The Impact of Location and number of Reviews on Airbnb Rental Prices: Applying the Ordinary Least Squares (OLS) as an analytical tool.

Armando Porras (6684211)

Duong Nguyen (6827209)

Gabriela Putri (6812732)

Hanna Kreitem (6816940)

Hassaan Rashid (6457207)

Lan Nguyen (6687547)

Team 5.2

Abstract

The phenomenal growth of Airbnb, an American vacation rental online marketplace firm, led to copious investigation in the company's keys to success. The adoption of dynamic pricing strategies is conceived as one of the most prevalent reasons. In this paper, we examine particular elements that influence Airbnb's price-setting procedure. In Section I, we explain our purpose of research and elaborate on the hypotheses that rental price has a negative relationship with the property's location but a positive relationship with the number of reviews on the property. Section II gives a detailed description of the data and the process of formulating the research equation. Section III discusses the relationship between the dependent variable and explanatory variables, based on which we perform several hypothesis testing to examine our postulations. The result is robust to conclude that the price of an Airbnb does not merely depend on location but also on several factors, including how many reviews it receives and the type of the room. However, there are several limitations of the paper that we want to explore in future research mentioned in the last section.

1 Introduction & Theory

1.1 Introduction

With the continued growth of internet applications and smartphones, as well as globalisation and urbanisation, there has been an advent of the *sharing economy* in recent years. In a broad perspective, the sharing economy can be defined as a business model designed to provide convenient and cost-effective access to take advantage of underutilised or redundant resources provided by digital platforms (Grinevich, Huber, Karatas-Özkan, Yavuz, 2019). The main features of the sharing economy include its online platform, idle capacity, collaborative consumption, non-ownership, accessibility of share, trust and network-based activity and for-profit activities (Ranjbari, Morales-Alonso, Carrasco-Gallego, 2018). People can share access to resources, such as transportation, accommodation, food, and skills. In the accommodation sector, one major player is Airbnb. The platform hosts more than 10 million listings and is present in more than 81,000 cities in 191 countries worldwide (Much Needed, 2020).

The all-too-familiar global COVID-19 pandemic is proving to be Airbnb's biggest challenge yet. In April 2017, Airbnb was valued at $31 billion, and three years later, it was worth a meagre $18 billion (Bosa, 2020). Yet, can Airbnb actually play an important role in fostering a sustainable and lasting recovery from this relentless pandemic?

New research from Oxford Economics highlights how Airbnb can play a key role in supporting the early recovery in tourism and jobs growth (Oxford Economics, 2019). The report studied Airbnb's impact in Australia and found that the company contributes approximately $10.38 billion to national GDP and supports almost 90,000 jobs in just 2019 alone. Furthermore, they also conveyed how Airbnb spread the benefits of tourism locally, by studying the impacts in Thailand in the same year. Airbnb guests spent over THB19.6 billion in local restaurants and shops, and for every THB1,000 spent on the online housing platform, an additional THB420 was spent on local businesses. Moreover, by offering unique listings and experiences, it can prompt tourism in less popular travel destinations with more than 9% of guest spending taking place outside key cities.

Therefore, the undisputed relevant contribution of Airbnb to present matters has led to the need to specify and estimate models targeted at determining factors underlying sharing economy pricing strategies. With this research, hosts will be able to more appropriately price their properties, thus a higher likelihood of Airbnb contributing to the pandemic recovery.

1.2 Theory

In recent years, the increasing popularity and exponential growth of Airbnb in the accommodation sector have attracted scholarly attention, and various studies have been conducted to explore the reasons behind this phenomenon. Several studies concentrated on Airbnb's impacts on the hotel industry, including hotel revenue and tourism industry development, while other researchers were more concerned about the relationship among service quality, satisfaction, and customers' loyalty (Zhang, 2017).

Among a wide range of factors contributing to the success of Airbnb, use of dynamic pricing strategies by Airbnb hosts is proven to be of the most influential (Gibbs & Guttentag, 2017). Realizing the indispensable role of the price-setting, we aim to look closer at specific elements that influence this process. Based on what we have searched, location and number of reviews are most relevant to the price-setting technique of Airbnb. Therefore, our research is to analyze the impact of these specific elements on Airbnb rental price by applying the Ordinary Least Squares (OLS) as an analytical tool.

According to Cui (2018), location has a significant influence on the offer price of each available-for-rent property. He argues that the company evaluates how accessibility, in terms of transport costs and disutility generated by other conveniences, affects customer's value on the property. However, since this topic is beyond our research we will not go in depth into it. The paper by Dan Hill (2015) gives strong evidence to the relationship between the number of reviews and the offer prices that we are interested in. He elucidates that the automated source of pricing information developed by Airbnb itself worked with all the quantifiable attributes they got about a listing and then looked to see which were the most highly correlated with the price a guest would pay for that listing. Airbnb came to a surprising result that people are willing to pay a premium for places with many reviews. For Airbnb, having a single review rather than no reviews makes a huge difference to a listing.

Based on these results, we are confident to formulate two main hypotheses above which our paper is concentrated. First, we predict that location has a negative relationship with the price for rent: the further the property is from the city center, the less does it cost. Second, we expect a positive relationship between the number of reviews and the price for rent: the more reviews on the accommodation, the higher the price is worth. However, given the data set we cannot distinguish or categorize the quality of reviews (either positive or negative). Therefore, we have to make a strong assumption that all the reviews are non-negative.

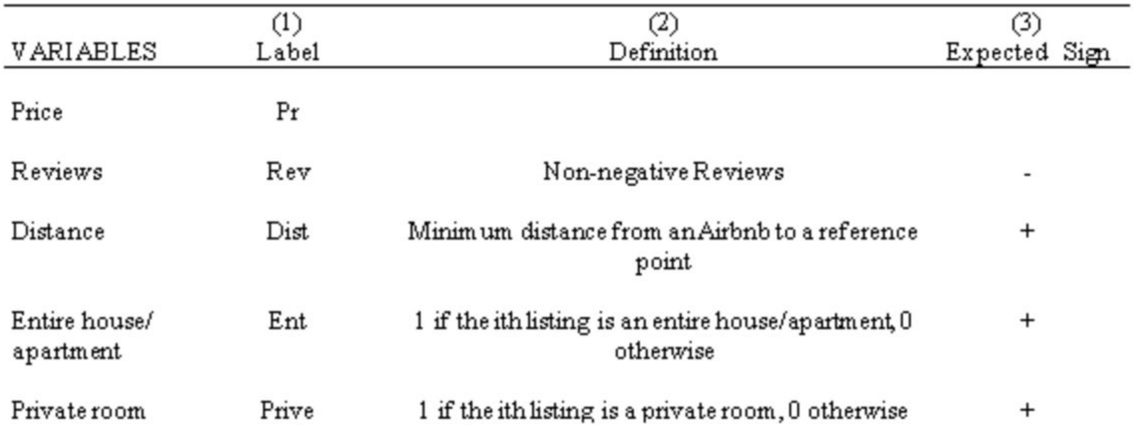
This assumption is supported by previous literature. It is found that the average Airbnb rating is 4.7 out of 5 stars (Zervas, Proserpio, Byers, 2016), which is overwhelmingly positive as the average rating of hotel reviews on TripAdvisor is only 3.9 out of 5 stars (Han et al, 2016). In addition, linguistic and large-scale sentiment analysis conducted on Airbnb reviews discovered that 98.1% of reviews and 76.4% of the sentences to be positive (Alsudais and Teubner, 2019). The staggering positivity can be attributed to psychological effects as this type of collaborative consumption is of a person-person nature (Bridges and Vásquez, 2016). People tend to be more careful in their complaints as they associate Airbnb reviews to reviewing another human (Zervas et al, 2015). These effects are accentuated if they developed a feeling of mutual trust and familiarity from the stay (Dayter and Rudiger, 2013). The positivity can also be due to Airbnb's review system which differs from hotels', as to maintain a Superhost status, the host has to achieve four 5-star reviews for every 4-star review (Campbell, 2018). Therefore, we believe the assumption that all reviews are non-negative is plausible.

While a wide range of scientific work was in support of the first hypothesis, many opposed the second. The papers by Ert, Fleischer, & Magen (2015) and Dogru & Pekin (2016) applied hedonic pricing analysis and attributed a lion share of Airbnb's dramatic growth to location whereas indicated that online reviews and ratings did not appear to have an effect on the listing price. In this paper, we want to investigate both hypotheses by applying the Ordinary Least Square estimator (OLS). We start with a bivariate model where price is an independent variable on the minimum distance - the minimum average distance from the properties to five different reference points. We then formulate a multivariate equation in which price is regressed on location, number of reviews, and room types. We will keep types of room as dummy variables and see how changes in either location or number of reviews would affect price behaviors, holding constant other variables. We also run the t-test on each independent variable and then the F-test to see whether these two variables should be included in the equation. In conclusion, we acknowledge that our model is consistent with current literature and can be deployed to estimate the main determinants underlying Airbnb pricing strategies and thus enable both hosts and customers to properly value the property.

2 Data description

2.1 Study area

The study area, New York, has been chosen with purpose, mainly due to the heterogeneity of each neighborhood or borough. New York is made up of five boroughs, each with its own distinct charm and character: Manhattan, Brooklyn, Queens, Bronx, and Staten Island. Manhattan is undoubtedly the most well-known, and is home to iconic skyscrapers, household-name attractions, as well as the headquarters of many major multinational corporations. Brooklyn is known for its cultural, social, and ethnic diversity, independent art scene, picturesque neighborhoods and notable architectural heritage. Queens is the most ethnically diverse urban area in the world, and is made up of a collection of historic small towns and villages founded by the Dutch (Weber, 2013). Other than being credited as the birthplace of hip-hop culture, Bronx has long been considered to be the "poorest borough", but since the 1990s, city policy has made it much more attractive (Chaffin, 2020). Finally, Staten Island is the most suburban in character, and is often known for being the greenest borough (Staten Island Advance, 2019). Therefore, this paper aims to discover and quantify the impact of location on Airbnb rental pricing, and how the relationships between the other independent variables vary with each borough.



	(1)	(2)	(3)
VARIABLES	Label	Definition	Expected Sign
Price	Pr		
Reviews	Rev	Non-negative Reviews	-
Distance	Dist	Minimum distance from an Airbnb to a reference point	+
Entire house/ apartment	Ent	1 if the ith listing is an entire house/apartment, 0 otherwise	+
Private room	Prive	1 if the ith listing is a private room, 0 otherwise	+

Table. 1. Variable names

For our research we defined our dependent variable as Pr in U.S. Dollars, with values ranging from a minimum of 10 and a maximum of 10,000. The independent variables that are hypothesized to affect "price" are location, number of reviews, and room type.

2.2 Variable description

Location is referred to as $Dist$, and it is defined as the minimum distance of an Airbnb rental to a reference point in kilometers. As there are five different neighbourhoods considered in the sample (Bronx, Staten Island, Queens, Brooklyn and Manhattan), we chose a main tourist attraction of each area as a benchmark and calculated the minimum distance of each rental listing from said benchmarks. After that, we selected the minimum distances and used them as a reference for location in our model. This should allow for a more logical interpretation of data rather than selecting only one benchmark position and redirecting all rental listings to it. The chosen benchmarks are: Central Park (Manhattan), Brooklyn Downtown (Brooklyn), Corona Park (Queens), Bronx park (Bronx), and Island Ferry (Staten Island). We then obtained a new variable with values ranging from .025km to 25.44km and a mean of 3.98km. The standard deviation indicates that most observations are within 2.05km of the average recorded.

Number of reviews is labelled Rev and measures reviews as a certain amount, regardless of length, quality or rating. However, as stated in the introduction, we assume that all reviews are non-negative. Total reviews may range from 0 to 629, averaging about 23 reviews per rental and with a standard deviation of 44.55.

The room types are represented as dummy variables in the model. There are three room types: entire houses or apartments, where there is full private availability with the chance of owners occupying a certain floor; private rooms, where guests have their own private room for sleeping, with the possibility of sharing other areas; and shared rooms, where bedrooms and common areas have to be shared (What do the different home types mean? - Airbnb Help

Center). We then obtained the variables Ent, $Prive\ and\ Shared$. When the room type is a private room, Ent equals 0 and $Prive$ equals 1, and if the room type is an entire home or apartment, viceversa. When both variables equal 0, then we are working with shared rooms. The means of the dummy variables represent the proportion of a certain type of accommodation in our data. Therefore, 51.97% of recorded observations correspond to entire homes and apartments, in contrast to the 45.66% corresponding to private rooms, and 2.37% to shared rooms.

