Distance from the Colosseum and Prices of Airbnb Listings in Rome, Italy

Ryan Mulcahey

4/28/2022 | Washington University in St. Louis | Introduction to Econometrics

Abstract

Soaring housing prices in Rome have led to a housing crisis in the historic city. Finding a relationship between the geographical distance from the Colosseum and pricing per night of an Airbnb listing would add an additional information layer to conversations about Rome's housing problems, in addition to providing valuable information to tourists looking for a place to stay. I found data from Inside Airbnb of 24,747 listings of Airbnb rentals in Rome, compiled on December 8, 2021. Using the longitude and latitude of the listings and the latitude=41.8902° E W and the longitude=12.4922° N of the Colosseum, I created a linear regression model of the price per night of a listing. Based on similar research and economic theory, I chose a semilog (lnY) functional form for the equation. The results show a statistically significant negative effect of the distance from the Colosseum on the price per night of an Airbnb listing in Rome. While severe multicollinearity and autocorrelation do not seem to be issues for the equation, heteroskedasticity and omitted variable bias are issues to be aware of.

Motivation and Introduction

Tourists and advocates for affordable housing in Rome have an interest in seeing what factors influence housing prices. Location is a major factor in determining housing prices. Unlike other major cities, with “nicer”, more expensive neighborhoods scattered throughout, Rome is very centralized. Famous tourist attractions, such as the Pantheon, Colosseum, Roman Forum, and the upscale Monti Neighborhood lie at the center of the city. So, how does the distance from the Colosseum of an Airbnb listing in Rome affect the price per night of the listing? In this project, I seek to answer this question, as well as determine what other variables affect the price per night of an Airbnb listing in Rome, and to what extent.

Chica-Olmo et al.’s study on the effects of location on airbnb apartment pricing in Málaga revealed that the distance from the center of the city, distance from the beach, and distance from the nearest place of interest have a statistically significant negative effect on Airbnb apartment pricing [1]. Zhang et al.’s study of key factors affecting the price of Airbnb listings in Metro Nashville, Tennessee similarly concluded that the distance from the nearest convention center has a statistically significant negative effect on the Airbnb listing price. Note that most of Nashville’s convention centers are in the downtown area [6]. From a theoretical standpoint, this indicates that as the distance of an Airbnb listing in Rome from the Colosseum increases, the price per night will decrease, holding all else constant.

Description of the Model

Theoretical Model:

$$\text{lnprice} = f(\overset{-}{\text{DISTclsm}}, \overset{-}{\text{DISTstptr}}, \overset{-}{\text{MINNIGHTS}}, \overset{+}{\text{PERCENT}}, \overset{+}{\text{REVIEWSRATE}},$$

$$\overset{+}{\text{HOSTLISTINGS}}, \overset{-}{\text{AVAILABILITY}}, \overset{+}{\text{ENTIRE}})$$

Discussion of the Variables:

- lnprice is the natural logarithm of the predicted price per night in euros of the listing.

1. DISTclsm is the distance in decimal degrees (geographical coordinate system unit of measurement) from the Colosseum as calculated from the latitude and longitude of the listing. As distance increases from the Colosseum, which is in the center of the city and near other historic sites, I hypothesize the price should decrease.

   a. To avoid multicollinearity I did not include distance variables for the Monti Neighborhood, the Pantheon, or other famous tourist sites in the center of Rome, as they are located very close to the Colosseum.
2. DISTstptr is the distance in decimal degrees from Saint Peter's Basilica as calculated from the latitude and longitude of the listing. As distance increases from Saint Peter's Basilica, which is in Vatican City and near other religious sites and tourist attractions, I hypothesize the price should decrease.
3. MINNIGHTS is the minimum number of nights the listings must be rented. As you must stay for more nights (increasing minimum number), the price per night should decrease.
4. PERCENT is the proportion of the reviews for the listing on Airbnb that were from the last month. A more in demand (and higher priced rental), likely should have had more of its reviews more recently.
5. REVIEWSRATE is the number of reviews per month for the listing on Airbnb. A more in demand (and higher priced rental), likely should have more reviews per month.
6. HOSTLISTINGS is the number of listings the host has on Airbnb. As a host has more properties listed, they would have more experience renting properties and ensuring a high quality of stay, so the quality of the listing would increase and thus the price.
7. AVAILABILITY is the number of days out of 365 that the listing is available on Airbnb. Fewer days available in the year, means the host is more selective about when they rent, which likely means that they only rent on expensive rental days like school vacation weeks, so one could theorize that price per night goes down as availability goes up.
8. ENTIRE is 1 if the rental is for an entire home/apt, 0 if not. This means that Entire is 0, if the listing is shared, private room or a hotel room. A whole house/apt is much more valuable than a single/shared/hotel room.
   a. To avoid multicollinearity, I've eliminated two additional dummy variables for shared and private room, as suggested when I asked in class.

