Can highly rated hotels charge higher prices?

Term Project for Data Analysis 2

Lili Márk, Ágoston Reguly

2018/12/17

Abstract

We address the question, whether hotels, whose rating is higher, can charge a higher price. Our results based on data from 2017, November, at Vienna among 3-4 stars hotels, suggests one unit higher ration is associated with 49% higher price. This can be important for hotel managers, while they probably can focus on achieving higher user rating (via better services), thus aiming for higher prices.

We are interested in the connection of average rating by the users and hotel prices for a night. We argue, that users feedback on hotels has a strong association on price. We believe that if users give positive feedback then potential users will be willing to pay extra amount of money to be sure that the quality of the overnight will be better on general in a hotel. This also give a comparison for hotel managers, while given their ratings (and other variables) they can have an idea how to price their service.
In order to address this question we select a specific data set from Vienna and propose a log-level linear model, which gives some insight for this question.

1 Data

We restrict our attention to hotels which are located in Vienna in 2017, November, without weekends. We scraped the data from the hotel comparison website. The quality of the data is rather good, therefore we do not need to worry about systematic measurement errors in our variables. However, beyond removing duplicates, we drop hotels with prices above 600 EUR. In our view for 3-4 star hotels, these must be some errors in the data.

Furthermore, we filtered our data a bit more. We choose hotels, we dropped hotels with a user rating below 3 as we only have 4 observations below 3 rating, and any finding in that range is likely to be random. We checked them and they are based on very few reviews, confirming our call. We argue for these hotels, there might be some different mechanism going on, and our main objective is to create a model which can explain the

majority of the hotels. We acknowledge that our finding is relevant for hotels with at least a 3 user rating.

We also excluded hotels which are referred as in Vienna, but in fact they are in Fischamend, Schwechat and Voesendorf. We argue, there are further away from the city and therefore there are other factors for an individual to choose such places, that we can not include in our model.

We want to model average ratings on prices. Our aim, is to compare hotels that are similar in many ways but differ in user rating. To this purpose, we also include controls in the analysis, such as the number of stars and distance from the city center.

The following table shows the descriptive statistics of these variable. For variables in

	Price	log(Price)	Ratings	Stars	Distance
Mean	110	4.64	4.06	3.57	1.53
Median	100	4.61	4.10	4.00	1.3
Std.	42	0.34	0.38	0.48	1.16
Min	50	3.19	2.20	3.00	0.0
Max	383	5.98	4.80	4.0	6.60

Table 1: Descriptive statistics of the variables

quantities, very often we are interested in relative differences, which is helped by a log transformation. We also look at histograms, as patterns are easier found when the right skewed data is transformed.

As prices usually log-normally distributed, we check for a logarithmic transformation and as 1. figure – by eyeballing – suggests it is favourable to do such transformation. Other control variables looks fine, therefore there is no need for transformations. Overall

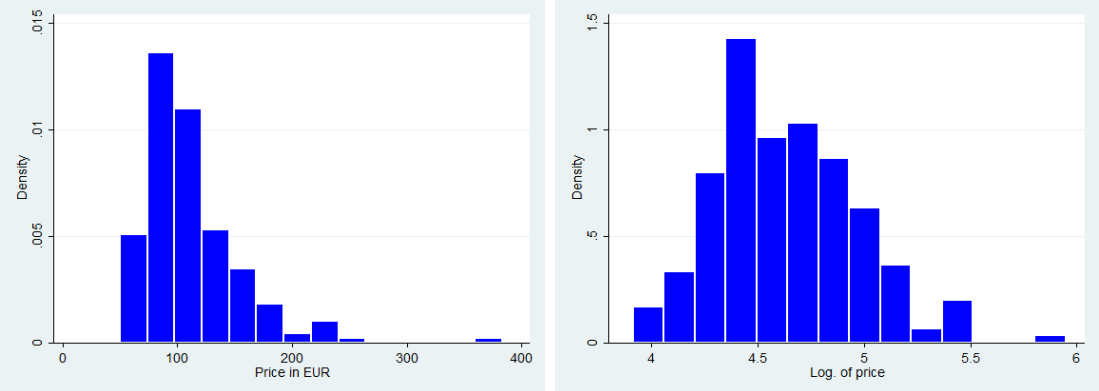

Figure 1: Histogram of prices (left) and log of prices (right)

we have 203 observations.

2 Model

We want to regress ratings on log-prices. First we check a non-parametric estimator – lowess smoother – to have a general idea about the functional form between these two variables. Between log-price and unit rating, here we see a clear linear up-warding trend.