2.3 Sample description

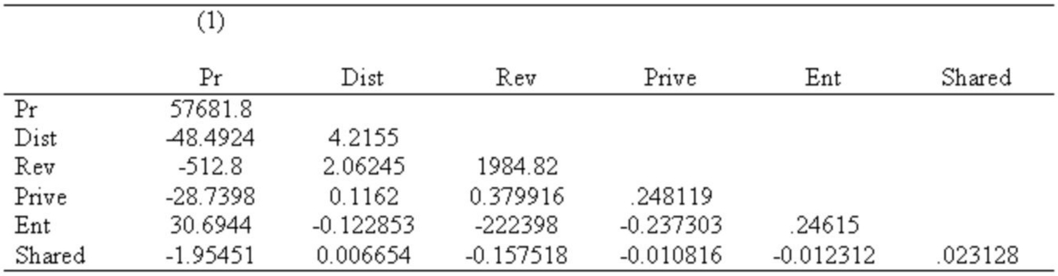
The complete sample that we are using contains 48,895 observations. However, since Airbnb listings cannot be free, 11 of them were considered missing values and omitted as they contained prices equal to 0. Such a reduced amount of omitted values should not cause any significant variation in the final results. Additionally, we noticed a negligible amount of unreasonably high prices for Airbnb listings in the dataset which rise up to 10,000$ per night. These could be considered as outliers, however there isn't sufficient information that shows whether these prices are justifiable or not. Therefore, we decided to keep these listings and not omit them to avoid any potential bias.

This sample is representative of Airbnb listings worldwide. Our population for the model is world-wide Airbnb listings. We decided on using New York City for our sample as it has particularly intriguing statistics with regards to its Airbnb presence. New York is among the most popular cities on Airbnb globally, and is ranked the second most popular in the United States in 2019 (iPropertyManagement, 2019). Although New York is Airbnb's largest market, as many as two-thirds of its listings are illegal. Due to fear of the effect of travelers on residential neighborhoods, some regions have passed laws and regulations about short-term rentals. New York is one such region, passing policies in which Airbnb users can only list one home at a time and are not allowed to rent out an entire apartment for less than 30 days (Tun, 2020). In addition, 72% of Airbnb hosts in New York rely on the revenue they earn from sharing their space to stay in their homes (Heath, 2015). Interestingly, there seems to be a "monopoly" in the New York Airbnb market, with the top 10% of hosts earning 48% of revenue in 2017 (iPropertyManagement, 2017). In combination, these thought-provoking statistics further fueled our curiosity in examining the determinants in Airbnb pricing strategies in one of the most prominent Airbnb markets globally. Some criticism for the use of NYC for our sample may be its failure to fully represent every city that contains Airbnb listings. New York City may be comparable to cities such as Paris or Madrid, but it may not reproduce the same conditions as other less attractive destinations. This may lead to a certain bias of results which might make them not applicable to certain locations. Despite this, it should not constitute a big issue for the interpretation and further real-world implementation of results.

Within our sample, we used data from five different neighborhoods. Observations from Manhattan and Brooklyn (total of 41,755) constitute 85.4% of total observations (48,884), whereas data for Queens, Bronx and Staten Island represents only 14.6% of total observations (with the latter containing only 373 observations). Manhattan and Brooklyn tend to be higher-status areas, hence leading to higher prices, which consequently causes a rise in average prices for our model.

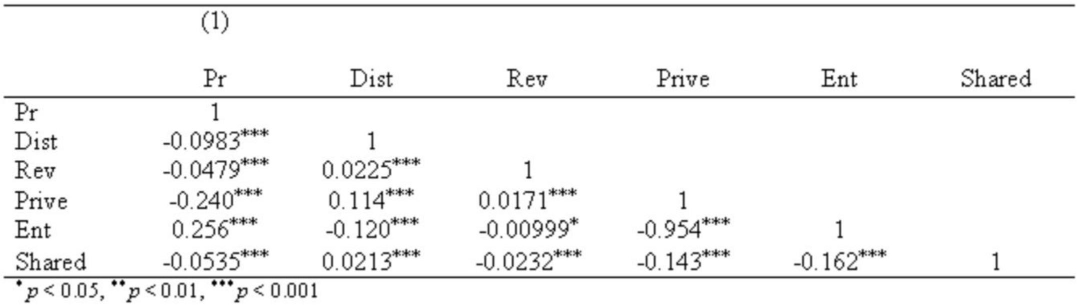
2.4 Bivariate analysis

We construct a covariance and correlation table to observe the relationship between price and the independent variables. From the covariance table, we can clearly see that the variance of price is 57861.8 while its covariance with number of reviews, private room, entire home or apartment and shared room is -48.4924, -512.8, -28.7398, 30.6944 and -1.95451 respectively. These results suggest a positive relationship between price and entire home or apartment while a negative relationship of price with number of reviews, private room and shared room. Correlation table is constructed to check the strength of relationship between these variables. From the correlation table, we can clearly infer the correlation between Pr and the dummy variables (Ent and Prive) is stronger than the correlation between Pr and the independent variables (Dist and Rev).



	(1)					
	Pr	Dist	Rev	Prive	Ent	Shared
Pr	57681.8					
Dist	-48.4924	4.2155				
Rev	-512.8	2.06245	1984.82			
Prive	-28.7398	0.1162	0.379916	.248119		
Ent	30.6944	-0.122853	-222398	-0.237303	.24615	
Shared	-1.95451	0.006654	-0.157518	-0.010816	-0.012312	.023128

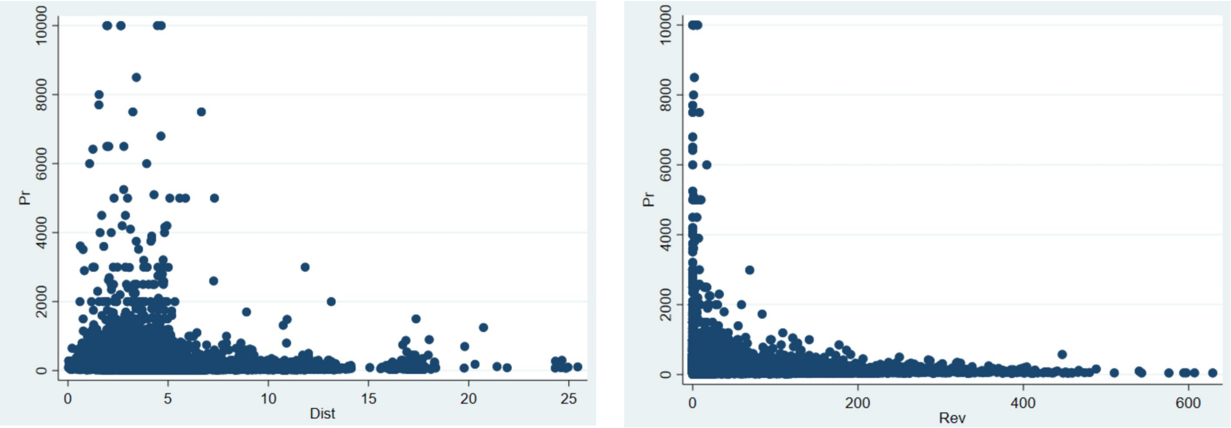
Table. 2. Covariance table for all variables.



	(1)					
	Pr	Dist	Rev	Prive	Ent	Shared
Pr	1					
Dist	-0.0983***	1				
Rev	-0.0479***	0.0225***	1			
Prive	-0.240***	0.114***	0.0171***	1		
Ent	0.256***	-0.120***	-0.00999*	-0.954***	1	
Shared	-0.0535***	0.0213***	-0.0232***	-0.143***	-0.162***	1

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Table. 3. Correlation table for all variables.



a. Price and Distance

b. Price and Number of reviews

Figure. 1. Scatter plots

Figure 1a and 1b show scatter plots between price and distance, and price and number of reviews. These diagrams also show fitted linear lines that are indicative of sample bivariate regression models.

3 Empirical Analysis

3.1 Regression model

As discussed in the theory section, location has a significant influence on the rental price of each available-for-rent property. We start with a bivariate model to study this relationship between an Airbnb accommodation's location and its price per night. Hence, the initial bivariate linear model used is:

$$Pr_i = \beta_0 + \beta_1 Dist_i + \varepsilon_i$$

The rental of a property does not depend solely on its location. There are many other factors that can have an impact on the rental price of an available-for-rent property. As discussed in the theory section, one such factor is the number of reviews for an Airbnb property. We extend the bivariate model to a multivariate model to include the effect of the number of reviews on the rental price. The multivariate linear model used is:

$$Pr_i = \beta_0 + \beta_1 Dist_i + \beta_2 Rev_i + \varepsilon_i$$

In our sample, we can identify three unique types of accommodations, namely shared room, private room, and entire apartment or house. Considering that previous studies have found that guests are willing to pay a premium for privacy and that the accommodation types vary in their degree of privacy, we expect there to be a significant difference between their listing prices

(Gibbs et al, 2017). Additionally, as Airbnb hosts who manage entire apartments or houses have a need to meet financing or lease obligations, they may be more concerned in ensuring sufficient demand leading them to charge lower prices to "fill their beds" (Gibbs et al, 2017). Moreover, nightly rates can also vary due to different fees charged by both Airbnb and the hosts[1]. For example, guests booking an entire apartment are charged higher rental compared to guests booking only a private room in the same apartment. This is to compensate for more costly cleaning fee and utility costs such as electricity and water. Therefore, we believe that the type of accommodation will have a significant impact on the listing price, which is why we include it into our regression model. We use the dummy variable for private room and the dummy variable for entire apartment or house in the model. As a result, we can check the partial effects of distance and number of reviews on rental price while controlling for the room type. We have excluded the dummy variable for shared rooms as the reference category to avoid the problem of perfect multicollinearity among the dummy variables for room type. The multivariate linear model becomes:

$$Pr_i = \beta_0 + \beta_1 Dist_i + \beta_2 Rev_i + \beta_3 Ent_i + \beta_4 Prive_i + \varepsilon_i$$

The choice of model for this research is a difficult task as there are not so many clear explanations on why a specific form is preferred to the others. However, among the famously-used models, namely linear model, log-level, or double log, we adopt the second because it serves well four important criteria: being consistent with the economy theory, flexible enough to fit the data, satisfy all regression assumptions, and simple to interpret (Bhurtel, 2015). The study Sirmans et al (2005) on 125 previous research concluded that "most researchers used log-linear model because this model allowed for variation in characteristic prices across different price ranges within the sample and also helped to minimize the problem of heteroscedasticity" (Bhurtel, 2015). In the joint efforts of Hea (2010) with other researchers, he estimated both level-level and semi-log regression equations and concluded that "regression results show that semi logarithmic regression (based on wavelet transformation and denoising) has more explanatory power than other methods".

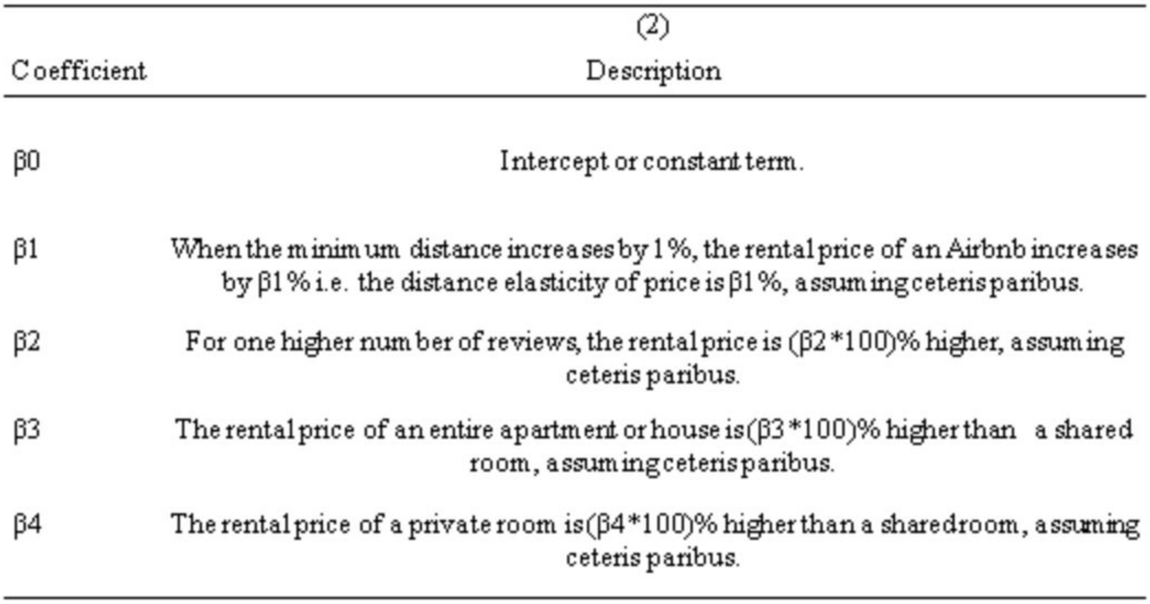
The functional form to be used in our model will include a log-log specification on the variables Pr, the Airbnb listing price per night, and Dist, minimum distance from the Airbnb listing to the respective benchmark position in each borough. We do this in light of evidence from Cui et al (2018), in which they found that the relationship between price and distance from the city center increases with significance the closer the listing is to the city center due to sensitive crime rate and safety concerns. A log-log specification on the variables Pr and Dist will allow for this effect for our study. This specification will also allow us to estimate a constant distance elasticity of price, as given by the estimated coefficient. Although we mostly rely on economic reasoning to decide which model to choose, we also study the regression results of different functional forms of the regression model, which are presented in Appendix B. We only

[1] See more at https://www.airbnb.com/help/article/1857/what-are-airbnb-service-fees

consider the last four functional forms with log specification on the dependent variable Pr in the Table B1. The other functional forms are not considered in light of the reasoning given earlier. Since the natural logarithm of zero is undefined, the specification of logarithm should not be used for a variable with zero values, otherwise regression results and hypothesis testing will be incorrect. As 20.56% of Airbnb properties have zero number of reviews in the sample, we do not consider a functional form with log specification on number of reviews. Although the R-squared of log-level specification on the variables Pr and Dist is higher than the R-squared of log-log specification on the variables Pr and Dist, we use the log-log specification on the variables Pr and Dist because the difference in R-squared values is quite small to have an adverse effect on hypothesis testing and, more importantly, this specification is also more consistent with the economic reasoning discussed earlier and will allow to calculate distance elasticity. So the final multivariate model we chose is:

$$ln(Pr_i) = \beta_0 + \beta_1 ln(Dist_i) + \beta_2 Rev_i + \beta_3 Ent_i + \beta_4 Prive_i + \varepsilon_i$$

The description of the coefficients β_0, β_1, β_2, β_3 and β_4 is given in the table below.



