Report on hotel ratings

Viktória Kónya

03 December 2021

Introduction

In this report, I examine the important predictors of high user ratings using data of different types of accommodations in Barcelona. My goal is to uncover how high ratings of the hotels are connected to other features of the hotels such as the distance from the city center, the number of stars and the type of the accommodation.

Descriptive statistics

The outcome variable of our analysis is the rating class of the hotels. Hotels with an average rating of 4 or higher were classified as 'highly rated' and those below this rating as 'not highly rated'. The below table summarizes the most important features of the variables of our interest by the rating category.

Table 1: Descriptive statistics

Rating category	Variable	N	Missing	Mean	SD	Min	Max	P05	Median	P95
Highly rated	Distance	436	0.00	1.27	0.73	0.10	4.60	0.30	1.10	2.62
	Stars	436	0.00	3.58	0.82	1.00	5.00	2.00	4.00	5.00
	Reviews	436	0.00	208.30	227.96	1	2303	6.00	142.00	577.50
Not highly rated	Distance	284	0.00	1.17	0.75	0.10	3.10	0.22	1.00	2.50
	Stars	284	0.00	2.80	0.90	1.00	5.00	1.00	3.00	4.00
	Reviews	284	0.00	112.37	141.98	1	1000	2.00	70.50	322.55

Almost 60% of the hotels have 4 or higher ratings. We can see that the average distance from the city center is about 0.1 kilometer higher in case of the highly rated hotels. The number of hotel stars range between 1 and 5 stars in both rating groups, however the median number of stars is 4 in case of the highly rated, and 3 in case of the hotels with lower ratings. We can also see that the number of rating reviews range in a wide scale including observations with very few ratings. For the subsequent analysis, I excluded hotels with less than 50 reviews to avoid this noise. I also kept observations of all types of accommodation in the dataset. Note that I will use hotels and accommodation interchangeably in the descriptions.

Regression analysis

In order to analyze the predictors of high ratings, I used LPM, logit and probit models to estimate the probability of being highly rated. The left hand side variable is the binary variable showing if the hotel had 4 or above rating, and the predictors are the number of stars, the distance from the city center and the type of the accommodation. Accommodation types with very few observations were grouped together in order to create meaningful categories.

In the LPM model the predicted coefficient of stars is 0.23 and statistically significant at 1%. Comparing hotels with the same distance from the city center and of the same type, hotels with one additional star are 23 percentage point more likely to be highly rated. Distance has positive coefficient suggesting that, conditional on other characteristics, hotels one kilometer farther away from the city center are 2.7 percentage point more likely to have high ratings. In case of the accommodation types, the control group is the hotels. We can see that apartments are 7.7, pensions are 3.7 and hostels are 9 percentage points less likely to have 4 or above ratings than hotels conditional on the other characteristics, however these coefficients are statistically not significant.

The marginal differences from the logit and probit models are almost the same and are very similar to the corresponding coefficients from the LPM model. If we look at the number of stars, we can see that hotels located at the

Table 2: Probability models

	LPM	Logit coeffs	Logit marginals	Probit coeffs	Probit marginals
Constant	−0.106	−3.413**		−2.027**	
	(0.082)	(0.528)		(0.301)	
Stars	0.230**	1.279**	0.213**	0.761**	0.216**
	(0.020)	(0.154)	(0.034)	(0.086)	(0.019)
Distance	0.027	0.162	0.027	0.090	0.025
	(0.025)	(0.154)	(0.027)	(0.089)	(0.026)
Apartment	−0.077	−0.577*	−0.097*	−0.324*	−0.093*
	(0.049)	(0.269)	(0.046)	(0.159)	(0.046)
Pension and B&B	−0.037	−0.105	−0.018	−0.062	−0.018
	(0.073)	(0.371)	(0.061)	(0.222)	(0.061)
Hostel	−0.090	−0.272	−0.047	−0.173	−0.051
	(0.082)	(0.450)	(0.077)	(0.267)	(0.078)
Num.Obs.	506	506	506	506	506

* p < 0.05, ** p < 0.01

same distance from the center and of the same type have 21.3 and 21.6 percentage points higher chance to be highly rated if they have one star higher hotel rating. The coefficients and the robust standard errors of the distance are technically the same as from the LPM model and similarly, they are statistically not significant. The logit and probit models indicate that apartments are 9.7 and 9.3 percentage point less likely to have high hotel ratings than hotels, conditional on the other attributes of the hotels and this difference is statistically significant. Overall, we can conclude that the three estimated models give very similar results.

Let's also compare the goodness of fit measures of the three probability models.

