Data Analysis Report

Data Analysis 2: Assignment 2

Mukhammed Zainidinov

Date: December 8, 2019

Table of contents

1 Part 1

In this part, we prepare our data for analysis using Stata. For the analysis we use hotels-europe_features.dta and hotels-europe_price.dta datasets.

1.1 Data cleaning

First, we use hotels-europe_features.dta to choose the city and features. We pick up the city with large number of observations – Paris – the second largest city after Rome in terms of observation amount. After dropping out other cities there left 2,184 observations. Then we focus only on those hotels that are not far from the city center and keep hotels that are actually in Paris (N = 1,676). Our purpose is only hotels and hostels, so we drop out other types of accommodation (N = 1,508). After filtering our data on hotel features, we merge our cleaned dataset with hotels-europe_price.dta dataset.

Next task for us is which year and month to choose. We choose the month November 2017, and prices for weekdays (N = 1242). As we know, the data contain prices for 1 night and for 4 nights as well. As we are using cleaned data all prices are for 1 night. Then we look for extreme values in prices and there two hotels with the price of 1,072 EUR and other two hotels with the price of 1,126. For the accuracy, we remove these hotels from our data (N=1,238). Since we are using already cleaned data there is no duplicated observation. Now our data is ready for the analysis. There are 16 hostels and 1,222 hotels. We save it as hotels-paris.dta.

1.2 Data description

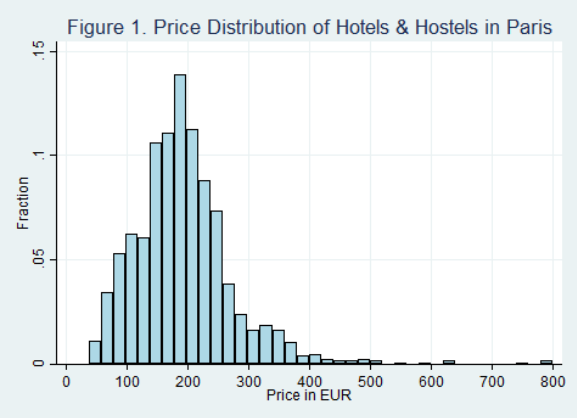
Table 1 shows the descriptive statistics on price, distance, user rating, and stars. Let us start from the price for hotels and hostels. Number of observations is 1,238 with the average price of 194 EUR. Median is smaller than mean, so the price distribution is skewed with a long right tail, it is also seen from the Figure 1. Minimum price is 39 EUR, while maximum one is 789 EUR. About 14% of hotels and hostels charge the price between 180-200 EUR.

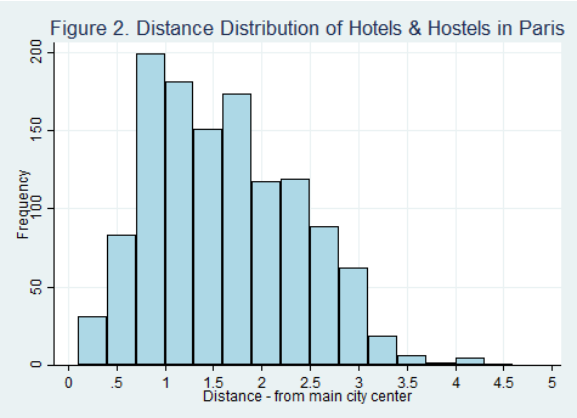
Table 1. Descriptive statistics of selected variables

variable	N	mean	p50	sd	min	max
price	1238	194.1155	186	82.99475	39	789
distance	1238	1.566236	1.5	.7649077	.1	4.5
rating	1233	3.863504	4	.5587684	1	5
stars	1091	3.289643	3	.8039313	1	5

Average distance from the city center to the hotel/hostel is 1.56 miles. Its mean slightly differs from its median, so the distribution should have long right tail. It's seen from Figure 2. The nearest hotel/hostel's distance to the city center is 0.1 miles, while the farthest one is located