	(2)
Coefficient	Description
β0	Intercept or constant term.
β1	When the minimum distance increases by 1%, the rental price of an Airbnb increases by β1% i.e. the distance elasticity of price is β1%, assuming ceteris paribus.
β2	For one higher number of reviews, the rental price is (β2*100)% higher, assuming ceteris paribus.
β3	The rental price of an entire apartment or house is(β3*100)% higher than a shared room, assuming ceteris paribus.
β4	The rental price of a private room is(β4*100)% higher than a sharedroom, assuming ceteris paribus.

Table. 4. Model coefficients and their respective descriptions.

Note. Coefficients are assumed to be positive in the description. Vice versa is true if a coefficient is negative.

3.2 OLS assumptions and multicollinearity

The classical assumptions for OLS are discussed in this section. First four assumptions must be satisfied for OLS to be an unbiased estimator, and if the fifth assumption is added, OLS will be the best linear unbiased estimator with the smallest variance. The sixth assumption must be satisfied to perform hypothesis testing.

Multicollinearity is also discussed in this section. Perfect multicollinearity is discussed under assumption 4. Imperfect multicollinearity is discussed under imperfect multicollinearity. Imperfect multicollinearity does not violate any OLS assumptions, but it still may be a concern.

Assumption 1: The population model is linear in parameters, is correctly specified and has an additive term.

The specification of the model is discussed in depth in regression model 3.1 and is assumed to be correct. The assumed model is linear in the coefficients β_0, β_1, β_2, β_3 and β_4 and the stochastic error term is included in the model.

Assumption 2: The error term has a zero population mean.

Intercept coefficient β_o is included in the model. This assumption is met as long as a constant coefficient is included in the model (Studenmund, 2016, p.95).

Assumption 3: All independent variables are uncorrelated with the error term.

This assumption states that the independent variables are exogenous. This assumption can fail due to omitted variable bias. The omitted variable problem will be discussed in the proceeding section.

Assumption 4: No independent variable is a perfect linear function of any other explanatory variable(s) (no perfect multicollinearity).

If perfect multicollinearity exists, the OLS estimator will be incapable of distinguishing one variable from the other, because every moment of one of the variables is matched by the relative movement in the other (Studenmund, 2016, p. 98). We check the correlation coefficient between the independent variables to diagnose perfect multicollinearity. Perfect multicollinearity exists if a correlation coefficient is approximately equal to 1 or -1. As it can be seen in the Table 6, none of the correlations between independent variables is approximately equal to 1 or -1. This is a proof that perfect multicollinearity does not exist in the model. Another way to diagnose perfect multicollinearity is to run an auxiliary regression of each independent variable on the other independent variables. R-squared equal to 1 for an auxiliary regression confirms that the perfect multicollinearity exists. Four different auxiliary regressions for the independent variables in the model are run and results are presented in Appendix C. The results show that none of the R-squared values is equal to 1. This is another proof that perfect multicollinearity does not exist in the model. Therefore, the 4th assumption is satisfied. Note that perfect multicollinearity exists between the dummy variables for room type. This is why the dummy variable for shared rooms is excluded from the model and is considered as a reference category.

Assumption 5: The error term has a constant variance (no heteroskedasticity).

If this assumption is violated, OLS is still an unbiased estimator of β_k, but, since the estimated variance of $\widehat{\beta}$ depends on variance of the error term Var (ε_i), it is a biased estimator of the estimated variance of $\widehat{\beta}$.

We test for heteroskedasticity using Breusch-Pagan test. The steps are follows:

1. Regress the following estimated model:

$$ln(Pr_i) = \beta_o + \beta_1 ln(Dist_i) + \beta_2 Rev_i + \beta_3 Ent_i + \beta_4 Prive_i + \varepsilon_i$$

2. Residuals are predicted from the estimated model. These residuals are squared to obtain squared residuals.
3. Regress squared residuals on the independent variables from the original model

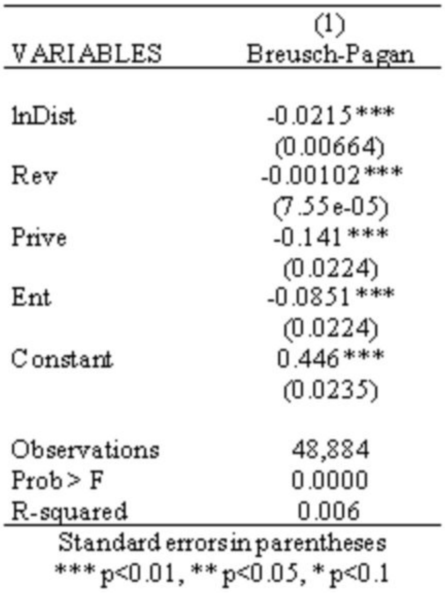
$$e_i^2 = \delta_0 + \delta_1 * \text{lnDist}_i + \delta_2 * \text{Rev}_i + \delta_3 * \text{Prive}_i + \delta_4 * \text{Ent}_i + v_i$$

4. Test whether the independent variables have a jointly significant impact on squared residuals e_i^2. If they do (i.e. H_0 is rejected), we have heteroskedasticity. If they do not (i.e. H_0 is not rejected), we do not have heteroskedasticity.

$$H_0 : \delta_1 = \delta_2 = \delta_3 = \delta_4 = 0 \text{ (homoskedasticity)}$$

$$H_A : H_0 \text{ not true (heteroskedasticity)}$$

The result of the regression of squared residuals on the independent variables is shown below. The explanatory variables are jointly significant, as seen from the model F-test (p-value = 0.0000 < 0.05). This means we can reject the null hypothesis of homoskedasticity: the errors are heteroskedastic. We will use heteroskedasticity-robust standard error for hypothesis testing in the proceeding section.



VARIABLES	(1) Breusch-Pagan
lnDist	-0.0215***
	(0.00664)
Rev	-0.00102***
	(7.55e-05)
Prive	-0.141***
	(0.0224)
Ent	-0.0851***
	(0.0224)
Constant	0.446***
	(0.0235)
Observations	48,884
Prob > F	0.0000
R-squared	0.006

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table. 5. Breusch-Pagan Test regression results.

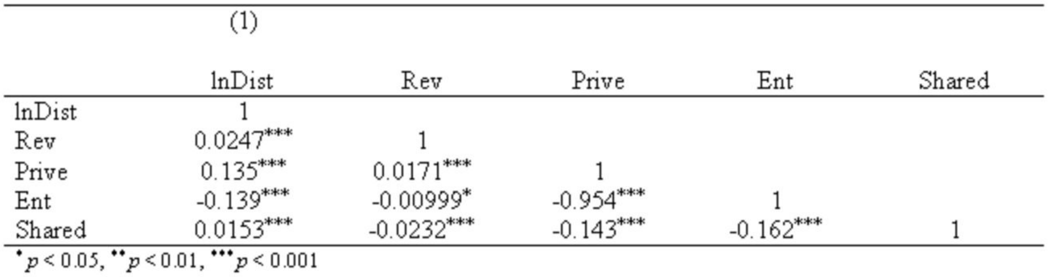
Assumption 6: Error term is normally distributed with a mean of 0 and a constant variance σ²

Central Limit Theorem (CLT) tells us that the sampling distributions will be normally distributed for large sample sizes. Since the dataset has a larger number of observations (more than 48,000 observations), we can assume that the sampling distribution of the error term is normally distributed.

Imperfect multicollinearity

Imperfect multicollinearity occurs when two or more explanatory variables are imperfectly linearly related. If multicollinearity exists, the expected variance and standard errors of the estimates will increase. This makes it more difficult to reject the null hypothesis that a particular independent variable has no impact on the dependent variable.

One way to diagnose multicollinearity is to check the correlation coefficient among independent variables. The correlation coefficients among all independent variables are low, except for the correlation coefficient between the dummy variable for private rooms and the dummy variable for entire apartment or house. The high correlation coefficient of 0.9535 between the two dummy variables for room type is an indication of severe multicollinearity. There is no certain rule that multicollinearity is problematic if the absolute value of the correlation coefficient exceeds a certain threshold. The correlation coefficient is problematic if it causes large variances in the coefficient estimates. Whether this is problematic in this model or not is discussed later.

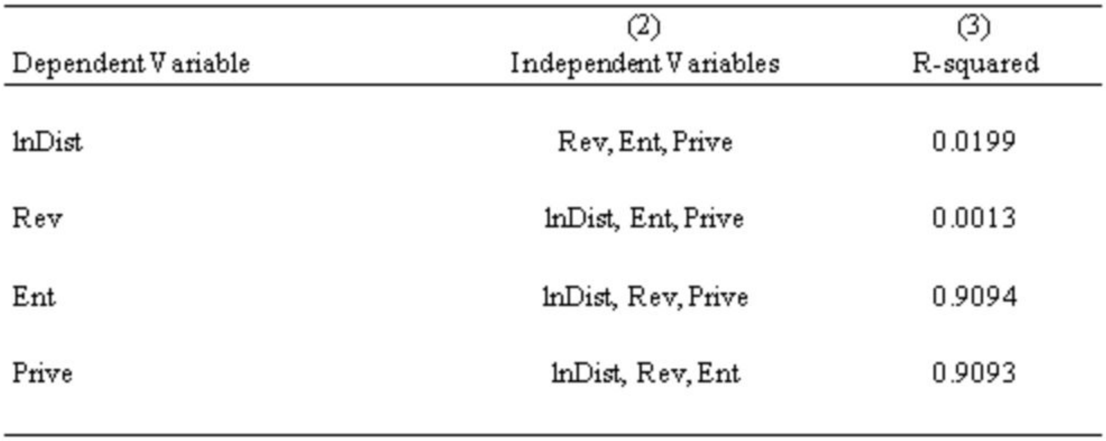


	(1)				
	lnDist	Rev	Prive	Ent	Shared
lnDist	1				
Rev	0.0247***	1			
Prive	0.135***	0.0171***	1		
Ent	-0.139***	-0.00999*	-0.954***	1	
Shared	0.0153***	-0.0232***	-0.143***	-0.162***	1

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table. 6. Correlation table for independent variables of new model.

Another method to diagnose multicollinearity is to estimate an auxiliary regression of each independent variable on the other independent variables. This method is considered a better mechanism to test multicollinearity, because it is possible that groups of independent variables act together to cause multicollinearity without any single simple correlation coefficient being high enough to indicate that multicollinearity is in fact severe. The complete result of auxiliary regressions are in Appendix C. The table below shows these auxiliary regressions and their corresponding R-squared. R-squared of the first auxiliary regression of lnDist on Rev, Ent and Prive and of the second auxiliary regression of Rev on lnDist, Ent and Price is 0.0199 and 0.0013 respectively. The low R-squared values show that little variation in the natural logarithm of minimum distance lnDist and number of reviews Rev is explained by the other independent variables and therefore multicollinearity for these models is not severe. R-squared of the third auxiliary regression of Ent on lnDist, Rev and Prive and of the fourth auxiliary regression of Prive on lnDist, Rev and Ent is 0.9094 and 0.9093 respectively. The high R-squared values are the result of high correlation between the dummy variable for private rooms and the dummy variable for the entire apartment or house. The high R-squared values show that the most of the variation in the dummy variable for private room Prive and the dummy variable for entire apartment or house is explained by the other independent variables and therefore multicollinearity for these models can be of concern.