Data Description and Model Estimation

The data is from Inside Airbnb, which is unfortunately the only source I could find for free Airbnb listings data. It has 24,737 observations of listings. Inside Airbnb is a website, which provides scrapped Airbnb, free of charge.

Irregular listings were dropped from the dataset. In the context of this project, a listing was deemed irregular if the price was greater than $500 per night. Or, if a listing received more than twenty reviews in the previous month, as this could indicate a host manipulating their listing with fake reviews. Or, if a listing was available for less than seven days in a calendar year, as this could indicate a listing only being available during especially expensive holiday weekends. After also dropping listings with missing data, data on 11,830 listings remained.

The mean distance from the Colosseum of these listings is 0.039 decimal degrees, or roughly 2.7 miles. The mean distance from Saint Peter's Basilica of these listings is 0.046 decimal degrees, or roughly 3.2 miles. The listings' prices range from 9 euros to 500 euros, the natural logarithm of each of these prices are 2.2 and 6.2, respectively. The mean natural logarithm of listing price is 4.3. On average, listings had a minimum night requirement of two nights and one review per month. Hosts of the listings had between 0 and 174 other listings on Airbnb. On average, the listings had 23% of their reviews in the last month.

Variable	Obs	Mean	Std. Dev.	Min	Max
lnprice	11,830	4.327806	.5537883	2.197225	6.214608
price*	11,830	81.51014	61.94386	9	500
DISTclsm	11,830	.0392198	.0417899	.0006563	.297821
DISTstptr	11,830	.0459986	.0423393	.001648	.3353328
MINNIGHTS	11,830	2.096619	1.119557	1	9
PERCENT	11,830	.2296168	.3329023	0	1
REVIEWSRATE	11,830	1.13124	1.046873	.03	7.42
HOSTLISTINGS	11,830	6.591462	16.89497	1	175
AVAILABILITY	11,830	241.2455	107.4516	7	365
ENTIRE	11,830	.708284	.4545715	0	1

Fig. 1. Summary statistics of the dependent and explanatory variables, in addition to price. Note that price is included in this summary statistics table, but is not part of the equation, as it is lnprice that is the dependent variable. The inclusion of price is simply to provide context to the summary statistics of lnprice.

Empirical Model:

$$\text{lnprice}_i = \beta_0 + \beta_1\text{DISTclsm}_i + \beta_2\text{DISTstptr}_i + \beta_3\text{MINNIGHTS}_i + \beta_4\text{PERCENT}_i + \beta_5\text{REVIEWSRATE}_i + \beta_6\text{HOSTLISTINGS}_i + \beta_7\text{AVAILABILITY}_i + \beta_8\text{ENTIRE}_i + \varepsilon_i$$

Functional Form Explanation:

The quality of fit of an estimated equation is not an all-encompassing measure of quality of a regression, so I chose to base the functional form choices based on theory. Adjusted R-squared can be used to compare the fits of equations with the same dependent variable, which is the case for all equations whose dependent variable is lnprice. However, they can not be used to compare the fits of equations with different dependent variables, which is the case for an equation with lnprice as the dependent variable and an equation with price as the dependent variable.

To avoid multicollinearity between NUMREVIEWS, the number of reviews for the listing on Airbnb, and NUMREVIEWSLTM, the number of reviews for the listing last month, I created the variable PERCENT, which is NUMREVIEWS/NUMREVIEWSltm, instead of including NUMREVIEWS and NUMREVIEWSltm individually in the equation.

In their spatial autoregressive model, Chica-Olmo et al. chose to set the dependent variable as the natural logarithm of apartment price, mentioning that this functional form is frequently chosen in Airbnb pricing models [1]. I similarly chose a semilog (lnY) functional form for the equation, setting the dependent variable to be the natural logarithm of the price per night, as I hypothesize that one unit increases in each of the explanatory variables, independently, will have an effect on the price of the listing in percentage terms. This means that the functional form of the dependent variable is logarithmic, while all explanatory variables are linear. I did not include any interaction terms, as I do not hypothesize that the change in price with respect to any of the independent variables depends on the value of different independent variable(s).

