 2"
## [1] 465
## [1] "Beta 3"
## [1] 721
## [1] "Beta 4"
## [1] 547
## [1] "Beta 5"
## [1] 662
```

Determining problematic observations, we see that both models have same number of outliers, no influence points using cooks distance, but tends to have more Leverage points, and similar amount of observations that influence the coefficients even though it has 1 less predictor in the fixed_model2. Therefore, we may lean more towards fixed_model.

MODULE 9

```
## [1] "Likelihood measures of fixed_model"
```

```
##           [,1]      [,2] [,3]
## [1,] -17663.62 -17610.67    7
## [2,] -17663.62 -17610.67   NA
```

```
## [1] "Likelihood measures of fixed_model2"
```

```
##           [,1]      [,2] [,3]
## [1,] -17410.93 -17365.54    6
## [2,] -17410.93 -17365.54   NA
```

We already saw in the summaries of each model that the original model (fixed_model) has higher R^2/Adj R^2 value. After computing AIC, BIC, AICc (first, second, third column respectively), we see that AIC and BIC both agrees that fixed_model is the better model while AIc states that fixed_model2 is the better model but just by 1. After these factors, we conclude fixed_model is better because even though it has an extra predictor, these likelihood measures return a better value for the fixed_model.

After using the automated selection algorithm to determine if there is another model with same pool of predictors that are better than our current best model fixed_model. Before applying this algorithm, we excluded Bedrooms, Bathrooms, and Cleanliness rating from the scope_list and instead included t_Bedrooms, t_Bathrooms, t_Cleanliness_rating which are the transformed values of each respective predictors. All of foward, backward, stepwise direction agrees on 1 model, which is including every predictor in the dataset except the categorical variable superhost. All directions forward, backward, and stepwise automated selection agrees on 1 model which is given by:

$$logPrice(Y) = 4.739 - 0.5055I(x_{EntirePlace} = Shared) - 0.03173x_{Bed} + 0.008942x_{Amenities} + 0.1320x_{Accommodates}$$

$$-0.03950x_{HostCommunicationRating} - 0.1179x^{-3}_{Bedrooms} - 0.1572x^{-2}_{Bathrooms} + 1.249e^{08}x^{-10}_{CleanlinessRating}$$

Discussion

As a result of the analysis using several statistical methods, we conclude that model obtained from the automated selection tool provides the best explanation for our research question. It contains the number of bedrooms, the number of bathrooms, the rating of cleanliness, the number of beds, the host communication rating, the number of accommodations, the number of amenities and whether or not the listing was shared as variables. In other words, these variables are significantly associated with the overall price of Airbnb listings. Based on the coefficients corresponding to these variables, an increase in the number of amenities, accommodates and cleaning rating causes the increase in the price of airbnb, while other variables decrease the price of airbnb when their values increase. Our best model is the descriptive rather than predictive Since it focuses on identifying variables related to the response. Thus, it is suitable to be used to answer our research question. We expected the Superhost' rating to be an important factor in the price, however, through the analysis, we got the unexpected result that the model does not include it.

There are some limitations to consider. The automated selection tool we used to get the best model is likely to exclude the significant predictor. In our case, 'Superhost' could be removed in our model even if it is actually significant in our analysis. Moreover, from the result section, we noticed severe multicollinearity between the number of bedroom and bathroom but our final model contains them. This is likely to cause some potential problems. For example, coefficients in the model could have the wrong sign compared to the literature, then it could lead our best model to explain the relationship between predictors and response incorrectly. Since it is often hard to prevent this problem, we decided to leave this issue as a limitation. However, despite of these limitations, our final model is still reasonable.

Ethics

When discussing the model selection, we encountered two noteworthy questions of negligence. Firstly, we had found that our initial model violated every assumption, and set out to correct it with a Box-Cox transformation, which would mean omitting 38 pieces of data that had properties of zero bathrooms or were rated a zero for their cleanliness rating. Having to omit this data presented a slight discussion, as it means that our model may not report as accurately when

evaluating situations that hold these values, and could hence result in misinformation being spread and result in negligence on our part. However, because this accounted for approximately 0.3% of the observations, we considered it to be negligible data. After performing our transformation, we analyzed our new model and found that there was a strong multi-collinear relationship between the number of bedrooms and bathrooms. While we found that this is likely typical, it did prove a point of discussion as this is a violation of multiple linear regressions. Our concern was that by leaving it in, it may result in a disproportionate positive or negative relationship between price and more severe cases of the numbers of bathrooms and bedrooms. This is yet another cause of negligence, as we acknowledge it could be problematic for those attempting to use this model for pricing and these numbers could result in disproportionate sways. However, we decided to keep them in our model, as AIC and BIC testing both confirmed a model with these predictors was best.