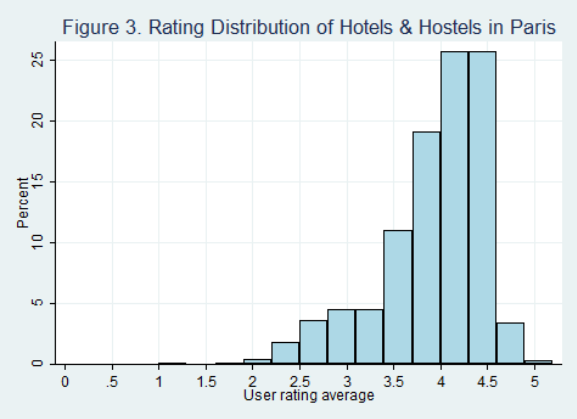
4.5 miles far from the center. Around 200 hundred hotels and hostels are located 0.7-1 miles far from the city center.

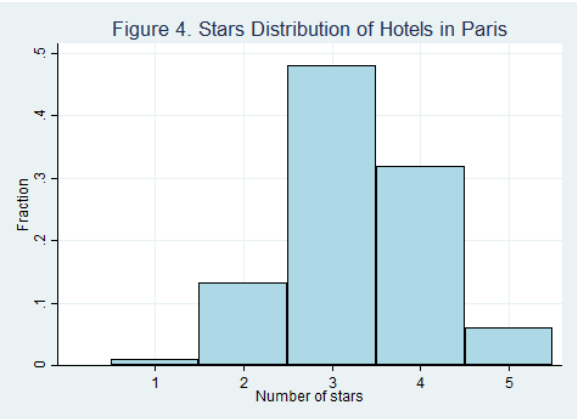




Only 1,233 hotels and hostels have rating available with average rating of 3.86. Its distribution has long left tail, so its median is greater than mean (look at Table 1). More than 50% of hotels and hostels have rating between 4 and 4.6.

Only hotels have stars. There are no hotel in Paris that has in-between stars. Median for stars is 3, while its average is 3.86. 48% of hotels have 3 stars, 32% have 4 stars. Only 6% of hotels have 5 stars.



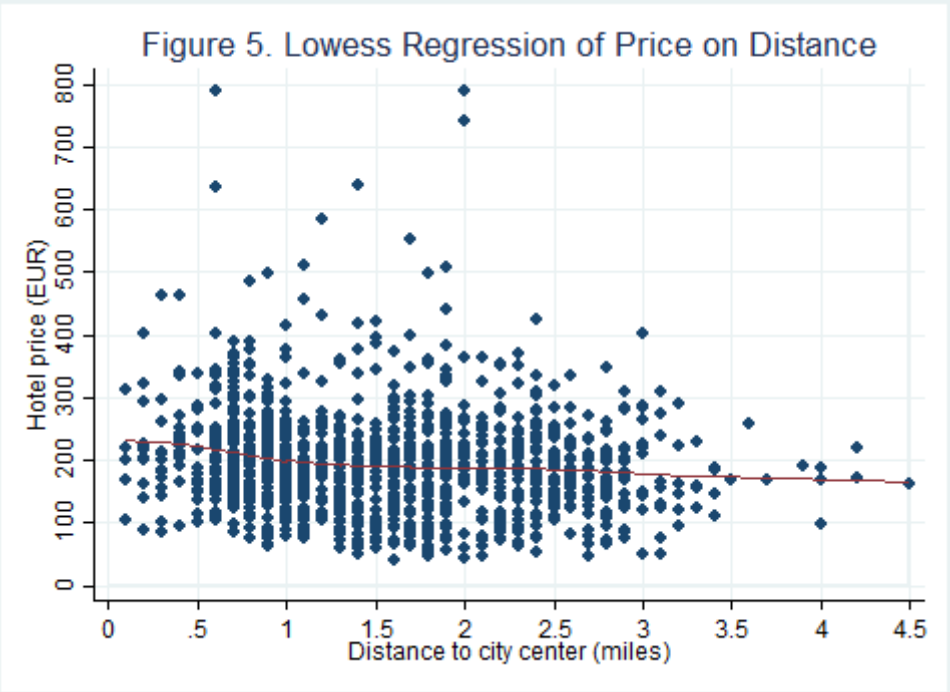


2 Part 2

In this part, we run linear, non-linear, nonparametric simple regressions of price on distance.