Preliminary Results

Initial Regression Results:

Source	SS	df	MS		Number of obs	=	11,830
Model	996.73325	8	124.591656		**F(8, 11821)**	**=**	**559.79**
Residual	2631.00162	11,821	.22257014		**Prob > F**	**=**	**0.0000**
Total	3627.73487	11,829	.30668145		R-squared	=	0.2748
					Adj R-squared	**=**	**0.2743**
					Root MSE	=	.47177

lnprice	**Coef.**	**Std. Err.**	**t**	P>|t|	[95% Conf.	Interval]
DISTclsm	-.72915	.1711696	-4.26	0.000	-1.064671	-.3936294
DISTstptr	-2.778335	.167804	-16.56	0.000	-3.107259	-2.449412
MINNIGHTS	-.0249785	.0040721	-6.13	0.000	-.0329604	-.0169965
PERCENT	-.0555414	.0135492	-4.10	0.000	-.0821	-.0289828
REVIEWSRATE	-.0772332	.0043473	-17.77	0.000	-.0857547	-.0687118
HOSTLISTINGS	.0018057	.0002616	6.90	0.000	.001293	.0023185
AVAILABILITY	.0004641	.0000407	11.40	0.000	.0003843	.000544
ENTIRE	.5300257	.0101345	52.30	0.000	.5101604	.549891
_cons	4.137415	.0170354	242.87	0.000	4.104023	4.170807

Fig. 2. Results of the initial regression, including parameter estimates, standard errors, and t-statistics of all independent variables, in addition to the overall F-statistic and adjusted R-squared.

The F-statistic of overall significance provides a formal hypothesis test of overfit. H_0: $\beta_1=\beta_2=\beta_3=\beta_4=\beta_5=\beta_6=\beta_7=\beta_8=0$, H_A: H_0 is not true. With an F-statistic equal to 559.79, and a p-value equal to 0.000, we can reject the null hypothesis at the 0.05 level.

The Adjusted R-squared of 0.2743, means that 27.43% of the variation of the natural logarithm of price around its mean can be explained by the explanatory variables, adjusted for degrees of freedom.

See the hypothesis tests of the explanatory variables below. All coefficients of the explanatory are statistically significant, however, some the coefficients of PERCENT, REVIEWSRATE, and AVAILABILITY do not have their hypothesized sign. Note that a one-tailed test at the 0.05 level, with eight explanatory variables and 11,830 observations has a t-critical value of 1.65. The absolute values of the t-statistics of the eight explanatory variables all far exceed this critical value, likely due to the low variances caused by the large sample size.

Hypothesis Tests:

1. DISTclsm, H_0: $\beta_1>=0$, H_A: $\beta_1<0$, p-value<0.05
   a. Since $|t|> t_c$ and has the hypothesized sign, we reject H_0.
2. DISTstptr, H_0: $\beta_1>=0$, H_A: $\beta_2<0$, p-value<0.05
   a. Since $|t|> t_c$ and has the hypothesized sign, we reject H_0.
3. MINNIGHTS, H_0: $\beta_2>=0$, H_A: $\beta_3<0$, p-value<0.05
   a. Since $|t|> t_c$ and has the hypothesized t sign, we reject H_0.

4. PERCENT, H_0: β_3<=0, H_A: β_4>0, p-value<0.05
    a. The coefficient does not have the hypothesized sign, so we can not reject H_0.
5. REVIEWSRATE, H_0: β_4<=0, H_A: β_5>0, p-value<0.05
    a. The coefficient does not have the hypothesized sign, so we can not reject H_0.
6. HOSTLISTINGS, H_0: β_3<=0, H_A: β_6>0, p-value<0.05
    a. Since |t|> t_c and has the hypothesized sign, we reject H_0.
7. AVAILABILITY, H_0: β_4>=0, H_A: β_7<0, p-value<0.05
    a. The coefficient does not have the hypothesized sign, so we can not reject H_0.
8. ENTIRE H_0: β_3<=0, H_A: β_8>0, p-value<0.05
    a. Since |t|> t_c and has the hypothesized sign, we reject H_0.