Bibliography

Chattopadhyay, Manojit, and Subrata Kumar Mitra. "Do airbnb host listing attributes influence room pricing homogenously?" International Journal of Hospitality Management, vol. 81, 2019, pp. 54–64, https://doi.org/10.1016/j.ijhm.2019.03.008.

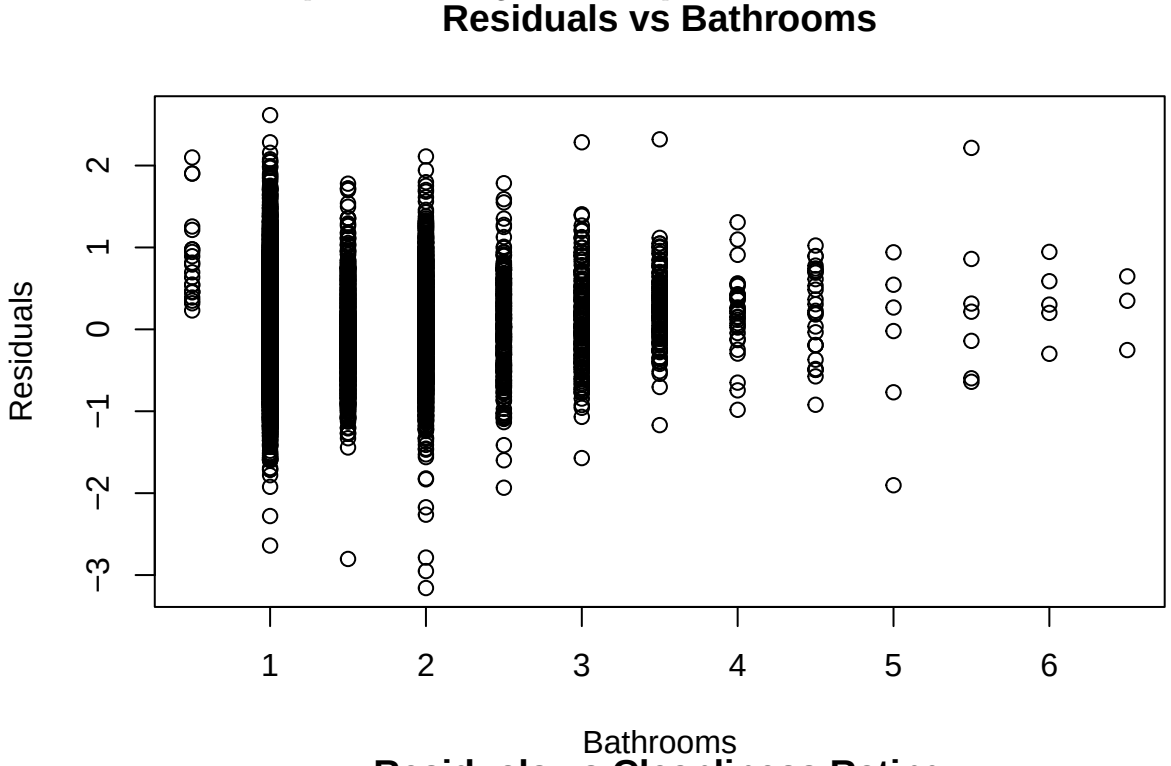
Kwok, Linchi, and Karen L. Xie. "Pricing strategies on airbnb: Are multi-unit hosts Revenue Pros?" International Journal of Hospitality Management, vol. 82, 2019, pp. 252–259, https://doi.org/10.1016/j.ijhm.2018.09.013.

