Assignment 1 Data Analysis 2

Anton Shestakov

November 2024

Executive summary

The main results of the analysis show a clear nonlinear relationship between price for one night in Munich hotels and longitude values. The statistics show a significant influence of the east-west location of the hotels on the price changing. Higher prices are explained by the behavior of a concave function, the top of which corresponds to the location in the center.

Introduction

(A)

This analysis focuses on determining the pattern between prices and longitude index of Munich hotels based on Airnbnb data for June 2024.

(B)

The following variables were used in the analysis:

- price ($) for one night
- longitude (°) shows how far east or west a hotel location is from 0° meridian
- room type (categorical variable)

The population of this analysis consists of landlords renting out premises in Munich.

Potential data quality issues might be connected with a great number of missing values, sample homogeneity (due to lack of randomization), or subjective distortion especially if questions are about the extent of satisfaction or another personal perception.

Data

(A)

The employed sample involved 3115 obs. Scaling modification was tailored to independent variable 'longitude'. To simplify the interpretation of the regression results, the variable was multiplied by 100 to obtain a change of 1 step instead of 0.01 step.

***Note:** descriptive analysis contains longtitude variable with original scale. The modification was implemented right before model building.

Some categories of variable 'room type' were dropped due to a small number of observations which create a risk of obtaining non-robust results. Eventually, only 'Entire home/apt' and 'Private room' were kept for comparison analysis of models.

(B)

Looking through descriptive table of our X (longitude) and Y (price) continuous variables, we can infer that the former is delimited by quite narrow range between 11.39 and 11.71. However, it's hard to find any abnormality.
The latter variable (price) has clear outliers that can effect our results. We see that max value is 11000 while mean value is around 172.1. In turn, it fairly explains high s.d value associated to obs. which are overjumping out of the average.

Table 1: Descriptive Statistics

Variable	**Count**	**Mean**	**SD**	**Min**	**25%**	**50%**	**75%**	**Max**
Price ($)	3115.0	172.1	252.5	10.0	79.0	120.0	195.0	11000.0
longitude (°)	3115.0	11.56	0.049	11.39	11.53	11.56	11.58	11.71

(C)

In the course of the analysis it was decided to modify the dependent variable price by logarithmic transformation. The original price distribution is right-skewed (pic. 1. Appendix) because of outliers and uneven values placement. Logarithmic transformation compressed our data making it more meaningful (pic. 2. Appendix). Moreover, it partially levels out the problem of heteroskedasticity. Logarithmic transformation creates more stable variance across the range making data more homoskedastistically.

Another transformation was made with longitude variable. In regression models squared values were used because of non-linear behavior with respect to dependent variable price (pic. 3. Appendix).

Model and evaluation

(A)

To justify the choice of the best model it is necessary to get back to the previous reasoning. First and foremost, it is meaningless to employ 'price' data w/o logarithmic transformation due to uneven distribution and outliers of original massive.
Furthermore, model's independent variable 'longitude' has likely non-linear effect according to the built plot (pic. 3. Appendix). It can be seen that values create the form of the concave function with peak point between 11.5 and 11.6 interval.
In fact, if regression model includes only 'longitude' var w/o any transformation, it has equivocal results with a low significance (Table 1. Appendix). Evidently, this should prompt the idea of using the quadratic effect.

(B)

Final version of the model:

$$\log(price) = \beta_0 + \beta_1 \cdot longitude_i + \beta_2 \cdot {longitude_i}^2 + \varepsilon_i,$$

with I. I. D. assumptions

Estimated parameters are shown in Table 2 (Appendix). As it's been said before, 'longitude' variable was multiplied by 100 to simplify interpretation. Totally, the marginal effect on 'price' can be inferred as:

$$\text{M.E} = \beta_1 + 2 \cdot \beta_2 \cdot longitude_i$$

To interpret the parameters it should be also taken into account that model type is log-linear that means: Y is 100 * beta coefficient (%) on average for observation with a unit higher X. In other words, on average price for hotels in Munich is changed by 564(%) - 0.24(%)*longitude(i).
Obviously, these results cannot be regarded as valid because the model doesn't take many other factors. However, the pattern of dependence is enough clear.