	(2)	(3)
Dependent Variable	Independent Variables	R-squared
lnDist	Rev, Ent, Prive	0.0199
Rev	lnDist, Ent, Prive	0.0013
Ent	lnDist, Rev, Prive	0.9094
Prive	lnDist, Rev, Ent	0.9093

Table. 7. Summary of auxiliary regressions.

Whether the severity of multicollinearity indicated by the results of correlation coefficients and auxiliary regression is problematic for hypothesis testing is discussed in this paragraph. A remedy for multicollinearity should be considered only if the consequences cause insignificant t-scores or unreliable estimated coefficients (Studenmund, 2016, p. 236). The results in the Table 8 show that the t-statistics for all four estimated coefficients are significant. Although multicollinearity among some of the independent variables is high, this result shows that it is not problematic for our model. The most likely explanation for this result is that we have a large sample with more than 48000 observations. The larger the sample the higher is the total sum of squares for an independent variable (TSS). As a result, the standard errors of the estimated coefficients are low which counteract the impact of the multicollinearity. Therefore, no remedy is needed for hypothesis testing.

3.3 Hypothesis testing

In the following hypothesis testing procedures, we use the significance level of 5%. The number of observations is 48,884

3.3.1 Testing to add variables

In this paper, our principal concern is to observe the relationship between an Airbnb accommodation’s location and its price per night. Hence, the initial assumed model is

$$ln(Pr_i) = \beta_o + \beta_1 ln(Dist_i) + \varepsilon_i$$

Most phenomena (outcomes) depend on multiple factors, and in a complex manner. The rental of a property does not depend solely on its location as there are many other factors that should also be considered in order to operate an available-for-rent accommodation. In order to decide if a certain independent variable should be added into the model, F-test is used to examine its helpfulness in explaining the sampling data, and therefore the population. Hence, a variable is added only when it is proved to have a significant impact on the dependent variable, or the price per night in this case.

First, as mentioned in the theory section and the regression model section, the number of non-negative reviews is expected to have a significant effect on the price per night.

Restricted model (1 restriction)	**Unrestricted model**
$ln(Pr_i) = \beta_o + \beta_1 ln(Dist_i) + \varepsilon_i$	$ln(Pr_i) = \beta_o + \beta_1 ln(Dist_i) + \beta_2 Rev_i + \varepsilon_i$
($\overline{R}^2$ = 0.0598; k = 1; n = 44884)	($\overline{R}^2$ = 0.0611; k = 2; n = 44884)

The incorporation of new variable Rev does improve the goodness-of-fit of our model, indicated by the higher adjusted R-squared (0.0598 > 0.0611). It drives us to conduct an F-test to confirm this hypothesis.

Hypothesis: *The number of non-negative reviews of an Airbnb accommodation has a significant impact on its price per night.*

$$H_0 : \beta_2 = 0$$
$$H_{A:} \beta_2 \neq 0$$

Critical value: $F_c = F_{M,\, n-k-1,\, \alpha} = F_{1,\ 4881,\ 0.05} = 3.84$

Test statistic: $F = \frac{(RSS_M - RSS)/M}{RSS/(n-k-1)} = \frac{(22405.52-22374.1405)/1}{22374.1405/(48884-2-1)} = 68.555$ [2]

Since $F > F_c$ (68.555 > 3.84), reject H_0.

Conclusion: The restricted model is not true. Variable Rev has a significant impact on price per night.

Moreover, the difference in rental between Airbnb properties is also subjected to different types of rooms offered. Accordingly, a test of adding the dummy variables of room type to the model should be conducted.

Restricted model (2 restrictions)	**Unrestricted model**
$ln(Pr_i) = \beta_o + \beta_1 ln(Dist_i) + \beta_2 Rev_i + \varepsilon_i$	$ln(Pr_i) = \beta_o + \beta_1 ln(Dist_i) + \beta_2 Rev_i + \beta_3 Ent_i + \beta_4 Prive_i + \varepsilon_i$
($\overline{R}^2$ = 0.0611; k = 2; n = 44884)	($\overline{R}^2$ = 0.4128; k = 4; n = 44884)

[2] See *Table A1* and *Table A2*.

In comparison with our previous model with adjusted R-squared of 0.0611, adding the two dummy variables does improve the overall fitness of the model, indicated by the new adjusted R-squared of 0.4128. The following hypothesis testing will be conducted to confirm our result.

Hypothesis: *The dummy variables for room type are relevant in the regression model.*

$$H_0 : \ \beta_3 = \beta_4 = 0$$
$$H_{A:} \ \beta_3 \neq 0 \ or \ \beta_4 \neq 0$$

Critical value: $F_c = F_{M, n-k-1, \ \alpha} = F_{2, \ 48879, \ 0.05} = 3$

Test statistic: $F = \frac{(RSS_M - RSS)/M}{RSS/(n-k-1)} = \frac{(22374.1405 - 13993.4525)/2}{13993.4525/(48884-4-1)} = 14636.83$ [3]

Since $F > F_c$ (14636.83 > 3), reject H_0.

Conclusion: The restricted model is not true. There is enough statistical evidence to conclude at 5% significance level that the dummy variables are relevant in our model.

Additionally, the question of whether there are differences in determining the nightly rate of an entire apartment and a single room (including private room and shared room) is not answered yet.

Restricted model	**Unrestricted models**
$ln(Pr_i) = \ \beta_o + \ \beta_1 ln(Dist_i) + \beta_2 Rev_i + \ \varepsilon_i$	$ln(Pr_i) = \ \beta_o^E + \ \beta_1^E ln(Dist_i) + \beta_2^E Rev_i \ + \varepsilon_i$ *(for entire apartment)* $ln(Pr_i) = \ \beta_o^S + \ \beta_1^S ln(Dist_i) + \beta_2^S Rev_i \ + \varepsilon_i$ *(for single room)*

Therefore, Chow test will be used for the following hypothesis testing.

Hypothesis: *Regression functions for an entire apartment and a single room are different.*

$$H_0 : \ \beta_o^E = \beta_o^S, \ \ \beta_1^E = \beta_1^S, \ \beta_2^E = \beta_2^S$$
$$H_{A:} \ H_0 \ not \ true$$

Critical value: $F_c = F_{M, n-k-1, \ \alpha} = F_{2, \ 48879, \ 0.05} = 3$

Test statistic: $F = \frac{(RSS_M - RSS_1 - RSS_2)/(k+1)}{(RSS_1 + RSS_2/(n_1 + n_2 - 2(k+1))} = \frac{(22374.14 - 7897.25 - 6185.23)/(2+1)}{(7897.25+6185.23)/(25407+23477-2(2+1))} = 9591.84$ [4]

Since $F > F_c$ (9591.84 > 3), reject H_0.

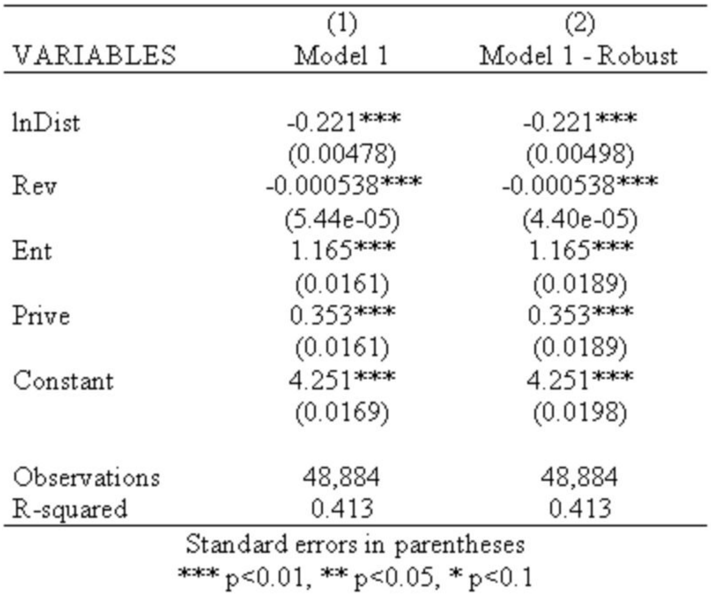
Conclusion: *Entire apartment and single room have significantly different regression functions.*

[3] See *Table A1* and *Table A3*.

[4] See *Table D1*, *Table D2* and *Table D3*.

3.3.2 *Testing partial effects of coefficients*

Based on OLS assumptions testing above, our model violated the assumption of homoscedasticity. Therefore, heteroskedasticity-robust standard errors are calculated to correct the significance of all coefficients.



VARIABLES	(1) Model 1	(2) Model 1 - Robust
lnDist	-0.221***	-0.221***
	(0.00478)	(0.00498)
Rev	-0.000538***	-0.000538***
	(5.44e-05)	(4.40e-05)
Ent	1.165***	1.165***
	(0.0161)	(0.0189)
Prive	0.353***	0.353***
	(0.0161)	(0.0189)
Constant	4.251***	4.251***
	(0.0169)	(0.0198)
Observations	48,884	48,884
R-squared	0.413	0.413

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table. 8. Semi-log model.

Hypothesis: *The distance between an Airbnb accommodation and its neighborhood's center has a significant impact on its price per night.*

$$H_0 : \beta_1 = 0$$
$$H_{A:} \beta_1 \neq 0$$

Critical value: $t_c = t_{n-k-1,\ \alpha/2} = t_{48879,\ 0.025} = 1.96$

Test statistic: $t = -44.31$

Since $|t| > t_c$ (44.31 > 1.96), reject H_0.

Conclusion: There is enough statistical evidence to conclude that the location has an effect on price per night at 5% level of significance, all else equal.

Hypothesis: *The number of non-negative reviews has a significant impact on the price per night.*

$$H_0 : \beta_2 = 0$$
$$H_{A:} \beta_2 \neq 0$$

Test statistic: $p - value = 0.000$

Since *p-value* $< \alpha$ (0.000 < 0.05)[5], reject H_0.

[5] See *Table 8*.

Conclusion: There is enough statistical evidence to conclude that the number of non-negative reviews has an effect on price per night at 5% level of significance, all else equal.

3.3.3 Testing overall significance of all coefficients

The model F-test is used for the overall significance test for the purpose of examining the joint impact of all included independent variables.
For now, our linear model is

$$Pr_i = \beta_o + \beta_1 Dist_i + \beta_2 Rev_i + \beta_3 Ent_i + \beta_4 Prive_i + \varepsilon_i$$

Hypothesis: *Independent variables have a joint impact on the price per night.*

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$$
$$H_{A:}\ H_0 \textit{ is not true}$$

Test statistic: $p - value = 0.0000$ [6]

Since $p\text{-}value < \alpha$ $(0.0000 < 0.05)$, reject H_0.

Conclusion: All independent variables $(Dist, Rev, Ent, Prive)$ are jointly statistically significant at 5% level of significance.

3.4 Interpretations

The final model is estimated as

$$ln(Pr_i) = 4.251 - 0.2207\ ln(Dist_i) - 0.00053\beta_2 Rev_i + 1.1653 Ent_i + 0.3528 Prive_i + \epsilon_i$$

Overall fitness

The R-squared of 0.4128 is observed in our final model, meaning that 41.28% of the variation in the dependent variable $ln(Pr)$ is explained by the joint impact of four independent variables.

Estimated coefficients & practical implications of the model

- The change of $ln(Dist)$ has a negative impact on $ln(Pr)$**.** The coefficient $\hat{\beta}_1$ (-0.2207) indicates that if the distance from an Airbnb to the nearest neighborhood's center increases by 1%, its price per night is 0.22% less expensive compared to other Airbnb listings, all else equal.
- There is a negative relationship between $ln(Pr)$ and the change in Rev**.** The coefficient $\hat{\beta}_2$ (-0.00053) indicates that if an Airbnb has one more non-negative review, its price per night is lower by 0.053%, all else equal.
- The coefficients of dummy variables Ent and $Prive$ interpret the increase in the dependent variable $ln(Pr)$ caused by the condition of an entire apartment/house or a

[6] See *Table A4*.

private room being met, compared to the omitted condition, which is when a shared room is booked. The coefficient $\hat{\beta}_3$ (1.1653) shows that an entire apartment/house is 116.53% more expensive than a shared room, while $\hat{\beta}_4$ (0.3528) suggests the price for a private room is 35.28% higher than that of a shared room, all else equal.

The negative relationship between the distance from an Airbnb to the nearest neighborhood's center and nightly rate closely follow our expectation of "the more central the place, the higher the rental". However, as discussed previously, Airbnb reviews are assumed to be immensely non-negative. Then, why is there a statistically significant negative relationship between the number of reviews and the Airbnb listing price?