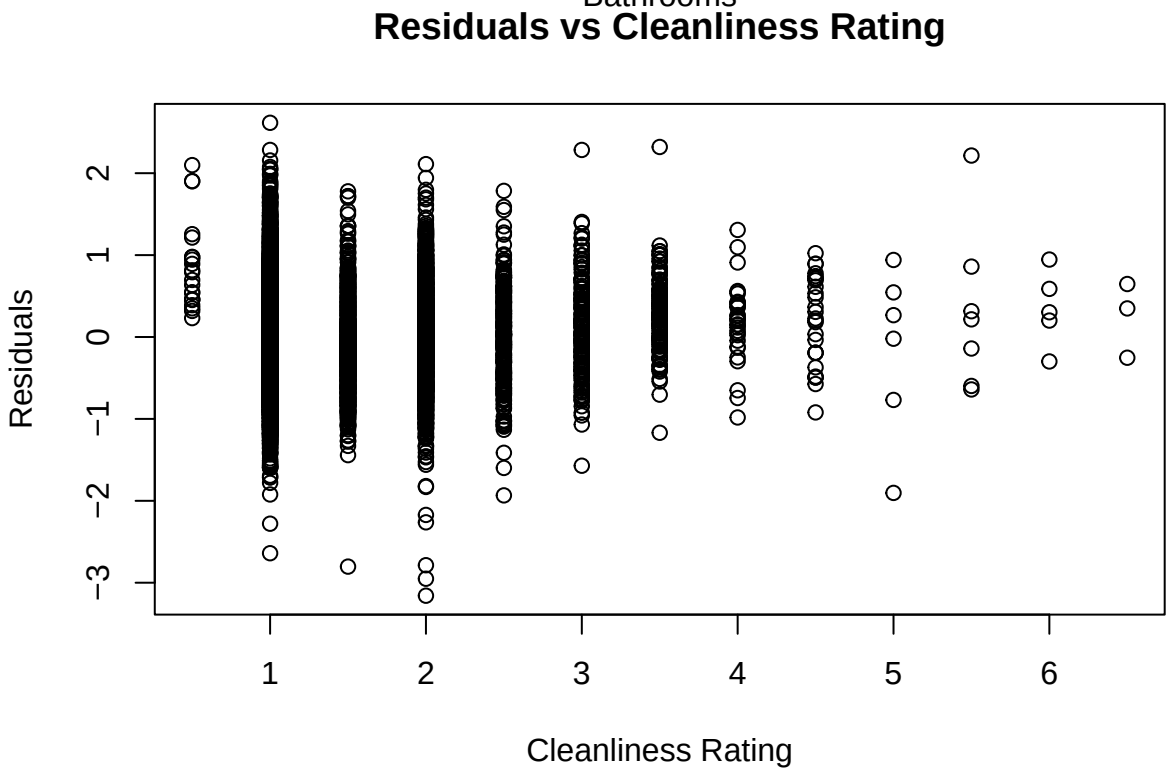
Voltes-Dorta, Augusto, and Agustín Sánchez-Medina. "Drivers of airbnb prices according toproperty/room type, season and location: A regression approach." Journal of Hospitality and Tourism Management, vol. 45,

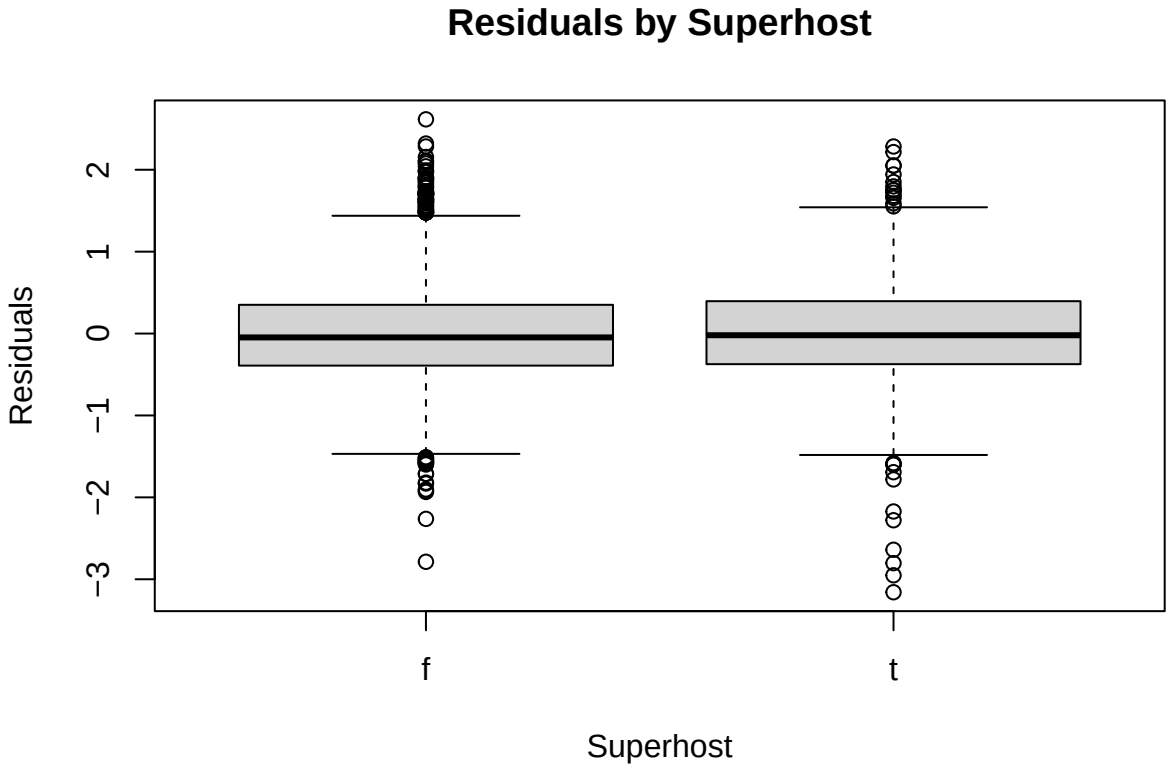
2020, pp. 266–275, https://doi.org/10.1016/j.jhtm.2020.08.015.