Tests for Econometric Issues

Multicollinearity:

I do not theorize any severe multicollinearity in the equation. However, multicollinearity exists to some degree in any equation. Therefore, I created a correlation matrix to display the correlation coefficients of the dependent variable, lnprice, and the eight independent variables. The only notably high correlation coefficient between two independent variables is that of DISTclsm and DISTstptr, 0.787. This is understandable as the Colosseum and Saint Peter's Basilica are located relatively close to each other, however they are still over 2 miles apart. However, the variances of the estimated coefficients do not seem to be "too high" as a result of this relatively high correlation coefficient. See Appendix Fig. 1 for further information on the correlation coefficients of the variables.

I also used variance inflation factors (VIFs) to detect the severity of multicollinearity in the equation. Despite there not being critical values for a VIF, a common rule of thumb is to note that multicollinearity may be severe if VIF > 5. However, the VIFs of all eight explanatory variables and the Mean VIF are less than 5, as. shown in Appendix Fig. 2. The highest VIFs are 2.72 and 2.68 for DISTclsm and DISTstptr, respectively. While it is possible to have severe multicollinearity without a large VIF, the relatively low VIFs of the explanatory variables and lack of theory supporting there being severe multicollinearity in the equation leads me to believe that severe multicollinearity is not an issue in the equation.

Heteroskedasticity:

From a theoretical standpoint, I believe that the equation suffers from heteroskedasticity, as the data set used has a large range of observed values for the dependent variable, lnprice, which is often associated with heteroskedasticity. Additionally, there appears to be heteroskedasticity in the equation, as shown in Appendix Fig. 3, a plot of the residuals in relation to lnprice. In summary, the plot visualizes how the error term variance is not constant and depends on the observation, as the residuals appear to increase as lnprice increases. Thus, I chose to perform the Breusch-Pagan and White tests to test for heteroskedasticity.

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
Ho: Constant variance
Variables: fitted values of lnprice

chi2(1) = 11.87
Prob > chi2 = 0.0006

The chi-square test-statistic is 11.87 with a p-value of 0.0006. Therefore at the 5-percent level, we reject the null hypothesis. Heteroskedasticity appears to be present, due to the results of the Breush-Pagan test.

White's test for Ho: homoskedasticity
against Ha: unrestricted heteroskedasticity

chi2(43) = 211.63
Prob > chi2 = 0.0000

The chi-square test-statistic is 211.63 with a p-value of 0.0006. There were 43 variables in the auxiliary regression and therefore 43 degrees of freedom for the White test. Therefore at the 5-percent level, we reject the null hypothesis and once again conclude that heteroskedasticity is present. Therefore, I will use heteroskedasticity-corrected standard errors to re-estimate the model.

Autocorrelation/Serial Correlation:

The data I used for the equation only includes one observation per listing, not multiple observations of a single listing over time. As this is cross-sectional data, not time series data, I am not concerned about serial correlation in the equation.

Final Estimated Model

$$\begin{array}{lllll} \text{lnprice}_i = 4.137 & - 0.729\text{DISTclsm}_i & - 2.778\text{DISTstptr}_i & - 0.025\text{MINNIGHTS}_i & - 0.056\text{PERCENT}_i \\ & (0.176) & (0.168) & (0.004) & (0.013) \\ t = & -4.14 & -16.58 & -5.76 & -4.35 \end{array}$$

$$\begin{array}{lcccc} & - 0.077REVIEWSRATE_i & + 0.002HOSTLISTINGS_i & + 0.0005AVAILABILITY_i & + 0.5300ENTIRE_i \\ & (0.004) & (0.0002) & (0.00004) & (0.010) \\ t = & -19.64 & 7.37 & 11.43 & 53.69 \end{array}$$

$$N = 11{,}830 \qquad \text{R-squared} = 0.2748$$

I found heteroskedasticity to be an econometric issue for my initial estimated equation. After examining the equation carefully for specification errors, I could not find any obvious specification errors, so I chose to use heteroskedasticity-correct (HC) standard errors to re-estimate the equation. These standard errors are calculated specifically to avoid the consequences of heteroskedasticity. Although these standard errors are biased, I used a large sample, which lends to heteroskedasticity-corrected standard errors to generally be more accurate than the incorrect standard errors I used in the initial estimated model. OLS estimates of standard errors are biased if heteroskedasticity is present, which causes the hypothesis testing and confidence intervals of the initial equation to be unreliable. The hypothesis tests and confidence intervals of this re-estimated model will be more reliable, as it uses heteroskedasticity-corrected standard errors. Note that the coefficients are the same for this equation and the initial equation. Simply the standard errors, and thus the t-statistics, p-values, and confidence intervals of each variable change, in addition to the overall F-statistic and its corresponding p-value. There were not any significant changes in any of these values using heteroskedasticity-correct standard errors, instead of uncorrected standard errors.