Another important step in the analysis is to employ the model for comparative analysis based on the room-type variable. Estimated results (Table 3. Appendix) demonstrates analogical patterns in the context of premise types.

(C)

$p < 0.001$ significance level is used to test hypothesis of zero betas effect. Z-score interpretation in our model is a bit nuanced because the model includes a quadratic effect of independent variable. Nevertheless, the significance is observed in both cases. Null hypothesis of zero linear effect between price and longitude is rejected on 99.9% confidence level with critical value of 3.291 as well as the null hypothesis of zero non-linear effect is also rejected on 99.9% c.l.

Even though the data was modified to dodge a high variation from the mean among observations, the model employed robust standard errors. Model's assumption is about non-constant variance which leads to heteroscedasticity.

(D)

Fitted model values are predominantly located in area of 4.8 and 4.92 (Picture 4. Appendix). It is hard to interpret the max and min residuals due to symmetric distribution with respect to fitted values. However, paying attention on Picture 5 (Appendix) it becomes clear that these values fit to outliers. Obviously, the built model doesn't include many other factors that can explain price on hotels. Nonetheless, the largest positive residuals (black point on the plot) are concentrated close to 4.9 area and testify that built model underestimates the effects of the 'invisible' factors that make prices for these observations more valuable.

Analogically, for the largest negative errors (green point on the plot): they're concentrated at the same area and testify about model's overestimation the effects of certain factors on the price.

Conclusion

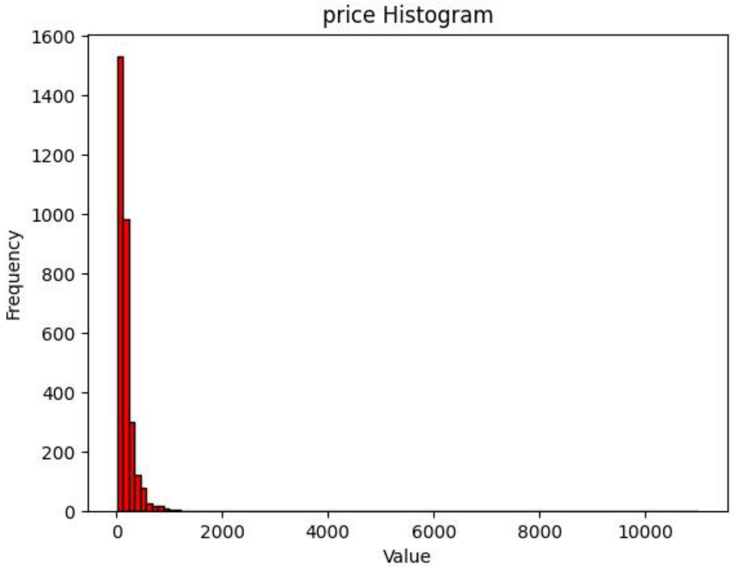
This analysis attempts to find pattern between price for one night in Munich hotels and longitude values through building simple regression with using OLS method. Eventually, the built model figures out the specifics of variables' relationship. First of all, longitude has non-linear effect on hotel price.

Secondly, given data modification, the price for one night in Munich on average is changed on 564% - 0.24%*longitude(i) for obs. with a unit higher X where X fits to 'longitude' with a range of [1139; 1171]. Taking into account the meaning of independent variable, results are reasonable. Median longitude values are associated with higher price because of center location of hotel, while border values are associated with city outskirts. It explains a concave nature of dependence between price and longitude.