2.1 Lowess non-parametric regression

Figure 5 depict the scatter plot of lowess regression. We see that till 1 mile the price for hotel/hostel decreases when distance increases. After 1 mile the slope is flatter.



2.2 Simple linear regression

For the simple linear regression our model:

$$price = \alpha + \beta distance$$

We run robust regression. Our estimated model:

$$\widehat{price} = 213.37 - 12.29 distance$$

(5.192) (2.797)

N = 1238, R^2 = 0.0128

A hotel/hostel located right at the center, on average, charges the price of 213.37 EUR for 1 night. As distance from the city center increases by 1 mile, hotel/hostel tends to have the price, on average, 12.29 EUR lower. R-squared is very low, we can conclude that there are other variables that affect the price.

2.3 Simple linear regression that captures potential nonlinearities

In this section, we consider various types of regression.

2.3.1 Log-level regression

Model:

$$\ln(price) = \alpha + \beta distance$$

Estimated model:

$$\widehat{\ln(price)} = 5.286 - 0.066\ distance$$

N = 1238, R^2 = 0.0142

Hotels/hostels that are 1 mile further away from the city center are, on average, 6.6% cheaper in our data.

2.3.2 Level-log regression

Model:

$$price = \alpha + \beta \ln(distance)$$

Estimated model:

$$\widehat{price} = 199.51 - 17.67 \ln(distance)$$

N = 1238, R^2 = 0.0153

As distance from the city center increases by 1%, on average, the price for hotel/hostel falls by 0.18 EUR. The hotel/hostel's price located right 1 mile from the center, on average, is 199.51 EUR.

2.3.3 Log-log regression

Model:

$$\ln(price) = \alpha + \beta \ln(distance)$$

Estimated model:

$$\widehat{\ln(price)} = 5.21 - 0.099\ln(distance)$$

N = 1238, R^2 = 0.0142

As distance from the city center increases by 1%, on average, the price for hotel/hostel falls by 0.1 EUR.

2.3.4 Regression with quadratic polynomial

Estimated model:

$$\widehat{price} = 227.37 - 32.417\ distance + 5.766\ distance^2$$

N = 1238, R^2 = 0.0153

In this model, coefficient of distance is not significant. As distance from the city center increases by 1 mile, on average, the price for hotel/hostel falls by 20 EUR.

2.3.5 Regression with piecewise linear spline

Model:

$$price = \alpha_1 + \beta_1\, distance[if\ distance < 1] + (\alpha_2 + \beta_2) distance[if\ 1 \leq distance < 3] + (\alpha_3 + \beta_3\, distance[if\ distance \geq 3]$$

Below the results of the regression. R-squared is 0.0205. We see that coefficients of distance_2 and distance_3 are not statistically significant. But, for practice we can interpret. When we compare hotels with distance less than 1 mile, price is 58.4 EUR less, on average, for observations 1 mile farther from the center. When we compare hotels with distance between 1 and 3 miles, price is 3.93 EUR less, on average, for observations 1 mile farther from the center. When we compare hotels with distance more than 3 miles, price is 14.3 EUR less, on average, for observations 1 mile farther from the center. The slope is steeper between 0-1 miles.

Table 2. Piecewise linear spline regression

price	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
distance_1	-58.40422	15.70914	-3.72	0.000	-89.2238	-27.58464
distance_2	-3.928129	3.795542	-1.03	0.301	-11.37456	3.5183
distance_3	-14.30233	12.46494	-1.15	0.251	-38.75714	10.15249
_cons	250.3375	13.88462	18.03	0.000	223.0975	277.5776