This could be explained by the simple Economic Law of Demand, as listings with lower prices will receive more customers. Hosts of the real estate manager type need to meet financing or lease obligations, thus they are more likely to charge lower prices to "fill their beds" and generate more demand (Gibbs et al, 2017). The negative relationship can also be explained by looking at the opposing host type which is casual hosts. Since these hosts need to carefully think through the risks and effort of hosting others in their property, to make it "worth their while" they may charge higher prices and more carefully screen guests thus are more likely to reject reservation requests (Gibbs et al, 2017). Furthermore, Airbnb guests are more likely to desire "unique" experiences. Unique properties such as villas and yachts, which have a lower number of reviews, are listed with higher prices as hosts are aware that guests will pay more for the experience (Dogru and Pekin, 2016). We may also have committed *Omitted Variable Bias* leading to the estimated effect of the number of reviews on price being much too negative than it actually is. For instance, criminality would have a positive relationship with the number of reviews as with higher crime rates in an area, there is likely to be a high number of reviews due to the complaints and warnings from previous guests. Criminality would also have a negative relationship with the price, as guests are less likely to stay in a property in a high crime rate area, thus hosts would have to reduce listing prices to maintain demand.

Finally, if solving the aforementioned omitting variable problem, by adding a new explanatory variable concerning criminality, does not help confirming our expectation, there might be a chance that our initial assumption of non-negative reviews only does not hold. Rather, it may increase the probability of customer complaints which eventually discourages potential newcomers and drives down the property's value. Nevertheless, in some extreme cases online reviews and ratings are proved to have no impact on the listing price: "*These results can be attributed to the fact that, on average, Airbnb hosts have a rating of 4.5 out of 5, which is very extreme compared to hotel firms' ratings (Zervas, Proserpio, & Byers, 2015)*.

5. Conclusion

Our research has given a hint to explain the price-setting strategies of the rental online marketplace company, Airbnb. Based on a number of economic theories suggested by Zhang (2017), Gibbs & Guttentag (2017), Cui (2018), Dan Hill (2015), and many other researchers, we generate two postulations regarding the relationship between the dependent variable, property price, and dependent variables, namely location, number of reviews, and room types. Through the process of analyzing data in Section 2 and Section 3.1, we formulate the general model equation in Section 3.2. Our option for the semi-log specification is robustly supported by previous findings by Bhurtel (2015), Hea (2010), and Cui et al (2018). We then perform several hypothesis testings to examine the relevance of included variables as well as the effects of explanatory variables on the dependent variable.

The outcomes of t-test and F-test reassure the unbiasedness of the estimated parameters and reliability of the OLS estimator. We, therefore, can infer three main conclusions out of our hypothesized model. First, our first hypothesis suggesting that location has a negative effect on the price of an Airbnb accommodation is correct. This indicates that the further distance from an Airbnb to the nearest neighborhood's center, the less it will cost, ceteris paribus. Much to our surprise, the second hypothesis is proven to be incorrect. The number of reviews, despite having a significant impact, is negatively correlated with the property price. Lastly, customers have to pay more for renting an entire apartment/house compared to a shared or private room.

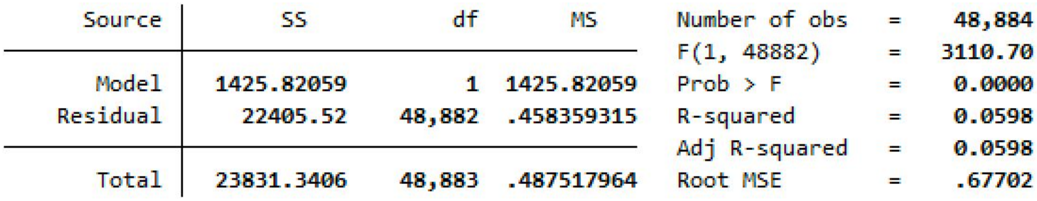
Nevertheless, we acknowledge important limitations of this study. Firstly, we are employing two dummy variables to consider the three types of Airbnb listings in our sample, i.e. private room, shared room, and entire apartment, whereas the Chow Test we learned could only accommodate one dummy variable. As such, we were unable to test whether the regression functions for private rooms, shared rooms, and entire apartments are statistically significantly different.

Secondly, the R-squared value obtained indicates that other determinants of Airbnb listing prices have not been fully explored. One important factor we excluded is the host's ethnicity, especially considering that one borough in particular, Queens, is the most ethnically diverse urban area in the world. It is also found that non-African-American Airbnb hosts charge approximately 12% more for the equivalent rental in New York City (Edelman and Luca, 2014). This limitation of excluding relevant determinants of Airbnb listing prices would be increasingly problematic if the omitted factor is correlated with one of our explanatory variables, as this would lead to an omitted variable bias. Although the impact of the borough was considered, the interactions between the borough and other pricing determinants have also not yet been fully explored. Further research would have to be conducted to investigate the excluded determinants and the borough interaction effects.

Another limitation would be regarding our reference points in each borough. We did not conduct research examining the population density and concentration of urban activities which may be of interest to tourists to determine the reference point. As such, the reference points used may not be entirely appropriate. The regression model used also assumes that Airbnb hosts cater to only one uniform consumer market. Visitors to the City can vary widely from business travelers to international tourists (New York Tourism Report, 2019). Furthermore, we recognise that Airbnb price determinant relationships vary significantly across cities due to the variation in city types and Airbnb presence, and across time periods due to the seasonality of Airbnb prices, thus our results may not be easily generalisable.

Taking data from Rome, one of the hardest hit cities by the COVID-19 pandemic, listing prices were 11% lower from March 2019 to March 2020 (Johnson and Ghiglione, 2020). A particularly interesting idea for future research may include examining the effects of the pandemic on Airbnb and other sharing economy platforms and the extent of the change in hosts' pricing strategies.

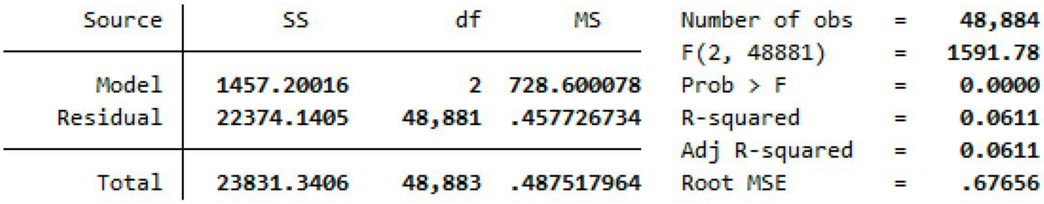
APPENDIX A. Multiple regression models



```
      Source |       SS           df       MS      Number of obs   =    48,884
-------------+----------------------------------   F(1, 48882)     =   3110.70
       Model |  1425.82059         1  1425.82059   Prob > F        =    0.0000
    Residual |    22405.52    48,882  .458359315   R-squared       =    0.0598
-------------+----------------------------------   Adj R-squared   =    0.0598
       Total |  23831.3406    48,883  .487517964   Root MSE        =    .67702
```

Note. Regress *logarithm of price per night* on logarithm of *minimum distance.*

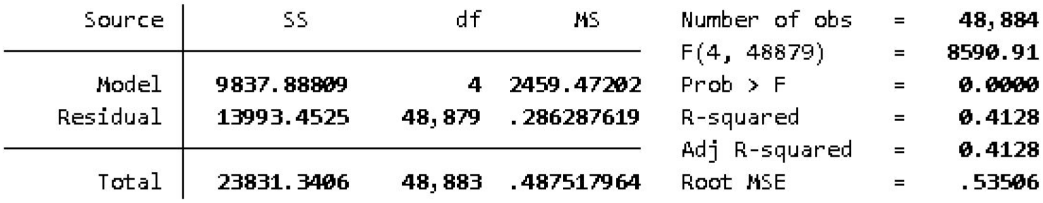
Table A1. Restricted model with 1 restriction



```
      Source |       SS           df       MS      Number of obs   =    48,884
-------------+----------------------------------   F(2, 48881)     =   1591.78
       Model |  1457.20016         2  728.600078   Prob > F        =    0.0000
    Residual |  22374.1405    48,881  .457726734   R-squared       =    0.0611
-------------+----------------------------------   Adj R-squared   =    0.0611
       Total |  23831.3406    48,883  .487517964   Root MSE        =    .67656
```

Note. Regress *logarithm of price per night* on *logarithm of minimum distance,* and *number of reviews.*

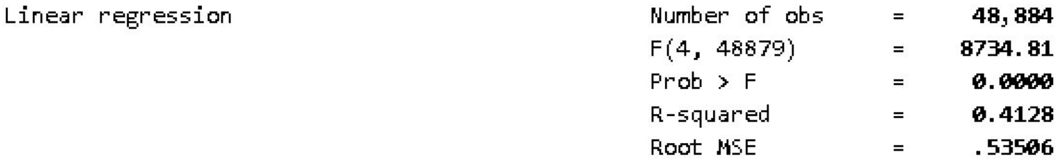
Table A2. Restricted model with 2 restrictions



```
      Source |       SS           df       MS      Number of obs   =    48,884
-------------+----------------------------------   F(4, 48879)     =   8590.91
       Model |  9837.88809         4  2459.47202   Prob > F        =    0.0000
    Residual |  13993.4525    48,879  .286287619   R-squared       =    0.4128
-------------+----------------------------------   Adj R-squared   =    0.4128
       Total |  23831.3406    48,883  .487517964   Root MSE        =    .53506
```

Note. Regress *logarithm of price per night* on *logarithm of minimum distance, number of reviews, and 2 dummies Ent and Prive.*

Table A3. Unrestricted model with heteroskedaticity.

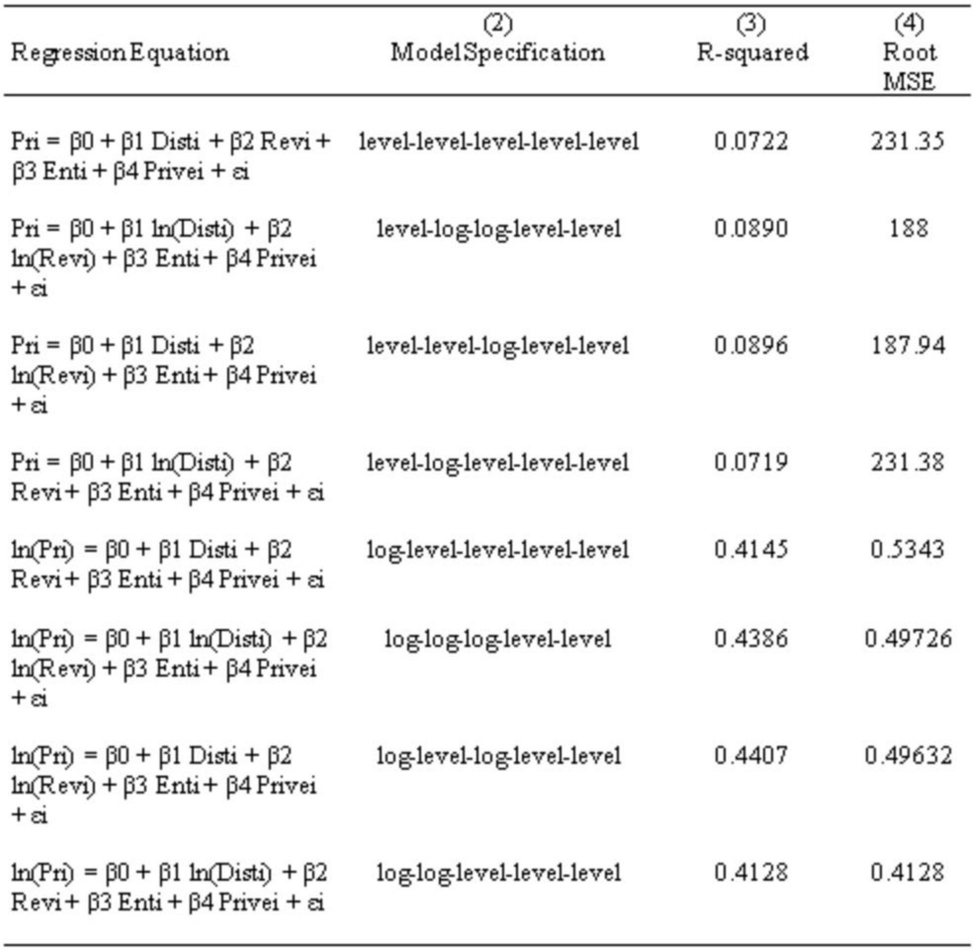


```
Linear regression                               Number of obs     =     48,884
                                                F(4, 48879)       =    8734.81
                                                Prob > F          =     0.0000
                                                R-squared         =     0.4128
                                                Root MSE          =     .53506
```

Note. Regress *logarithm of price per night* on *logarithm of minimum distance, number of reviews, and 2 dummies Ent and Prive, after correction for heteroskedasticity.*

Table A4. Unrestricted model without heteroskedaticity.