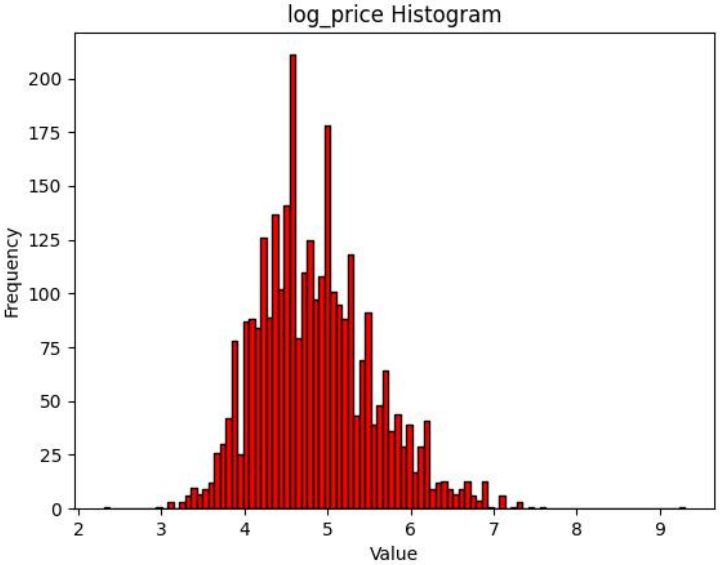
Thirdly, building additional models with room type categorization proves the same pattern effect of independent variable on price.

Appendix

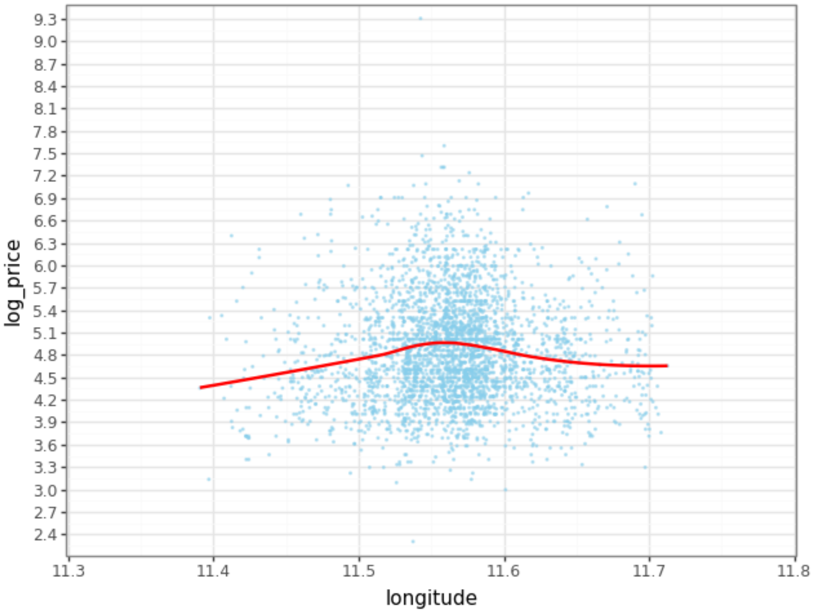
Picture 1.



Picture 2.



Picture 3.



Picture 4.

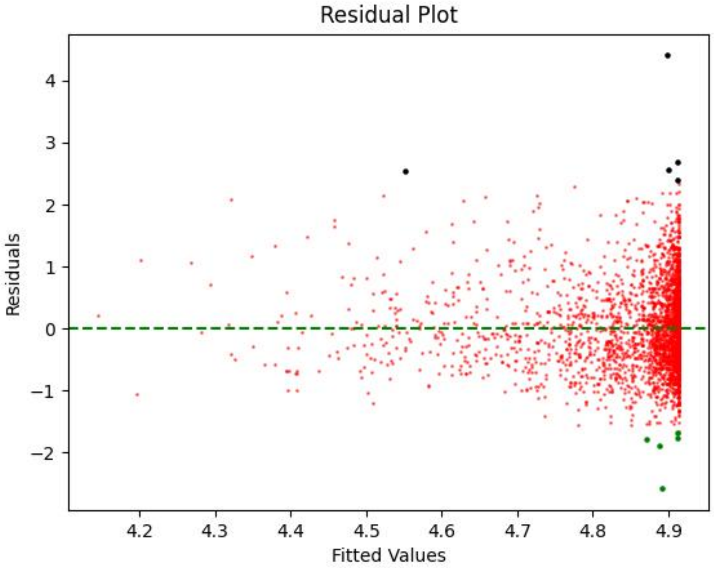


Table 1.