Apendix

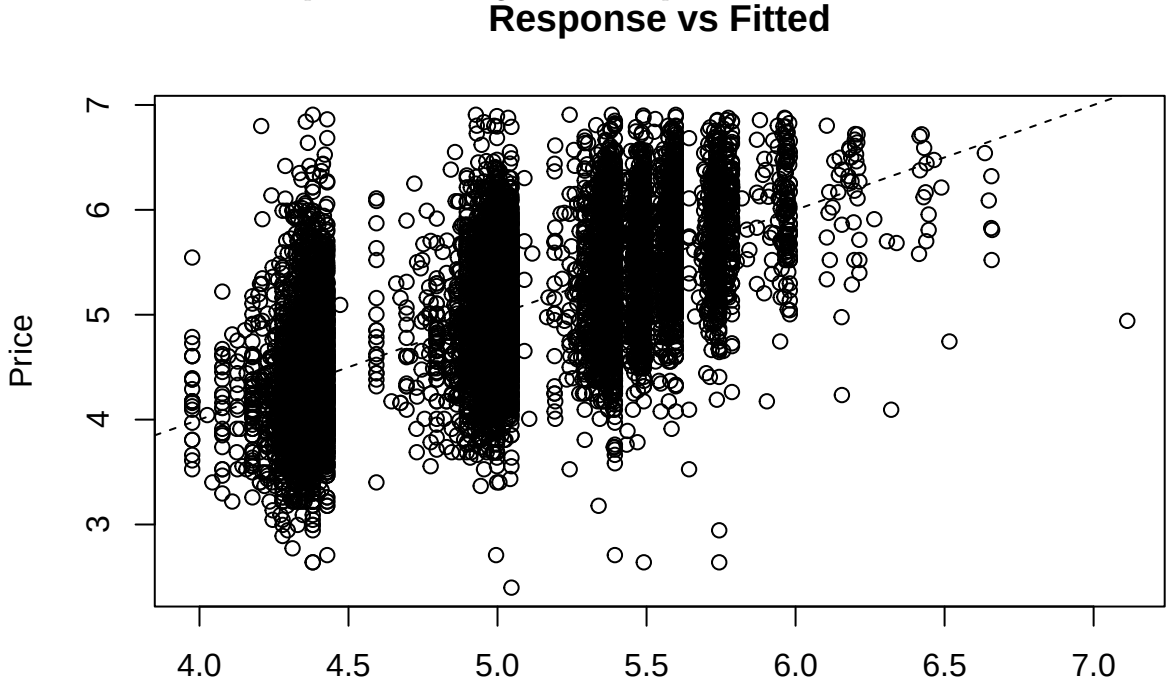
Below are the rest of the plots for checking linear assumptions of fixed_model:

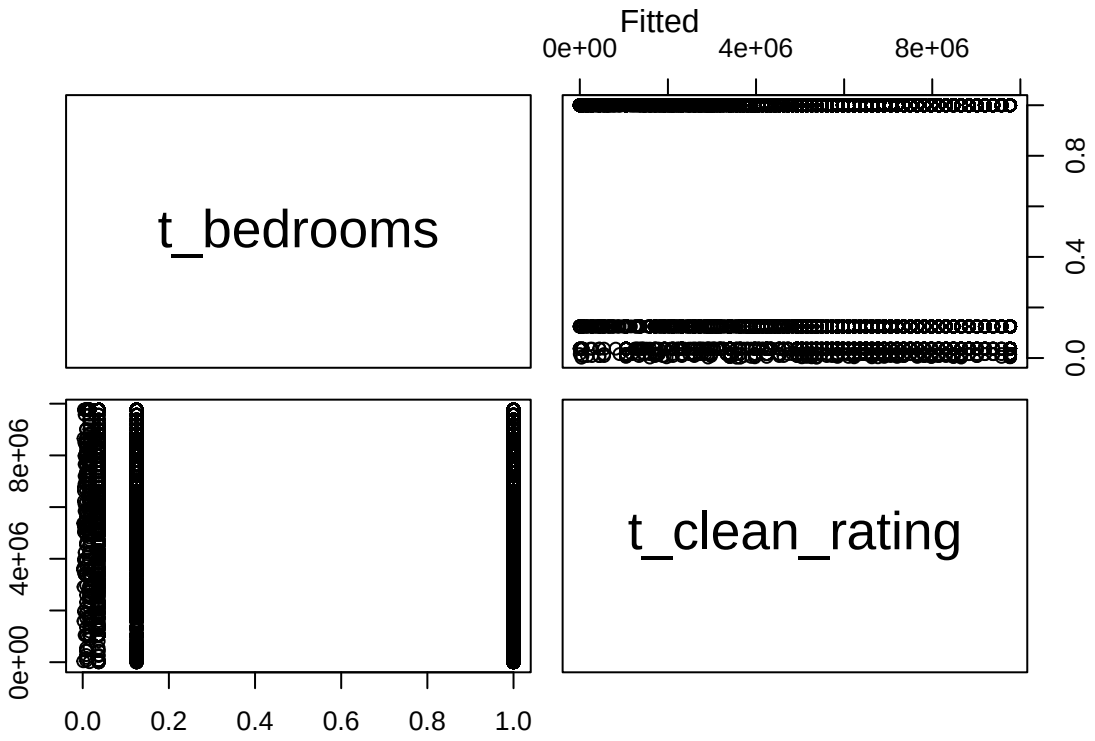


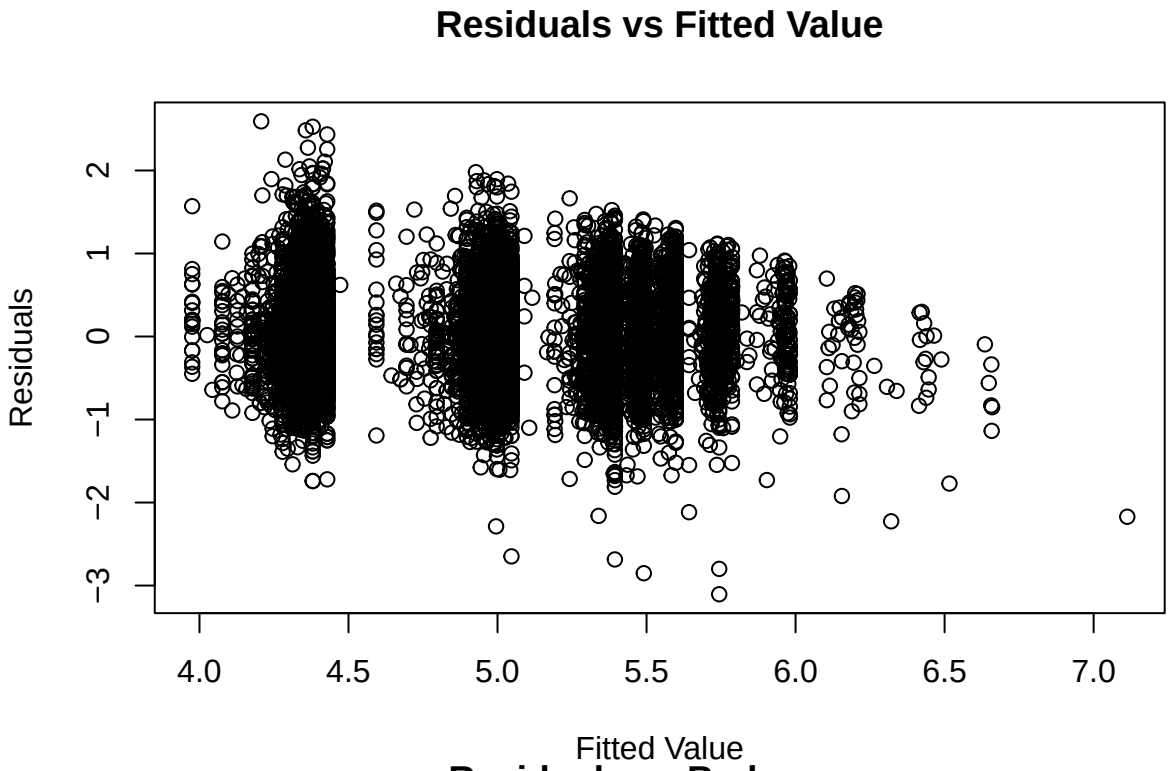