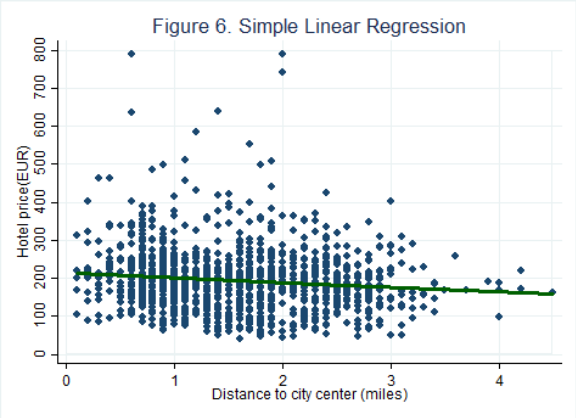
2.4 Discussion of overall findings

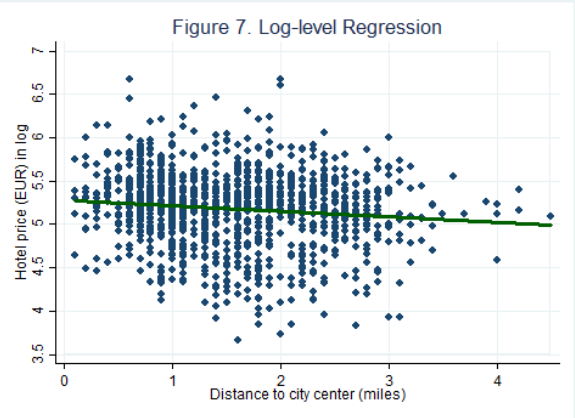
Table 3 depicts the results of level-level and log-log simple linear regressions, and with quadratic polynomial. All coefficients are statistically significant at 1% level except the coefficient for squared distance. If we rely on R-squared, then log-log model explains better the pattern between price and distance.

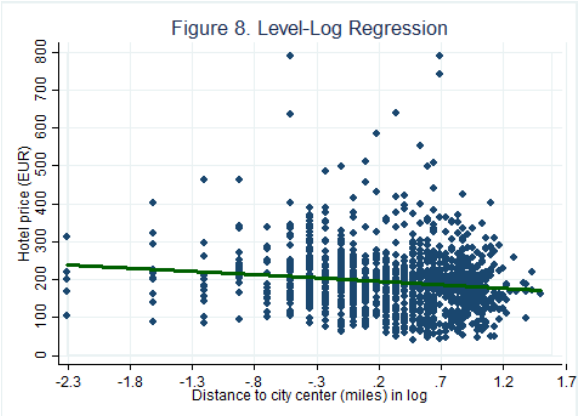
Table 3. Regression results

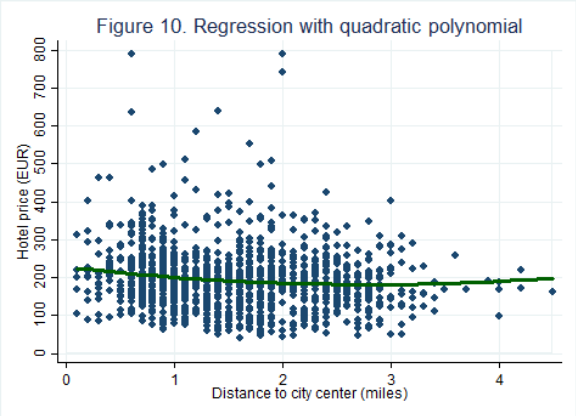
VARIABLES	(1) Price in EUR	(2) Natural logarithm of price	(3) Price in EUR
Distance - from main city center	-12.29***		-32.42***
	(2.797)		(11.303)
Natural logarithm of distance		-0.10***	
		(0.019)	
Squared distance			5.77**
			(2.893)
Constant	213.37***	5.21***	227.37***
	(5.192)	(0.013)	(9.598)
Observations	1,238	1,238	1,238
R-squared	0.013	0.018	0.015

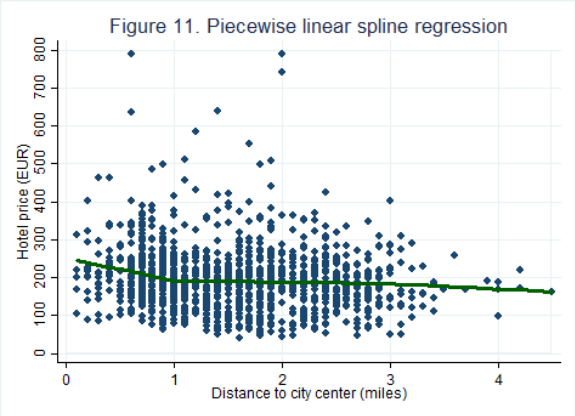
Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1











3 Part 3

3.1 Multiple regression

Our model for multiple regression:

$$price = \beta_0 + \beta_1\, distance + \beta_2\, stars + \beta_3\, rating$$

Below in the table results of the regression. N = 1,091 and R-squared is 0.5099, which means 51% of all true values lie on the fitted model. All coefficients are statistically significant at 5% level. Hotels that are 1 mile further away from the city center are, on average, 4.4 EUR cheaper in our data, holding stars and rating constant.