Number of obs	=	11,830
F(8, 11821)	**=**	**577.51**
Prob > F	**=**	**0.0000**
R-squared	=	0.2748
Root MSE	=	.47177

lnprice	**Coef.**	**Std. Err.**	**t**	P>|t|	[95% Conf.	Interval]
DISTclsm	-.72915	.1762485	-4.14	0.000	-1.074626	-.383674
DISTstptr	-2.778335	.1675258	-16.58	0.000	-3.106713	-2.449957
MINNIGHTS	-.0249785	.0043374	-5.76	0.000	-.0334805	-.0164764
PERCENT	-.0555414	.0127586	-4.35	0.000	-.0805503	-.0305324

REVIEWSRATE	-.0772332	.0039318	-19.64	0.000	-.0849402	-.0695263
HOSTLISTINGS	.0018057	.000245	7.39	0.000	.0013256	.0022859
AVAILABILITY	.0004641	.0000406	11.43	0.000	.0003845	.0005437
ENTIRE	.5300257	.0098712	53.69	0.000	.5106765	.5493749
_cons	4.137415	.0179031	231.10	0.000	4.102322	4.172508

Fig. 3. Results of the final regression, including parameter estimates, standard errors, and t-statistics of all independent variables, in addition to the overall F-statistic and adjusted R-squared.

Interpretation and Hypothesis Tests:

1. DISTclsm, H_0: $\beta_1>=0$, H_A: $\beta_1<0$, p-value<0.05
   a. Since $|t|> t_c$ and has the hypothesized sign, we reject H_0.
   b. The coefficient of -0.729 implies that if the distance to the Colosseum of an Airbnb listing in Rome increases by one decimal degree, the price per night of the listing will decrease by roughly 72.9%, holding all else constant.
2. DISTstptr, H_0: $\beta_1>=0$, H_A: $\beta_2<0$, p-value<0.05
   a. Since $|t|> t_c$ and has the hypothesized sign, we reject H_0.
   b. The coefficient of -2.778 implies that if the distance to Saint Peter's Basilica of an Airbnb listing in Rome increases by one decimal degree, the price per night of the listing will decrease by roughly 277.8%, holding all else constant.
3. MINNIGHTS, H_0: $\beta_2>=0$, H_A: $\beta_3<0$, p-value<0.05
   a. Since $|t|> t_c$ and has the hypothesized t sign, we reject H_0.
   b. The coefficient of -0.025 implies that if the minimum number of nights required to rent an Airbnb listing in Rome increases by one night, the price per night of the listing will decrease by roughly 2.5%, holding all else constant.
4. PERCENT, H_0: $\beta_3<=0$, H_A: $\beta_4>0$, p-value<0.05
   a. The coefficient does not have the hypothesized sign, so we can not reject H_0.
   b. The coefficient of -0.056 implies that if the proportion of the reviews that were from the last month of an Airbnb listing in Rome increases by one unit, the price per night of the listing will decrease by roughly 5.6%, holding all else constant.
5. REVIEWSRATE, H_0: $\beta_4<=0$, H_A: $\beta_5>0$, p-value<0.05
   a. The coefficient does not have the hypothesized sign, so we can not reject H_0.
   b. The coefficient of -0.077 implies that if the number of reviews per month of an Airbnb listing in Rome increases by one review, the price per night of the listing will decrease by roughly 7.7%, holding all else constant.
6. HOSTLISTINGS, H_0: $\beta_3<=0$, H_A: $\beta_6>0$, p-value<0.05

    a. Since $|t|> t_c$ and has the hypothesized sign, we reject H_0.
    b. The coefficient of 0.002 implies that if the number of total listings of a host of an Airbnb listing in Rome increases by one listing, the price per night of the listing will increase by roughly 0.2%, holding all else constant.