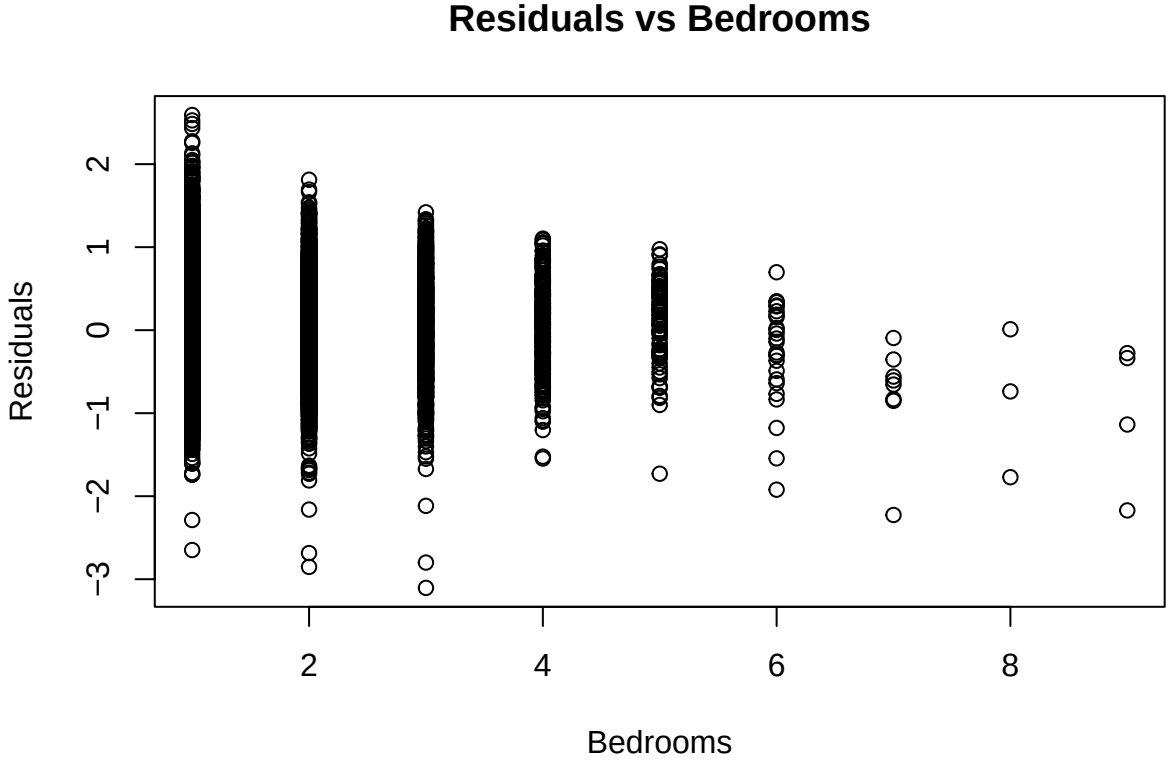

Residuals by Superhost
Residuals
2
1
0
−1
−2
−3
f
t
Superhost

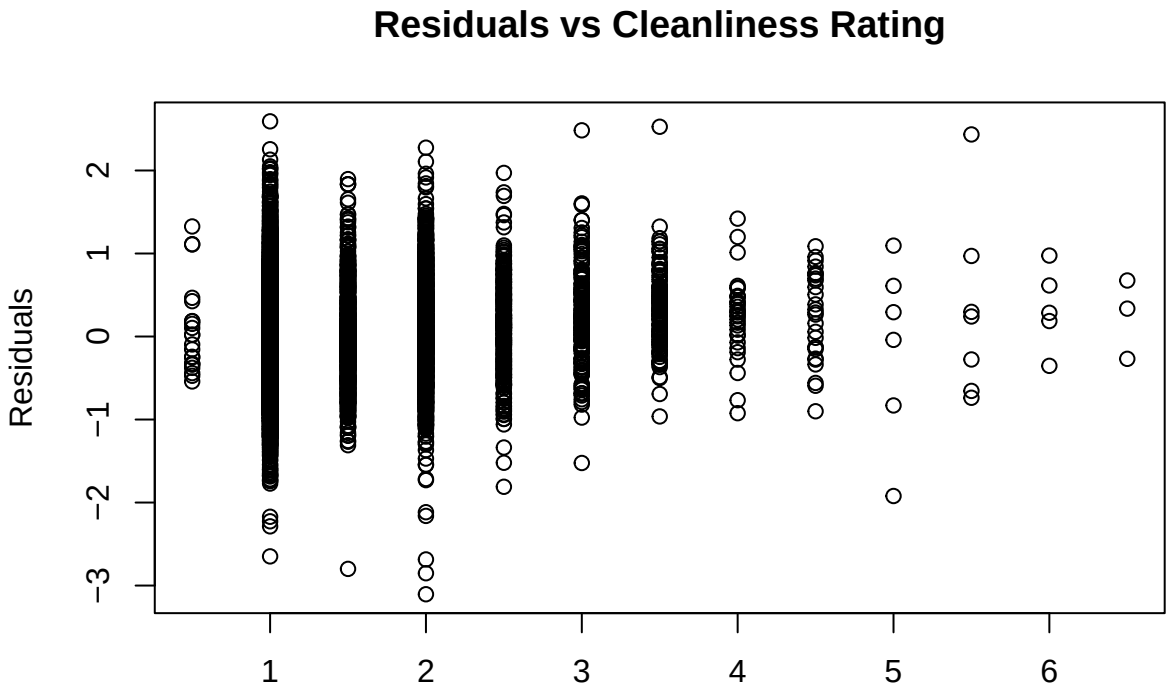
Below are the rest of the plots for checking linear assumptions of fixed_model2:

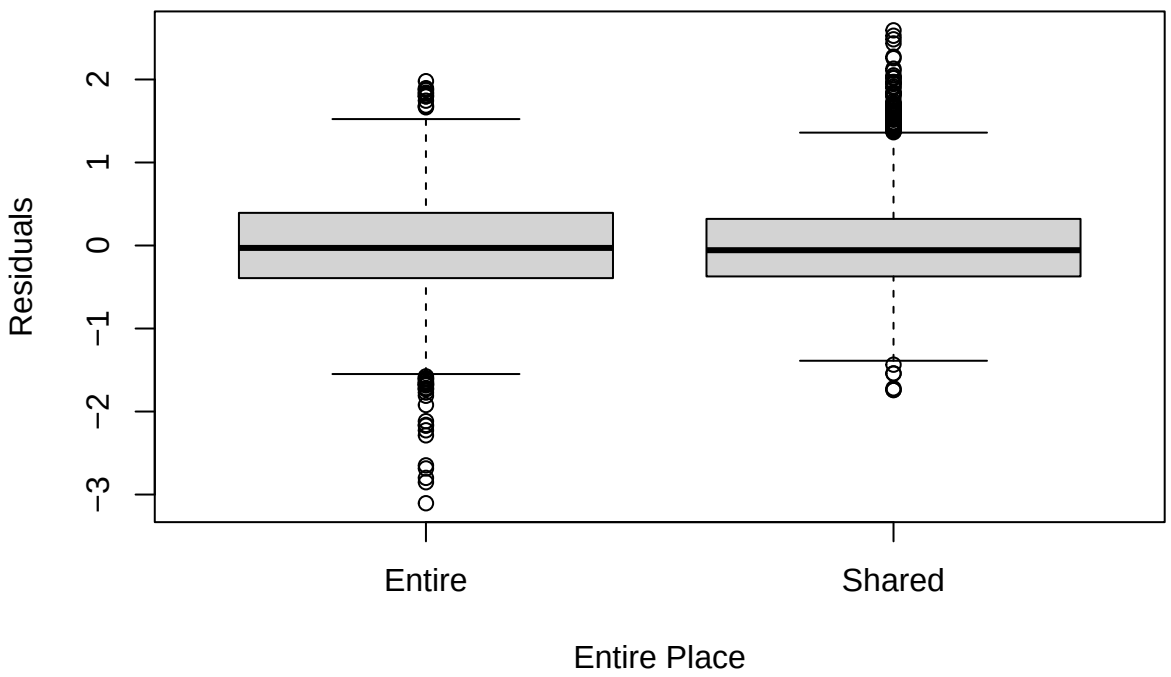


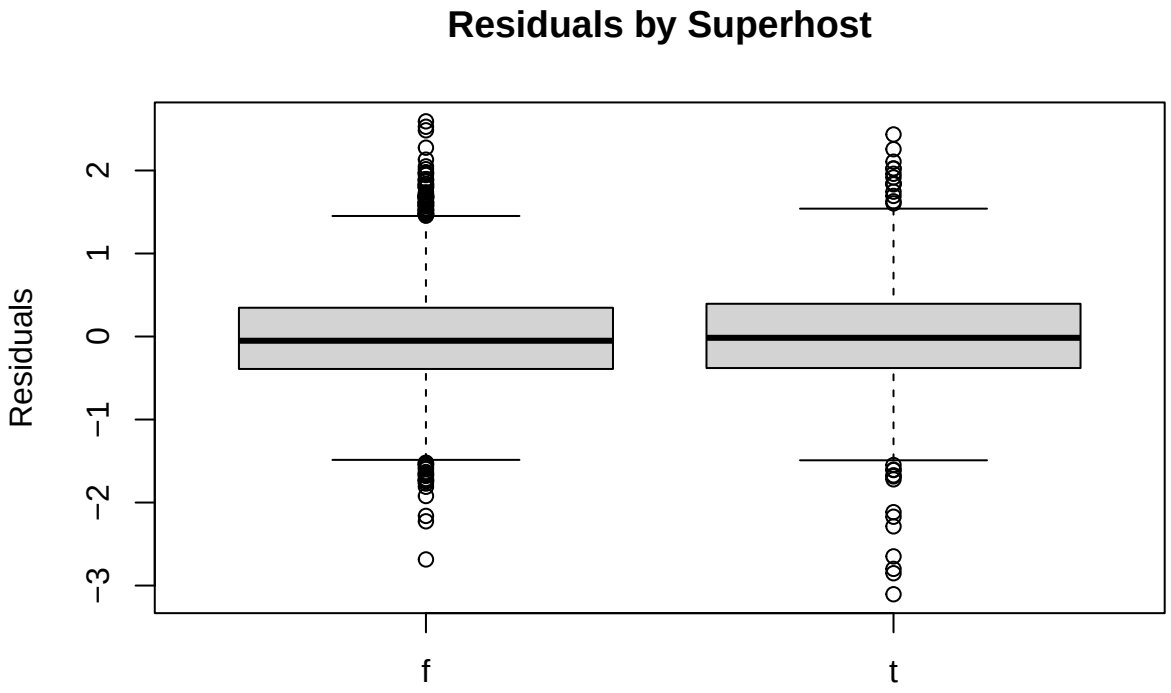


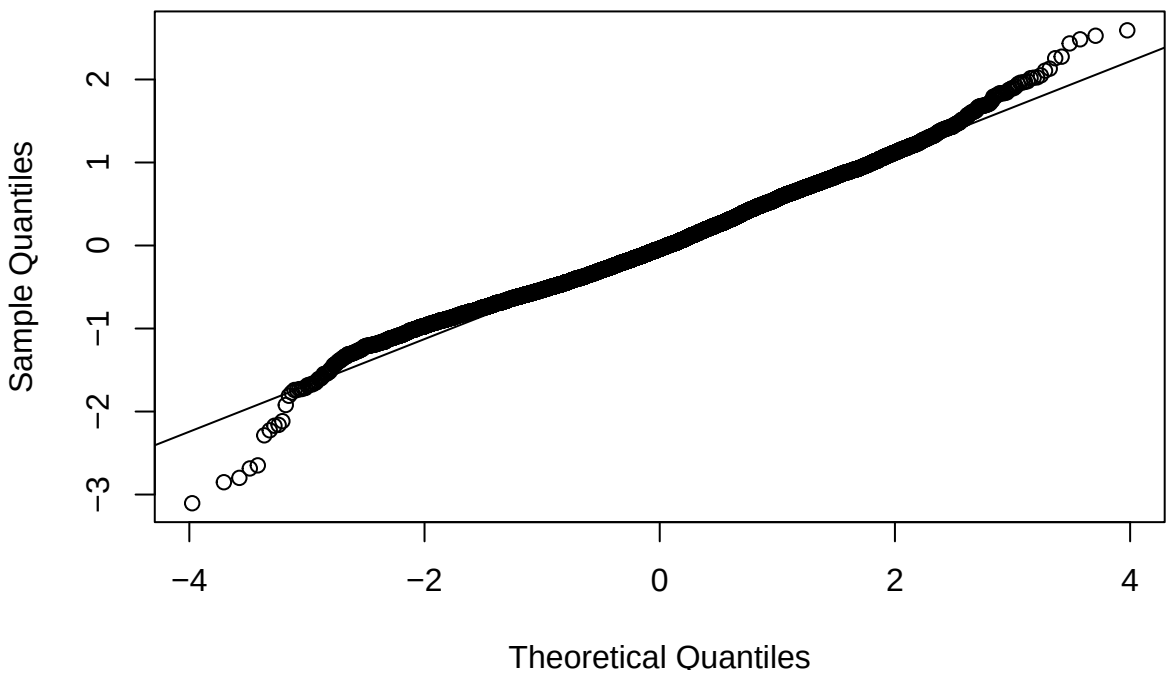

Residuals vs Fitted Value
Residuals
2
1
0
−1
−2
−3
4.0
4.5
5.0
5.5
6.0
6.5
7.0
Fitted Value


Residuals vs Bedrooms
Residuals
2
1
0
−1
−2
−3
2
4
6
8
Bedrooms


Residuals vs Cleanliness Rating
Residuals
−3
−2
−1
0
1
2
1
2
3
4
5
6
Cleanliness Rating


Residuals by Entire Place
Residuals
−3
−2
−1
0
1
2
Entire
Shared
Entire Place


Residuals by Superhost
Residuals
2
1
0
−1
−2
−3
f
t
Superhost


Normal Q−Q Plot
Sample Quantiles
2
1
0
−1
−2
−3
−4
−2
0
2
4
Theoretical Quantiles