Dep. Variable:	log_price	**R-squared:**	0.000
Model:	OLS	**Adj. R-squared:**	-0.000
Method:	Least Squares	**F-statistic:**	0.4751
Date:	Mon, 18 Nov 2024	**Prob (F-statistic):**	0.491
Time:	16:22:54	**Log-Likelihood:**	-3310.2
No. Observations:	3115	**AIC:**	6624.
Df Residuals:	3113	**BIC:**	6636.
Df Model:	1		
Covariance Type:	HC1		

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
Intercept	2.7557	3.047	0.904	0.366	-3.217	8.728
longitude	0.1816	0.263	0.689	0.491	-0.335	0.698

Omnibus:	205.580	**Durbin-Watson:**	1.897
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	265.302
Skew:	0.605	**Prob(JB):**	2.46e-58
Kurtosis:	3.762	**Cond. No.**	2.74e+03

Notes:

[1] Standard Errors are heteroscedasticity robust (HC1)

[2] The condition number is large, 2.74e+03. This might indicate that there are strong multicollinearity or other numerical problems.

Table 2.

Dep. Variable:	log_price	**R-squared:**	0.021
Model:	OLS	**Adj. R-squared:**	0.020
Method:	Least Squares	**F-statistic:**	30.42
Date:	Mon, 18 Nov 2024	**Prob (F-statistic):**	8.22e-14
Time:	16:59:06	**Log-Likelihood:**	-3277.4
No. Observations:	3115	**AIC:**	6561.
Df Residuals:	3112	**BIC:**	6579.
Df Model:	2		
Covariance Type:	HC1		

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
Intercept	-3259.0321	418.659	-7.784	0.000	-4079.588	-2438.476
longitude	5.6429	0.724	7.795	0.000	4.224	7.062
squared_longitude	-0.0024	0.000	-7.794	0.000	-0.003	-0.002

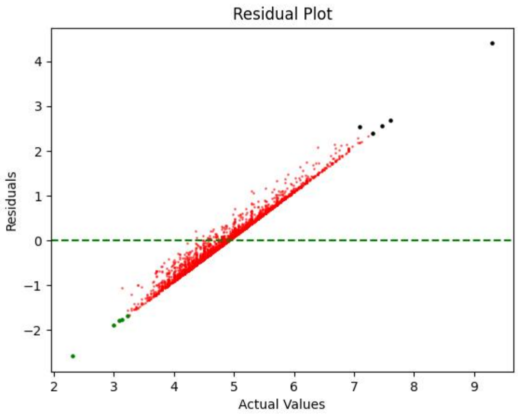
Omnibus:	223.639	**Durbin-Watson:**	1.907
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	297.486
Skew:	0.629	**Prob(JB):**	2.52e-65
Kurtosis:	3.843	**Cond. No.**	4.32e+10

Notes:
[1] Standard Errors are heteroscedasticity robust (HC1)
[2] The condition number is large, 4.32e+10. This might indicate that there are strong multicollinearity or other numerical problems.

Table 3.

	Dependent variable: log$_p$rice	
	Entire home/apt	Private room
Intercept	-1997.306***	-2219.664***
	(610.119)	(411.019)
longitude	3.463***	3.840***
	(1.055)	(0.711)
squared$_l$*ongitude*	-0.001***	-0.002***
	(0.000)	(0.000)
Observations	1984	1131
R^2	0.007	0.020
Adjusted R^2	0.006	0.018
Residual Std. Error	0.649 (df=1981)	0.579 (df=1128)
F Statistic	5.387*** (df=2; 1981)	16.669*** (df=2; 1128)
Note:		*p<0.1; **p<0.05; ***p<0.01

Picture 5.


Residual Plot
Residuals
Actual Values