For hotels that are located in the same place and have the same ratings, having 1 more star, on average, the price is higher by 57.51 EUR.

For hotels that are located in the same place and have the same stars, having 1 more rating, on average, the price is higher by 32.43 EUR.

Table 4. Results of multiple regression

price	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
distance	-4.402199	2.064199	-2.13	0.033	-8.452466	-.3519328
stars	57.50701	3.299579	17.43	0.000	51.03275	63.98128
rating	32.43014	4.259458	7.61	0.000	24.07245	40.78783
_cons	-107.3222	16.80827	-6.39	0.000	-140.3025	-74.34182

Standard error of distance coefficient is large. True values falls between -8.45 and -0.35 with 95% confidence. It is close to 0, but it is significant at 5% level.

Standard error of stars is small. True values falls between 51.03 and 63.98 with 95% confidence.

Standard error of stars is small. True values falls between 24.07 and 40.79 with 95% confidence.

3.2 Multiple regression having some variables in a non-linear form

Our model for multiple regression:

$$price = \beta_0 + \beta_1\, distance + \beta_2\, stars + \beta_3\, stars^2 + \beta_4\, rating + \beta_5\, rating^2$$

In this model β_1 and β_4 is not statistically significant. Other coefficients are significant at 5% level. N = 1,091 and R-squared is 0.5612. Holding other variables constant, hotels that are 1 mile further away from the city center are, on average, 2.76 EUR cheaper in our data.

For hotels that are located in the same place and have the same ratings, having 1 more star, on average, the price is lower by 31.54 EUR.

For hotels that are located in the same place and have the same stars, having 1 more rating, on average, the price is lower by 33.41 EUR.

Having one more star and one more rating decreases the price. Therefore, this model is not true.

Table 5. Results of multiple regression having some variables in a non-linear form

price	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
distance	-2.761977	1.978333	-1.40	0.163	-6.643768	1.119814
stars	-68.96115	16.52114	-4.17	0.000	-101.3782	-36.54415
stars2	18.71287	2.660871	7.03	0.000	13.49184	23.93391
rating	-60.10051	37.73533	-1.59	0.112	-134.143	13.94198
rating2	13.34766	5.267804	2.53	0.011	3.01142	23.68389
_cons	245.8724	73.11744	3.36	0.001	102.4048	389.34

3.3 Multiple regression with interaction term

Our model for multiple regression:

$$price = \beta_0 + \beta_1\, distance + \beta_2\, rating + \beta_3\, four\_stars + \beta_4\, rating * four\_stars$$

After considering only hotels with 3 and 4 stars, number of observations is 871. There is 3.5 stars hotel. R-squared is 0.3461. Only β_1 and β_4 are statistically significant.

β_0: 3-star hotels that are located right at the center and have 0 rating, on average, charges 23.1 EUR.

β_1: Hotels that are 1 mile further away from the city center are, on average, 4.48 EUR cheaper in our data, holding rating and stars constant.

β_2: 3-star hotels for one additional rating, on average, charges 42.34 EUR more, holding distance constant.

β_3: Holding distance and rating constant, 4-star hotels charge 82.811 EUR more than 3-star hotels.

β_4: 4-star hotel's price for one additional rating is 9.65 EUR cheaper than for 3-star hotels, on average.

Table 6. Multiple regression with interaction term

price	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
distance	-4.479659	1.847515	-2.42	0.016	-8.10579	-.8535278
rating	42.34705	4.853877	8.72	0.000	32.82031	51.87379
four_stars	82.81124	44.40488	1.86	0.063	-4.342541	169.965
ratingXfour_stars	-9.649344	10.66005	-0.91	0.366	-30.57189	11.27321
_cons	23.09758	19.29179	1.20	0.232	-14.76654	60.96171