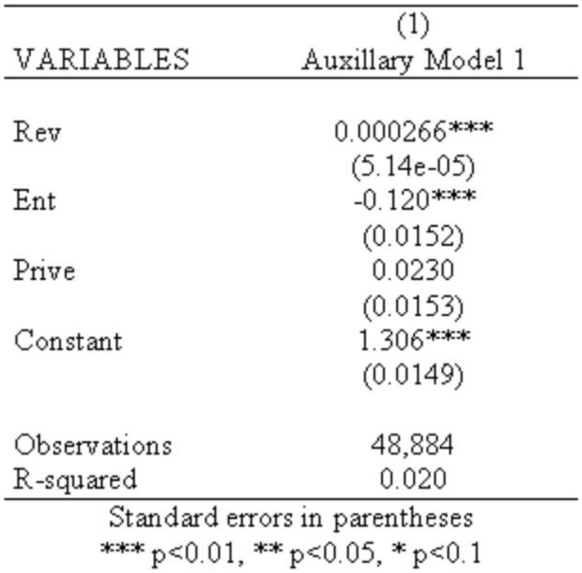
APPENDIX B. Functional forms of the regression model



	(2)	(3)	(4)
Regression Equation	Model Specification	R-squared	Root MSE
Pri = β0 + β1 Disti + β2 Revi + β3 Enti + β4 Privei + εi	level-level-level-level-level	0.0722	231.35
Pri = β0 + β1 ln(Disti) + β2 ln(Revi) + β3 Enti + β4 Privei + εi	level-log-log-level-level	0.0890	188
Pri = β0 + β1 Disti + β2 ln(Revi) + β3 Enti + β4 Privei + εi	level-level-log-level-level	0.0896	187.94
Pri = β0 + β1 ln(Disti) + β2 Revi + β3 Enti + β4 Privei + εi	level-log-level-level-level	0.0719	231.38
ln(Pri) = β0 + β1 Disti + β2 Revi + β3 Enti + β4 Privei + εi	log-level-level-level-level	0.4145	0.5343
ln(Pri) = β0 + β1 ln(Disti) + β2 ln(Revi) + β3 Enti + β4 Privei + εi	log-log-log-level-level	0.4386	0.49726
ln(Pri) = β0 + β1 Disti + β2 ln(Revi) + β3 Enti + β4 Privei + εi	log-level-log-level-level	0.4407	0.49632
ln(Pri) = β0 + β1 ln(Disti) + β2 Revi + β3 Enti + β4 Privei + εi	log-log-level-level-level	0.4128	0.4128

Table B1. R^2 of different functional forms of the regression model:

APPENDIX C. Auxiliary Regression

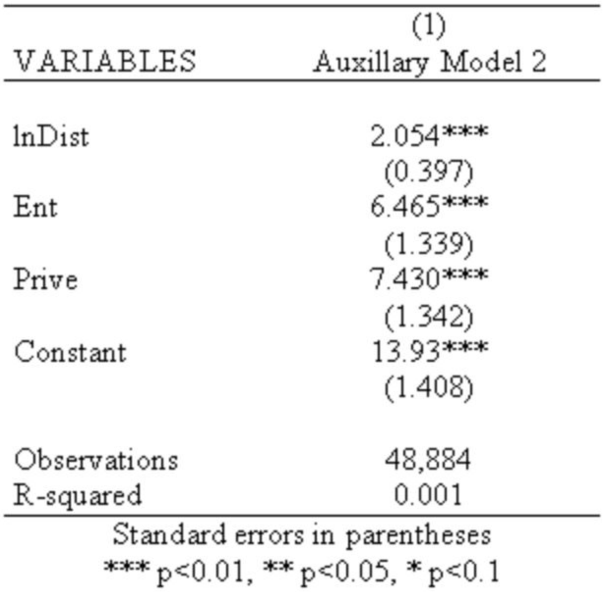


VARIABLES	(1) Auxillary Model 1
Rev	0.000266***
	(5.14e-05)
Ent	-0.120***
	(0.0152)
Prive	0.0230
	(0.0153)
Constant	1.306***
	(0.0149)
Observations	48,884
R-squared	0.020

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table C1. Auxiliary regression model

Note: Auxiliary regression of the natural logarithm of minimum distance on number of reviews and the two dummy variables for room type.

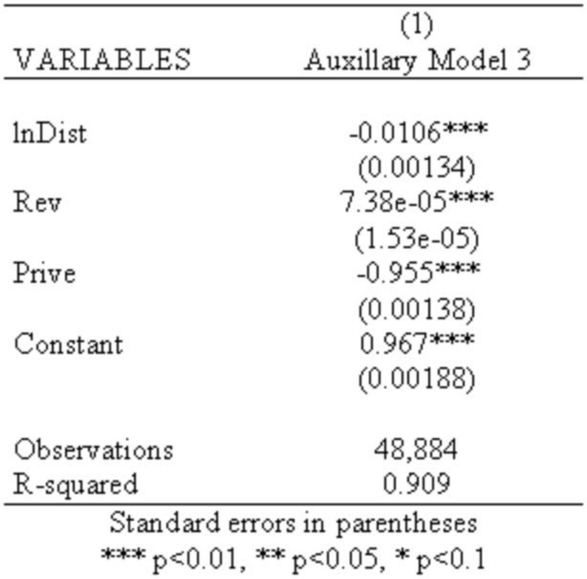


VARIABLES	(1) Auxillary Model 2
lnDist	2.054***
	(0.397)
Ent	6.465***
	(1.339)
Prive	7.430***
	(1.342)
Constant	13.93***
	(1.408)
Observations	48,884
R-squared	0.001

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table C2. Auxiliary regression model

Note: Auxiliary regression of number of reviews on the natural logarithm of minimum distance and the two dummy variables for room type.

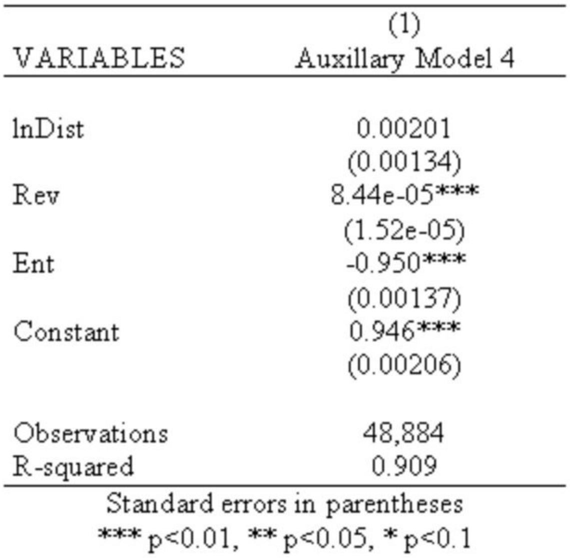


VARIABLES	(1) Auxillary Model 3
lnDist	-0.0106***
	(0.00134)
Rev	7.38e-05***
	(1.53e-05)
Prive	-0.955***
	(0.00138)
Constant	0.967***
	(0.00188)
Observations	48,884
R-squared	0.909

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table C3. Auxiliary regression model

Note: Auxiliary regression of the dummy variable for entire apartment or house on the natural logarithm of minimum distance, number of reviews and the dummy variable for private room.



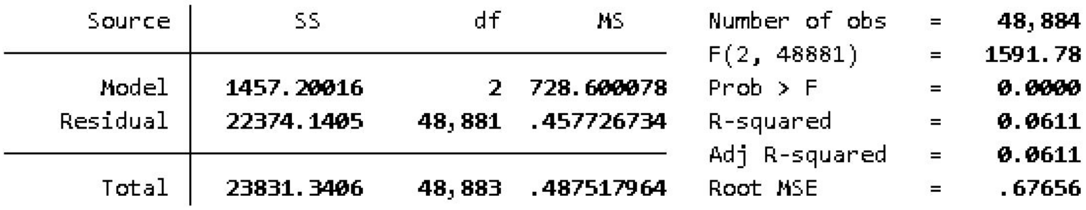
VARIABLES	(1) Auxillary Model 4
lnDist	0.00201
	(0.00134)
Rev	8.44e-05***
	(1.52e-05)
Ent	-0.950***
	(0.00137)
Constant	0.946***
	(0.00206)
Observations	48,884
R-squared	0.909

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table C4. Auxiliary regression model

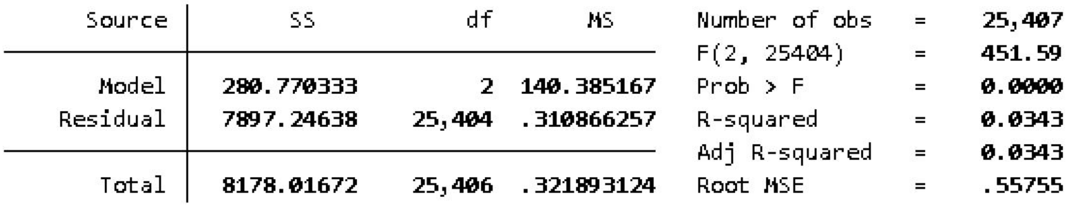
Note: Auxiliary regression of the dummy variable for private room on the natural logarithm of minimum distance, number of reviews and the dummy variable for entire apartment or house.

APPENDIX D. Chow test



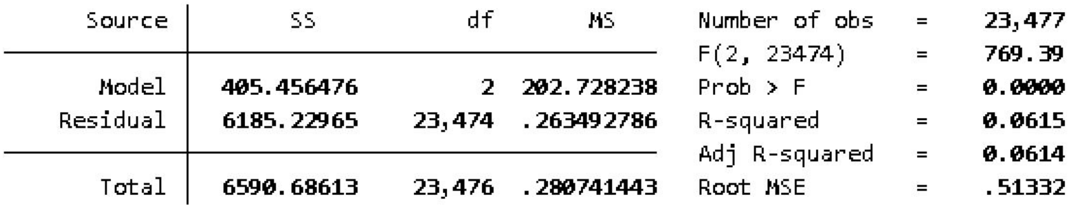
```
      Source |       SS           df       MS      Number of obs   =    48,884
-------------+----------------------------------   F(2, 48881)     =   1591.78
       Model |  1457.20016         2  728.600078   Prob > F        =    0.0000
    Residual |  22374.1405    48,881  .457726734   R-squared       =    0.0611
-------------+----------------------------------   Adj R-squared   =    0.0611
       Total |  23831.3406    48,883  .487517964   Root MSE        =    .67656
```

Table D1. Restricted model.



```
      Source |       SS           df       MS      Number of obs   =    25,407
-------------+----------------------------------   F(2, 25404)     =    451.59
       Model |  280.770333         2  140.385167   Prob > F        =    0.0000
    Residual |  7897.24638    25,404  .310866257   R-squared       =    0.0343
-------------+----------------------------------   Adj R-squared   =    0.0343
       Total |  8178.01672    25,406  .321893124   Root MSE        =    .55755
```

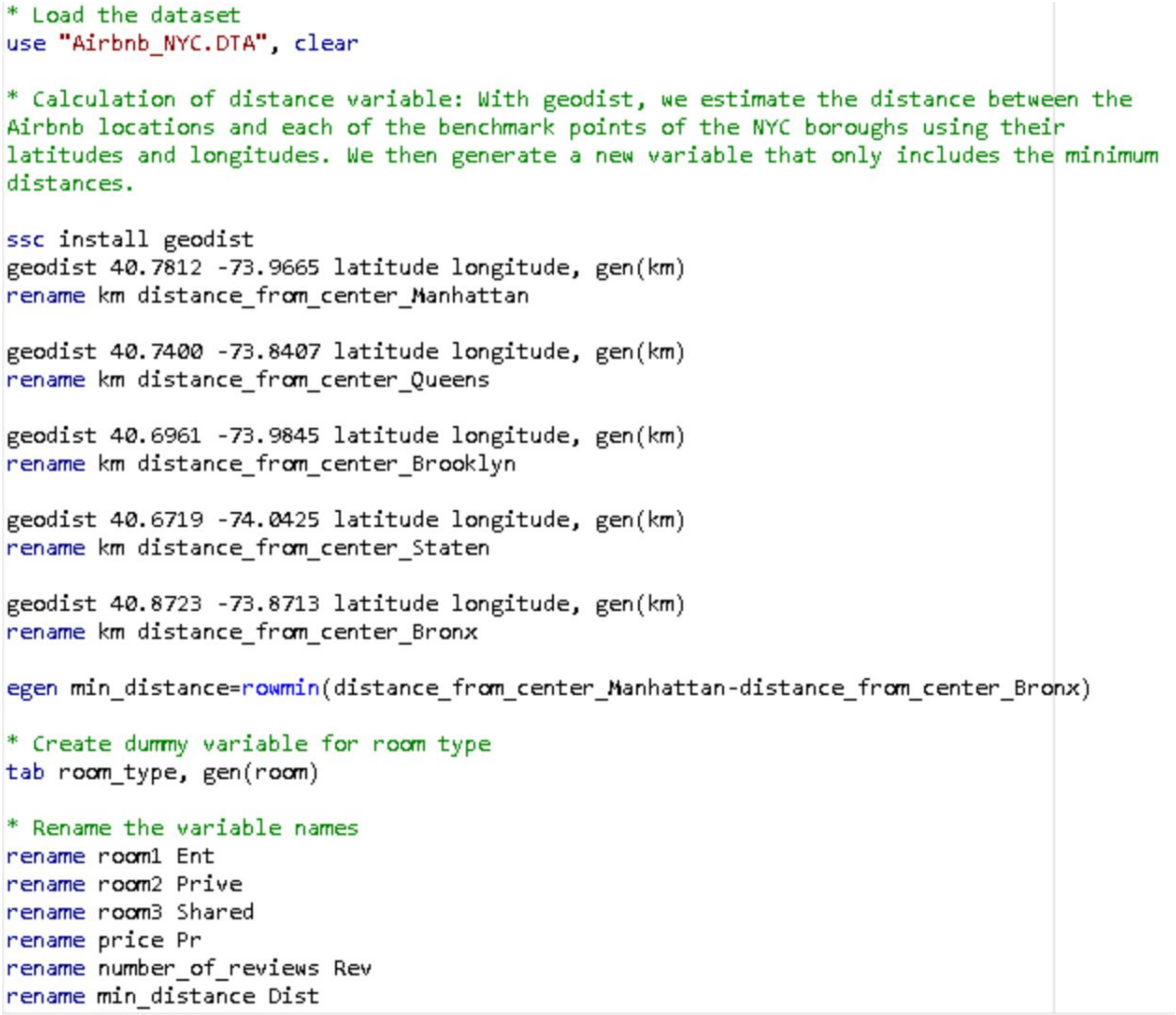
Table D2. Unrestricted model for entire apartment.