7. AVAILABILITY, H_0: β_4>=0, H_A: β_7<0, p-value<0.05
    a. The coefficient does not have the hypothesized sign, so we can not reject H_0.
    b. The coefficient of 0.0005 implies that if the number of days available in 365 days of an Airbnb listing in Rome increases by one day the price per night of the listing will increase by roughly 0.05%, holding all else constant.
8. ENTIRE H_0: β_3<=0, H_A: β_8>0, p-value<0.05
    a. Since $|t|> t_c$ and has the hypothesized sign, we reject H_0.
    b. The coefficient of .5300 implies that the price per night of the listing is roughly 53% higher if the listing is for an entire home or apt, than if not, holding all else constant.

Omitted Variable Bias:

It seems that a lot of variance is not captured by the explanatory variables, but this is a property of the data/topic itself. There is theoretical concern for omitted variable bias, as there's a lack of explanatory variables that would theoretically have a significant impact on the price per night of a listing, such as quality of listing (renovated/new or old property/building, amenities available, etc) and distance from the metro. Data for these explanatory variables could not be collected for this project, as it is not included in this data set and other datasets on Airbnb listings are not free. To test this theory of omitted variable bias, I ran the Ramsey RESET test.

```
Ramsey RESET test using powers of the fitted values of lnprice
     Ho:  model has no omitted variables
          F(3, 11818) =     17.84
            Prob > F =      0.0000
```

Running the Ramsey Reset test on the final estimated models, led to an F-statistic of 17.84 and a p-value of 0.0000. Therefore, we reject the null hypothesis of the Ramsey RESET test at the 0.05 level, meaning we have some evidence that the coefficient estimates on the additional parameters are statistically different from zero. This is evidence of a potential specification error, potentially from an omitted relevant explanatory variable(s) or an incorrect functional form. While it is certainly possible that there is an error in the functional form, I am led to conclude that there are certainly relevant omitted variables, due to the hypothesized variables above, quality of listings and distance from the metro. It is also quite possible that other relevant omitted variables exist. Unfortunately, I did not have access to data to include these in this project.

Conclusions

Understanding the effect of distance from the Colosseum and other factors on Aribnb housing prices in Rome is useful information for tourists looking for a place to stay and for those looking to ameliorate Rome's housing problems. Consulting similar studies into the effect of distance from prominent places on Airbnb listing prices, I chose a semilog (lnY) functional form for my equation. Upon discovering heteroskedasticity in my equation, I re-estimated my equation with heteroskedasticity-corrected standard errors. My final model revealed that all eight explanatory variables were statistically significant at the 0.05 level, with DISTclsm, DISTstptr, MINNIGHTS, PERCENT, and REVIEWSRATE having a negative effect on the natural logarithm of price. Meanwhile, HOSTLISTINGS, AVAILABILITY, and ENTIRE had a negative effect on the natural logarithm of price. However, PERCENT, REVIEWSRATE, and AVAILABILITY did not have the hypothesized effect in the final estimated model.

Omitted variable bias seems to be a major issue for the final estimated model. This is a finding backed both by theory and the Ramsey RESET test. Future research on Airbnb pricing in Rome should attempt to find listing data on the quality of listing (renovated/new or old property/building, amenities available, etc) and distance from the metro, in addition to any other hypothesized relevant variables. While straying from econometrics and towards the field of machine learning, future research could look into applying the k-nearest neighbor algorithm to predict Airbnb listing prices in Rome, by predicting the price of listings based on the prices of the "k" listings most similar to it. Regardless of future methodology, the current model could be improved by accounting for omitted variables. This econometric issue should be the focus of future research.