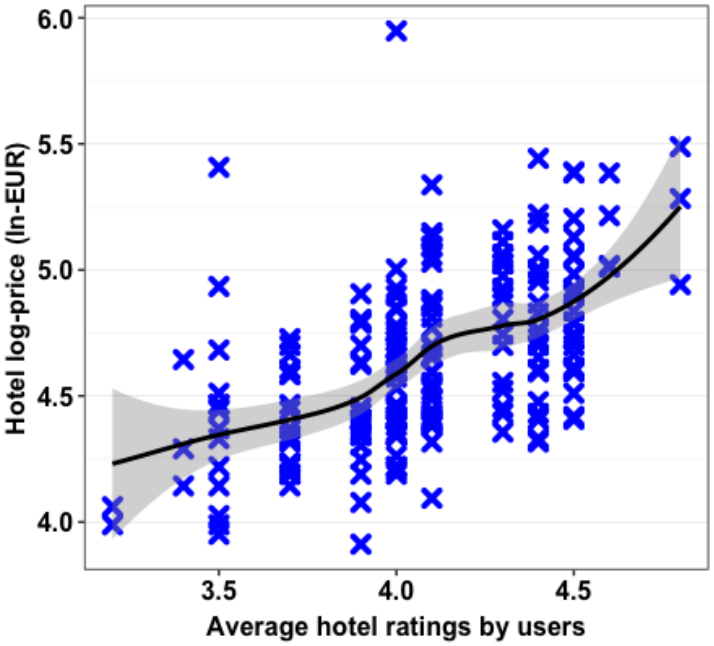

Figure 2: Lowess estimator for log-prices and user rating average

We will capture this connection with a linear model.
For the "simple model" we only use the rating on the right hand side. We use the following extended model:

$$\log(price_i)^E = \beta_0 + \beta_1 rating_i + \beta_2 stars_i + \beta_3 distance_i$$

Dep. var.	Log(Price)	Log(Price)
VARIABLES	Simple	Extended
rating	0.56***	0.40***
	(0.062)	(0.061)
stars		0.17***
		(0.041)
distance		-0.10***
		(0.021)
Constant	2.37***	2.54***
	(0.254)	(0.246)
Observations	203	203
R-squared	0.290	0.460

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Table 2: Linear regression results

The results suggest that *without* controlling for any other variable, higher rating is associated with 75% higher price (0.56 log units)[1]. In other words, Vienna hotels with a unit higher user rating are, on average, 75% more expensive.

To learn more about the role of ratings, let us try comparing hotels that are similar to each other but differ in terms rating. To this end, let us to control for stars and distance. When we compare hotels with the same distance and stars, we find that hotels with a unit higher rating will have on average 49% higher prices.

Based on our extended model we can state with 95% confidence the association is between 40-58%. It seems stars and distance can explain valuable part of variation in (log) prices. Stars has positive, while distance has negative association with prices. Both are significant at 1%. The R^2 increased from 29% to 46%, thus the extended model is a better fit.

2.1 Robustness analysis

We might be worried, that we have missed some important patterns or our analysis is true only for a specific sample. Therefore we check for four alternative specification:

1. we include those observations which has ratings below 3, thus we add the flat part to our analysis

2. hotels with only 3 stars

[1]While change is not close to zero, we need to use the following formula to express the change in percentages: $e^{\beta_1} - 1$

3. hotels with only 4 stars

4. use piece-wise linear spline for distance, using threshold at 2km. This is reasoned by a scatter plot between log-price and distance: there is a clear change in the slope at 2 km.

. The first 3 specification is investigating whether the results are sensitive to the sample, while the fourth is changing one functional form. Table 3. shows the estimation results. The main interest is the coefficient on rating. Apart from specification (2), there is no

	(1)	(2)	(3)	(4)	(5)
Dep. var.	Log(Price)	Log(Price)	Log(Price)	Log(Price)	Log(Price)
VARIABLES	Main	All	3-star	4-star	Dist.spline
rating	0.40***	0.29***	0.41***	0.45***	0.38***
	(0.061)	(0.063)	(0.072)	(0.109)	(0.050)
stars	0.17***	0.17***			0.15***
	(0.041)	(0.041)			(0.038)
dist1					-0.24***
					(0.034)
dist2					0.03
					(0.030)
distance	-0.10***	-0.09***	-0.05*	-0.13***	
	(0.021)	(0.021)	(0.029)	(0.026)	
Constant	2.54***	2.98***	2.97***	3.07***	2.83***
	(0.246)	(0.257)	(0.295)	(0.481)	(0.218)
Observations	203	207	78	111	203
R-squared	0.460	0.417	0.246	0.458	0.531

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

Table 3: Alternative specification

significant change in the parameter. Testing for parameter equality, among models, we has the conclusion that these are the same values at 95% significance level. One exception, is model specification (2). As we have argued before there might be some omitted variables, which have effect on both log-prices and rating (e.g. being a youth hostel). Thus without modeling for those as well, we can not say anything about the parameter value. Other parameter values are also stable.

3 Causality and external validity

Does higher user rating make hotel price go up? is there a causal relationship?

So far we made a small step towards causality when we added control variables. Indeed for hotels that are of the same distance from the center and have equal stars, we found a positive correlation between price and user rating.

However, there could be many other factors at play, most likely there are confounders. Hotels that have some extra services such as a swimming pool are liked by hotels, but also cost a lot raising price demanded. In nice neighborhoods, users may enjoy the ambiance, but rent prices may be also higher.

Hence, we believe we found a useful pattern, but we would need much more control variables to believe we are close to causality.

An other factor is that we have used a specific sample from Vienna in 2017 November. Even though results suggests strong internal validity, the external validity may be weak, customs and user behavior may vary across cities. Further investigation is needed, if we would like to extend our suggestions to other cities or to other time periods.

4 Summary

We have analyzed the relationship between hotel rating and hotel prices. We have used a log transformation for prices, and extended our model with distance and number of stars. Using a linear model we arrived to the conclusion, that one unit higher rating is associated with 49% higher prices. Changing the model specification does not seems to have a significant effect on the parameter of interest, if we restrict our attention to ratings above 3.

Despite not being of causal nature, these results might be beneficial for hotel managers: higher rating may well give the possibility to set higher prices.