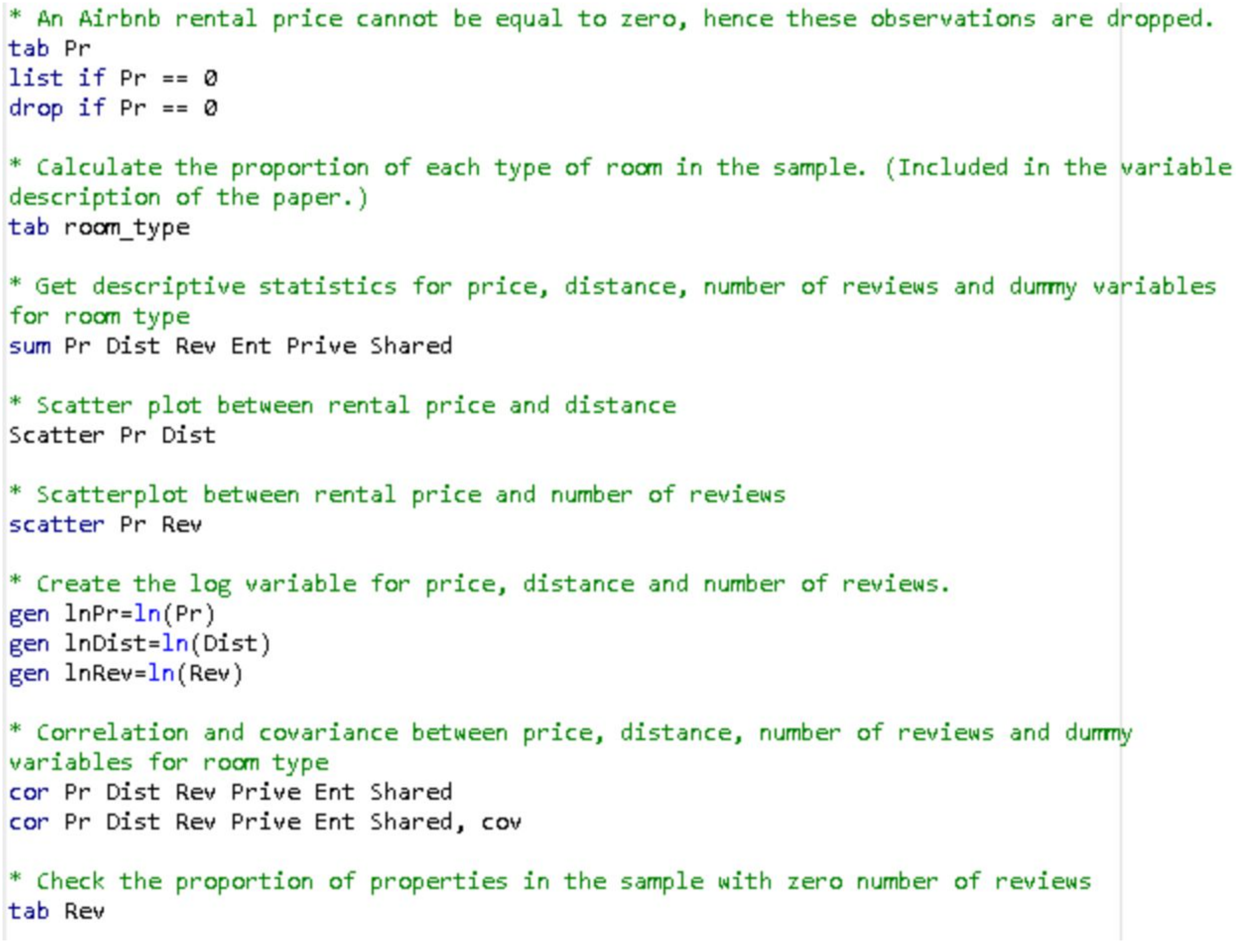
```
      Source |       SS           df       MS      Number of obs   =    23,477
-------------+----------------------------------   F(2, 23474)     =    769.39
       Model |  405.456476         2  202.728238   Prob > F        =    0.0000
    Residual |  6185.22965    23,474  .263492786   R-squared       =    0.0615
-------------+----------------------------------   Adj R-squared   =    0.0614
       Total |  6590.68613    23,476  .280741443   Root MSE        =    .51332
```

Table D3. Unrestricted model for single room.

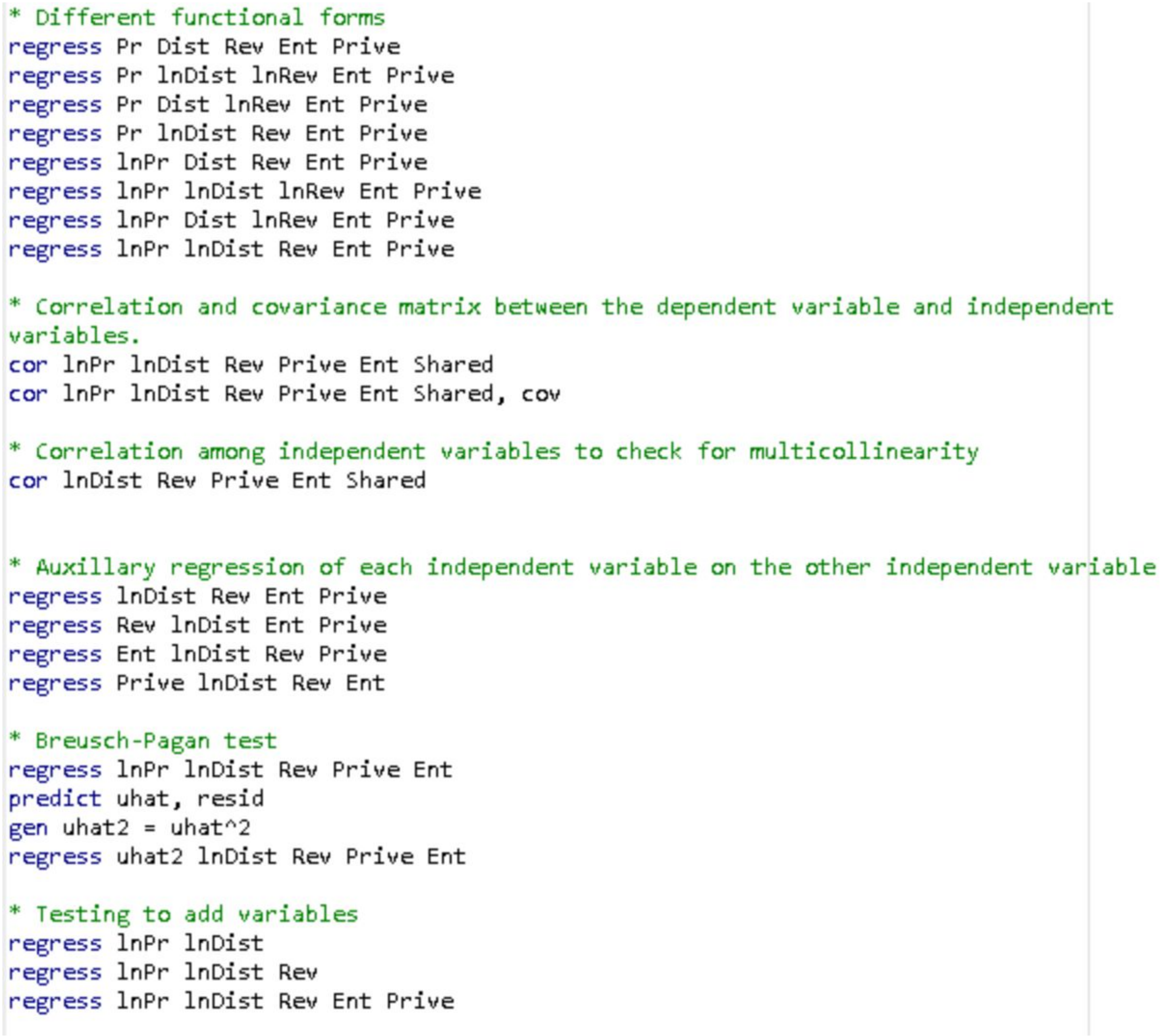
APPENDIX E. Code used on Stata.



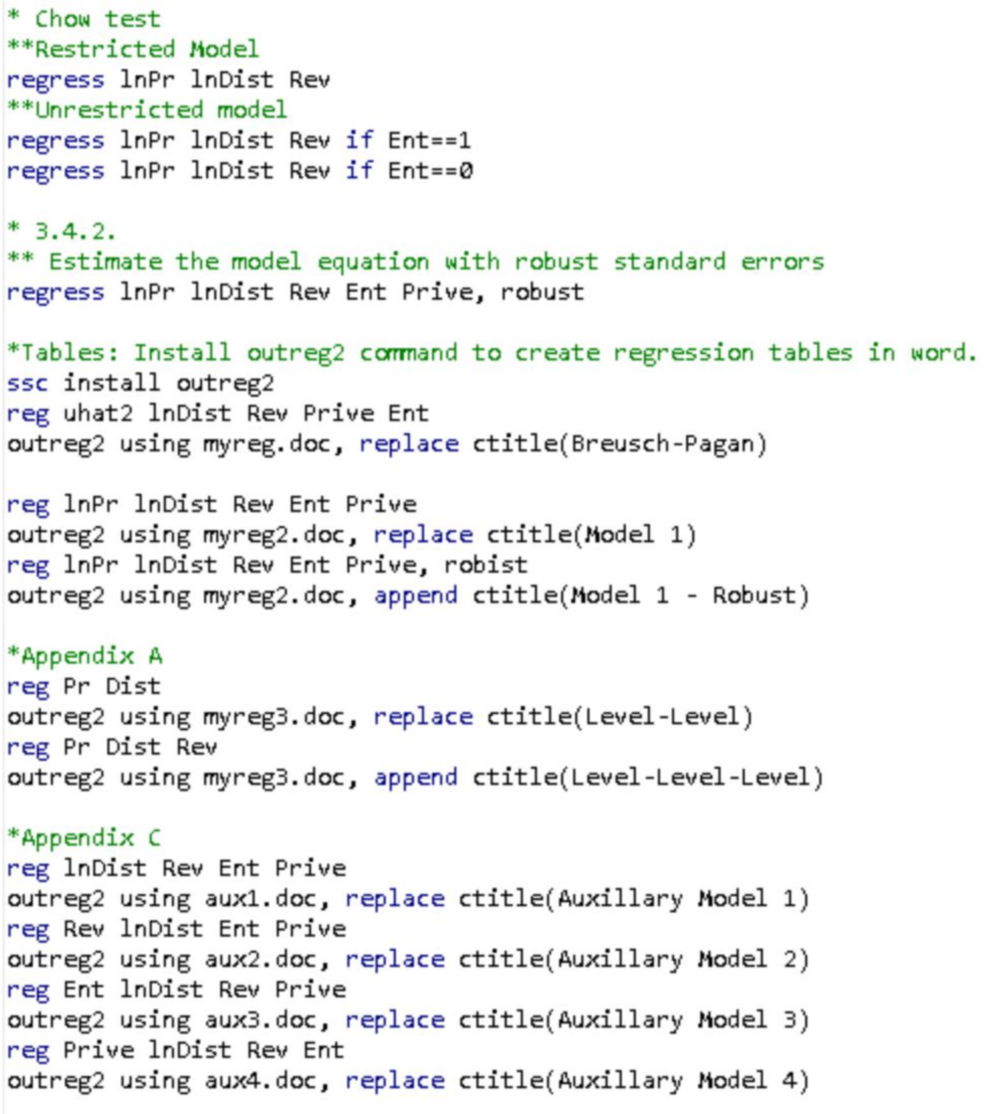
```
* Load the dataset
use "Airbnb_NYC.DTA", clear

* Calculation of distance variable: With geodist, we estimate the distance between the
Airbnb locations and each of the benchmark points of the NYC boroughs using their
latitudes and longitudes. We then generate a new variable that only includes the minimum
distances.

ssc install geodist
geodist 40.7812 -73.9665 latitude longitude, gen(km)
rename km distance_from_center_Manhattan

geodist 40.7400 -73.8407 latitude longitude, gen(km)
rename km distance_from_center_Queens

geodist 40.6961 -73.9845 latitude longitude, gen(km)
rename km distance_from_center_Brooklyn

geodist 40.6719 -74.0425 latitude longitude, gen(km)
rename km distance_from_center_Staten

geodist 40.8723 -73.8713 latitude longitude, gen(km)
rename km distance_from_center_Bronx

egen min_distance=rowmin(distance_from_center_Manhattan-distance_from_center_Bronx)

* Create dummy variable for room type
tab room_type, gen(room)

* Rename the variable names
rename room1 Ent
rename room2 Prive
rename room3 Shared
rename price Pr
rename number_of_reviews Rev
rename min_distance Dist
```