Appendix

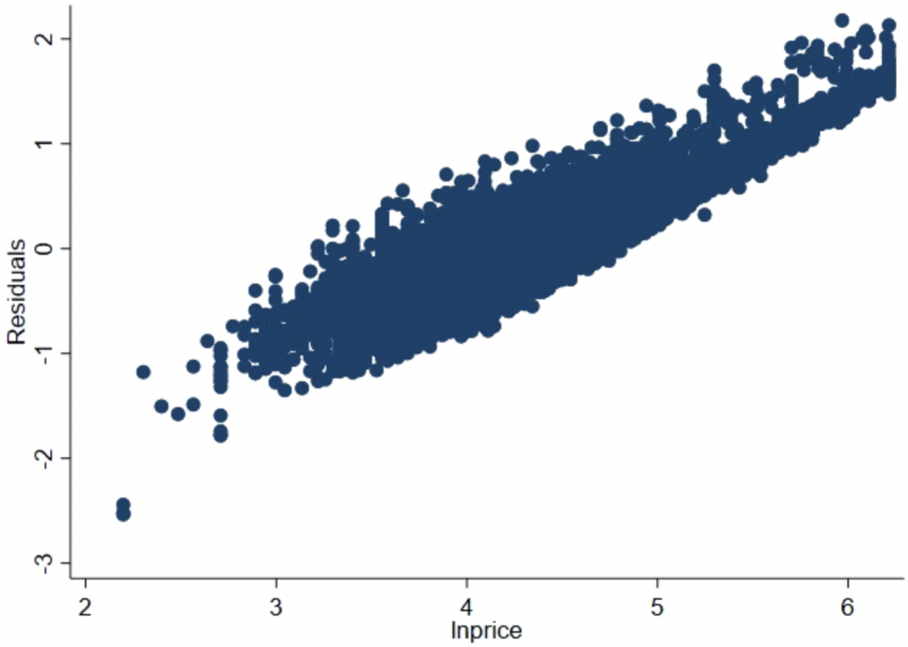
	lnprice	DISTclsm	DISTstptr	MINNIGHTS	PERCENT	REVIEWSRATE	HOSTLISTINGS	AVAILABLITY	ENTIRE
lnprice	1.000								
DISTclsm	**-0.205**	1.00							
DISTstptr	**-0.282**	**0.787**	1.00						

MINNIGHTS	0.062	0.028	0.009	1.00					
PERCENT	-0.039	-0.004	-0.003	-0.089	1.00				
REVIEWSRATE	-0.074	-0.162	-0.117	-0.034	0.221	1.00			
HOSTLISTINGS	0.098	-0.105	-0.082	-0.044	0.125	0.009	1.00		
AVAILABILITY	0.075	-0.019	-0.026	-0.096	0.047	-0.032	-0.038	1.00	
ENTIRE	**0.420**	0.005	-0.082	0.273	0.025	0.107	0.062	-0.064	1.00

Appendix Fig. 1. Correlation Matrix, simply showing correlation coefficients between all variables of the equation.

Variable	VIF	1/VIF
DISTclsm	2.72	0.367724
DISTstptr	2.68	0.372757
ENTIRE	1.13	0.886562
MINNIGHTS	1.10	0.905306
REVIEWSRATE	1.10	0.908429
PERCENT	1.08	0.924827
HOSTLISTINGS	1.04	0.963317
AVAILABILITY	1.02	0.982505
Mean VIF	**1.48**	

Appendix Fig. 2. Variance Inflation Factors (VIFs) and Tolerance table for all eight explanatory variables. The table also includes the Mean VIF.



Appendix Fig. 3. Plot of the residuals, the difference between predicted and actual value of the dependent variable, and lnprice, the dependent variable.

Bibliography

[1] Chica-Olmo, Jorge, et al. *Effects of Location on Airbnb Apartment Pricing in Málaga.* Tourism Management, Apr. 2020, https://www.sciencedirect.com/science/article/pii/S0261517719301797?casa_token=qd8OmBrlLykAAAAA%3ALX3zSKwwQHz6haf8wkM5f1qD6Q6Tt9NPyYXlE8OSDmUJWJm1OUC-h-eFbLAdKK0zk5BX1NpK.

[2] “Colosseum Latitude and Longitude.” *distancesto.com*, https://www.distancesto.com/coordinates/it/colosseum-latitude-longitude/history/80236.html.

[3] Gould, William. “Creating Dummy Variables.” *Stata*, https://www.stata.com/support/faqs/data-management/creating-dummy-variables/.

[4] “Inside Airbnb: Get the Data.” *Inside Airbnb*, http://insideairbnb.com/get-the-data/.

[5] Studenmund, Arnold H. *Using Econometrics: A Practical Guide*. 7th Edition, Pearson Addison Wesley, 2016.

[6] Zhang, Zhihua, et al. *Key Factors Affecting the Price of Airbnb Listings: A Geographically Weighted Approach.* Multidisciplinary Digital Publishing Institute, 14 Sept. 2017, https://www.mdpi.com/2071-1050/9/9/1635.