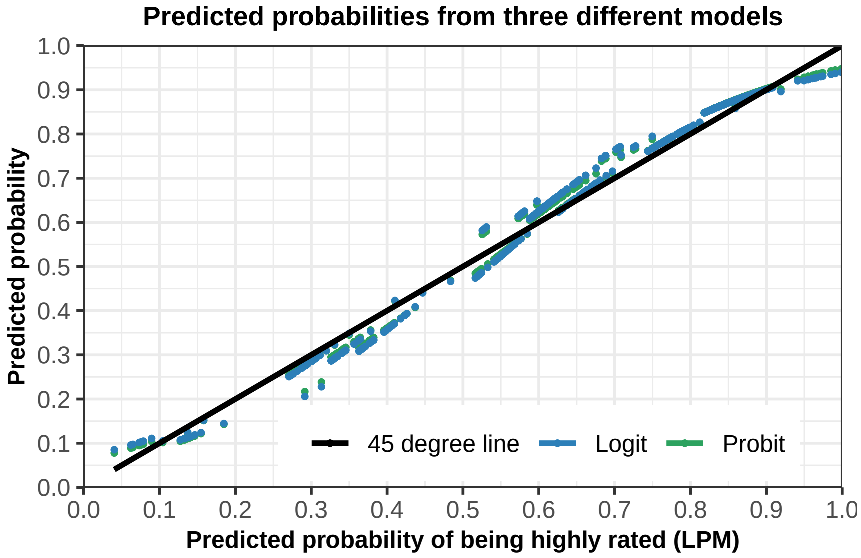
Table 3: Statistics of goodness of fit for the probability predictions

Statistic	LPM	Logit	Probit
R-squared	0.2504	0.2532	0.2533
Pseudo R-squared	Na	0.2109	0.2120
Brier score	0.1682	0.1676	0.1675
Log-loss	-0.5033	-0.5058	-0.5051

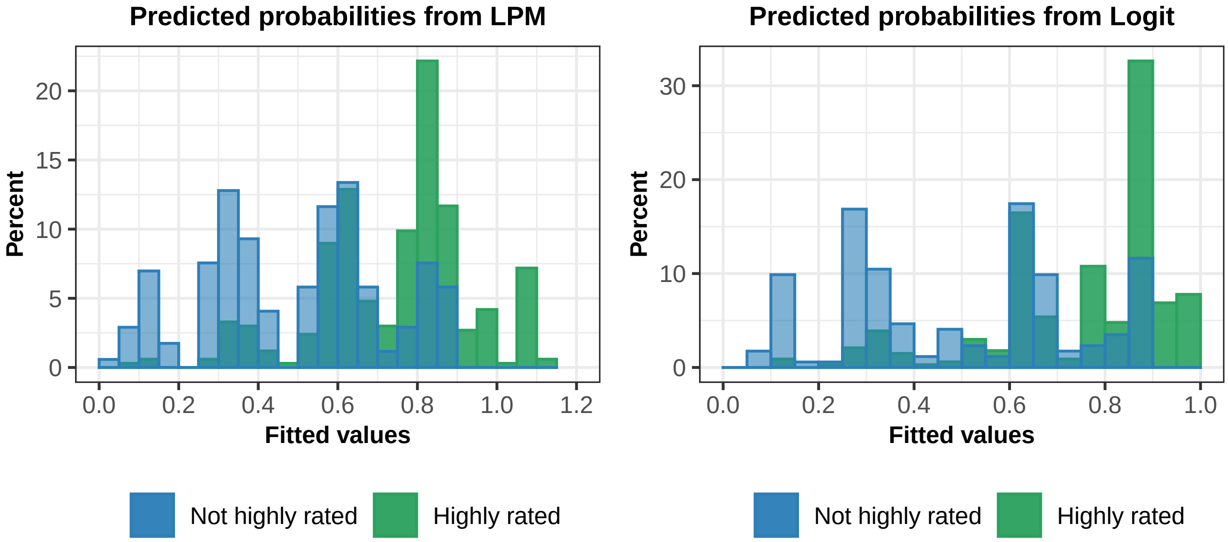
As expected, the performance of the models is very similar. The predictions of the logit and the probit models are slightly better in case of the Brier score and a little bit worse in case of the Log-loss compared to the LPM model, but the differences are not considerable.

Predictions

The following figure plots the predictions from the logit and probit models on the y axis against the predictions from the LPM model on the x axis. We can clearly see that the logit and probit predictions move closely together. If we compare them to the LPM model predictions, we can see that the scatterplot of the predicted values slightly vary from the LPM predictions which is more visible in the tails.



Finally, let's take a quick look at the predicted probabilities of being highly rated by the actual rating categories.



The LPM model has predictions out of the 0-1 range for 5.3% of the observations. The histograms suggest that in case of both models the distribution of the fitted probabilities among hotels with high ratings are are shifted towards the higher edge of the probability range. However, the distributions are overlapping which suggests that the predictive power of the two models are not that strong.

Summary

To sum up, our analysis showed that hotels with more stars are more likely to be highly rated after conditioning on the location and the accommodation type. Moreover, the marginal differences in the logit and probit models produced very similar results to the simplest linear probability model. We can conclude that using the LPM model to uncover the associations is just as fine as the more complicated logit or probit models, but if we care about prediction we might choose to use the latter two models instead.