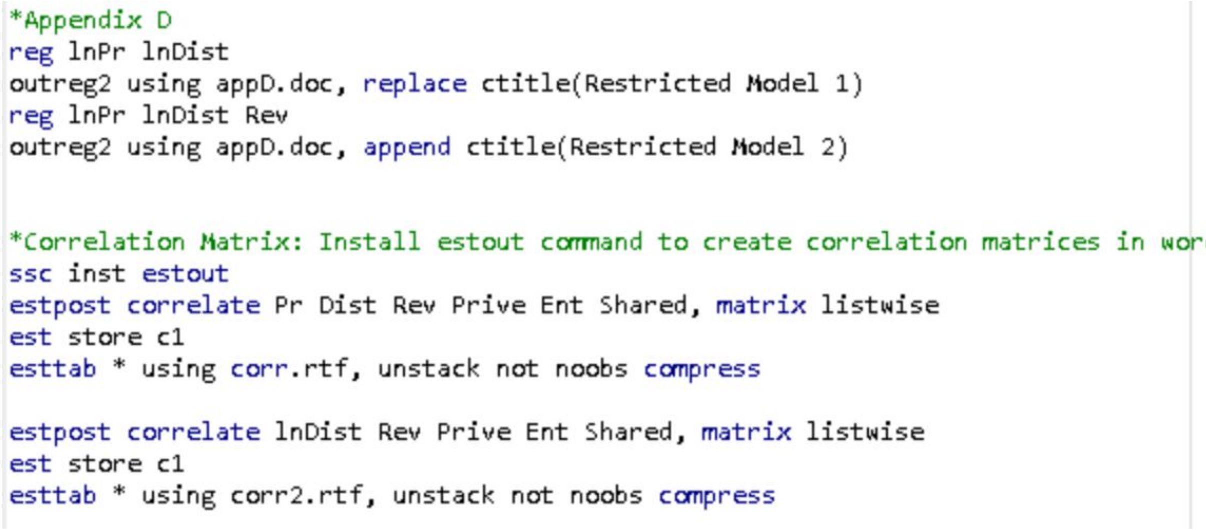
```
* An Airbnb rental price cannot be equal to zero, hence these observations are dropped.
tab Pr
list if Pr == 0
drop if Pr == 0

* Calculate the proportion of each type of room in the sample. (Included in the variable
description of the paper.)
tab room_type

* Get descriptive statistics for price, distance, number of reviews and dummy variables
for room type
sum Pr Dist Rev Ent Prive Shared

* Scatter plot between rental price and distance
Scatter Pr Dist

* Scatterplot between rental price and number of reviews
scatter Pr Rev

* Create the log variable for price, distance and number of reviews.
gen lnPr=ln(Pr)
gen lnDist=ln(Dist)
gen lnRev=ln(Rev)

* Correlation and covariance between price, distance, number of reviews and dummy
variables for room type
cor Pr Dist Rev Prive Ent Shared
cor Pr Dist Rev Prive Ent Shared, cov

* Check the proportion of properties in the sample with zero number of reviews
tab Rev
```



```
* Different functional forms
regress Pr Dist Rev Ent Prive
regress Pr lnDist lnRev Ent Prive
regress Pr Dist lnRev Ent Prive
regress Pr lnDist Rev Ent Prive
regress lnPr Dist Rev Ent Prive
regress lnPr lnDist lnRev Ent Prive
regress lnPr Dist lnRev Ent Prive
regress lnPr lnDist Rev Ent Prive

* Correlation and covariance matrix between the dependent variable and independent
variables.
cor lnPr lnDist Rev Prive Ent Shared
cor lnPr lnDist Rev Prive Ent Shared, cov

* Correlation among independent variables to check for multicollinearity
cor lnDist Rev Prive Ent Shared


* Auxillary regression of each independent variable on the other independent variable
regress lnDist Rev Ent Prive
regress Rev lnDist Ent Prive
regress Ent lnDist Rev Prive
regress Prive lnDist Rev Ent

* Breusch-Pagan test
regress lnPr lnDist Rev Prive Ent
predict uhat, resid
gen uhat2 = uhat^2
regress uhat2 lnDist Rev Prive Ent

* Testing to add variables
regress lnPr lnDist
regress lnPr lnDist Rev
regress lnPr lnDist Rev Ent Prive
```



```
* Chow test
**Restricted Model
regress lnPr lnDist Rev
**Unrestricted model
regress lnPr lnDist Rev if Ent==1
regress lnPr lnDist Rev if Ent==0

* 3.4.2.
** Estimate the model equation with robust standard errors
regress lnPr lnDist Rev Ent Prive, robust

*Tables: Install outreg2 command to create regression tables in word.
ssc install outreg2
reg uhat2 lnDist Rev Prive Ent
outreg2 using myreg.doc, replace ctitle(Breusch-Pagan)

reg lnPr lnDist Rev Ent Prive
outreg2 using myreg2.doc, replace ctitle(Model 1)
reg lnPr lnDist Rev Ent Prive, robist
outreg2 using myreg2.doc, append ctitle(Model 1 - Robust)

*Appendix A
reg Pr Dist
outreg2 using myreg3.doc, replace ctitle(Level-Level)
reg Pr Dist Rev
outreg2 using myreg3.doc, append ctitle(Level-Level-Level)

*Appendix C
reg lnDist Rev Ent Prive
outreg2 using aux1.doc, replace ctitle(Auxillary Model 1)
reg Rev lnDist Ent Prive
outreg2 using aux2.doc, replace ctitle(Auxillary Model 2)
reg Ent lnDist Rev Prive
outreg2 using aux3.doc, replace ctitle(Auxillary Model 3)
reg Prive lnDist Rev Ent
outreg2 using aux4.doc, replace ctitle(Auxillary Model 4)
```



```
*Appendix D
reg lnPr lnDist
outreg2 using appD.doc, replace ctitle(Restricted Model 1)
reg lnPr lnDist Rev
outreg2 using appD.doc, append ctitle(Restricted Model 2)


*Correlation Matrix: Install estout command to create correlation matrices in word.
ssc inst estout
estpost correlate Pr Dist Rev Prive Ent Shared, matrix listwise
est store c1
esttab * using corr.rtf, unstack not noobs compress

estpost correlate lnDist Rev Prive Ent Shared, matrix listwise
est store c1
esttab * using corr2.rtf, unstack not noobs compress
```

REFERENCES

Alsudais, A., & Teubner, T. (2019). Large-scale sentiment analysis on airbnb reviews from 15 cities. Retrieved from https://www.researchgate.net/publication/333114485_Large-Scale_Sentiment_Analysis_on_Airbnb_Reviews_from_15_Cities

Bosa, D. (2020). Airbnb raising another $1 billion in debt as coronavirus ravages tourism business. Retrieved from https://www.cnbc.com/2020/04/14/airbnb-raises-another-1-billion-in-debt.html

Bridges, J., & Vásquez, C. (2018). If nearly all airbnb reviews are positive, does that make them meaningless? *Current Issues in Tourism, 21*(18), 2065-2083. doi:10.1080/13683500.2016.1267113

Campbell, G. (2018). How to avoid the dreaded 4-star review: A guide for AirBnB hosts. Retrieved from https://medium.com/@campbellandia/how-to-avoid-the-dreaded-4-star-review-a-guide-for-airbnb-hosts-cdf482d083fe

Chaffin, J. (2020). New York's poorest borough flares as it suffers highest death rate. Retrieved from https://www.ft.com/content/a8b086f1-00b5-4965-9c98-2381ea48fda7

Chengjie Hea , Zhen Wanga , Huaicheng Guoa*, Hu Shenga , Rui Zhoub , Yonghui Yanga, (2010). *Driving Forces Analysis for Residential Housing Price in Beijing.* Procedia Environmental Sciences 2, 925–936.

Dayter, D., & Rudiger. (2013). Speak your mind, but watch your mouth: Complaints in CouchSurfing references. Retrieved from http://data.europeana.eu/item/2048441/item_CHBN7G3M3HCGH7EH5HW7SIGHZWJ62JID

Dogru, T., & Pekin, O. (2017). *What do guests value most in airbnb accommodations? an application of the hedonic pricing approach* Boston Hospitality Review.

Edelman, B. G., & Luca, M. (2014). Digital discrimination: The case of airbnb.com. *SSRN Electronic Journal,* doi:10.2139/ssrn.2377353

Zervas, G., Proserpio, D., Byers, J. (2017). The rise of the sharing economy: Estimating the impact of airbnb on the hotel industry. *Journal of Marketing Research, 54*(5), 687-705. doi:10.1509/jmr.15.0204

Gibbs, C., Guttentag, D., Gretzel, U., Morton, J., & Goodwill, A. (2018). Pricing in the sharing economy: A hedonic pricing model applied to airbnb listings. *Journal of Travel & Tourism Marketing, 35*(1), 46-56. doi:10.1080/10548408.2017.1308292

Grinevich, V., Grinevich, V., Huber, F., Huber, F., Karataş-Özkan, M., Karataş-Özkan, M., Yavuz, Ç. (2019). Green entrepreneurship in the sharing economy: Utilising multiplicity of institutional logics. *Small Business Economics, 52*(4), 859-876. doi:10.1007/s11187-017-9935-x

Han, H. J., Mankad, S., Gavirneni, N., & Verma, R. (2016). *What guests really think of your hotel: Text analytics of online customer reviews* Unpublished. doi:10.13140/rg.2.1.3963.6244

iPropertyManagement. (2017). Airbnb statistics. Retrieved from https://ipropertymanagement.com/research/airbnb-statistics

iPropertyManagement. (2019). Airbnb statistics. Retrieved from https://ipropertymanagement.com/research/airbnb-statistics

Johnson, M., & Ghiglione, D. (2020). *Rome's airbnb landlords suffer after tourism collapse*. London: The Financial Times Limited.

Much Needed. (2020). Airbnb by the numbers: Usage, demographics, and revenue growth. Retrieved from https://muchneeded.com/airbnb-statistics/

New York Tourism Report. (2019). *NYC travel and tourism trend report*. Retrieved from https://adobeindd.com/view/publications/e91e777a-c68b-4db1-a609-58664a52cffd/7r7x/publication-web-resources/pdf/NYC_Travel&TourismTrendReport_Oct2019.pdf

Oxford Economics. (2019). *Oxford Economics Landmark Report.* Retrieved from https://news.airbnb.com/en-au/airbnb-can-play-a-critical-role-in-tourism-recovery-oxford-economics/

Pankaj Bhurtel (2015). *An econometric analysis of housing market in Stockton, CA.* The University of Queensland.

Ranjbari, M., Morales-Alonso, G., & Carrasco-Gallego, R. (2018). Conceptualizing the sharing economy through presenting a comprehensive framework. *Sustainability (Basel, Switzerland), 10*(7), 2336. doi:10.3390/su10072336

Staten Island Advance. (2019). Parks: Staten island is the greenest new york city borough. Retrieved from https://www.silive.com/guide/2010/04/parks_staten_island_is_the_greenest_borough.html

Studenmund, A. H. (2016). Using econometrics: A practical guide. Pearson.

Tun, Z. T. (2020). Top cities where airbnb is legal or illegal . Retrieved from https://www.investopedia.com/articles/investing/083115/top-cities-where-airbnb-legal-or-illegal.asp

Weber, A. (2013). Queens. Retrieved from https://web.archive.org/web/20150513065643/http://www.newyork.com/articles/neighborhoods/queens-72876/

What do the different home types mean? - Airbnb Help Center. Retrieved from https://www.airbnb.com/help/article/317/what-do-the-different-home-types-mean

Zervas, G., Proserpio, D., & Byers, J.A first look at online reputation on airbnb, where every stay is above average. *SSRN Electronic Journal,* doi:10.2139/ssrn.2554500