Data Analysis - Assignment 2 - Regression

Fatima Arshad

Introduction

What features make hostels acquire a high rating? Are the hotels which are rated high and are nearer to the city center comparatively expensive? This study is conducted to answer these questions by estimating linear probability, logit, and probit models with distance and stars as explanatory variables. We join the **Hotel Features dataset** and **Hotel Price dataset** for Europe. For the purpose of this study the city we will choose is ***Paris***.

Data Transformation and initial findings

For this analysis, we will create the following new dependant variables: 1. A binary variable ***high_rating*** with a value of 1 when rating>4 and a value of 0 otherwise. 2. A binary variable ***more_stars*** with a value of 1 if stars>4 and a value of 0 otherwise. The visualize Lowess regression for high rated hotels and distance. By examining the kinks in Figure 1 we put two knots at 1.2 and 3 miles. We introduce two control variables i.e. the natural log of price (log_price). Finally, we will remove null and duplicated values from the dataset.

Analysis and Interpretation

The calculations in the summary table tells that we have huge number of hotels with a high rating since the mean lies above 0.5. Six regression models are shown in which are: 1. lpm0, 2. lpm, 3. logit, 4. marginal logit, 5. probit and 6. marginal probit. The lmp0 constant tells us that hotels with more stars are 42.9 percent more likely to be highly rated. Looking at the 95% confidence interval around the [0.411 ,0.447] slope parameter tells us that we are 95% confident having more stars in the sample means being highly rated. Interpreting the predicted probability graph, firstly as the unit distance from the city center increases (i.e. 1 - 1.2 miles) this probability of high rated hotels is observed to decrease by 10.4 percent. Secondly, as the unit distance increases between the 1.2 to 3 miles radius does not have an effect on the high rating hotels. Moreover, the probability for any distance beyond a 3 miles radius from the city center decreases by 22.9 percent.Next, the model's logit and probit estimates suggest that the probability of high rated with more stars, distance price are similar to the linear model. The logit coefficients in the table are approximately five times the values of the corresponding logit marginal differences. Moreover, the probit coefficients are three times the values of respective logit marginal differences. Furthermore, the logit and probit marginal differences have the same values, which is why we will interpret the coefficients of both, logit and probit marginal differences. Figure 2 visualized all the models. The y-axis lists the predicted probability of logit and probit whereas the x-axis shows the predicted probability of LPM.As visible in the figure 2 graph, the logit and probit represented by the s-shaped curve, lie close to the LPM represented by the 45 degree line. We can generalize that hotels with more than four stars have 26% higher probability to be high rated given that all other variables are kept constant. By looking at the logit and probit estimates for the model, the probability of highly rated to top stars, distance and conditional on price are same variables as linear model. By looking to the column 3 and 4, the Logit Coefficients are almost five times the size of corresponding logit marginal differences. Furthermore, in the column 5 and 6, probit coefficient is almost three times the size of corresponding probit marginal differences. It is interesting to observe that the two marginal differences, logit and probit, are the same and they are the same with LMP coefficients in column 2 which is applicable

Table 1: Summary Statistics

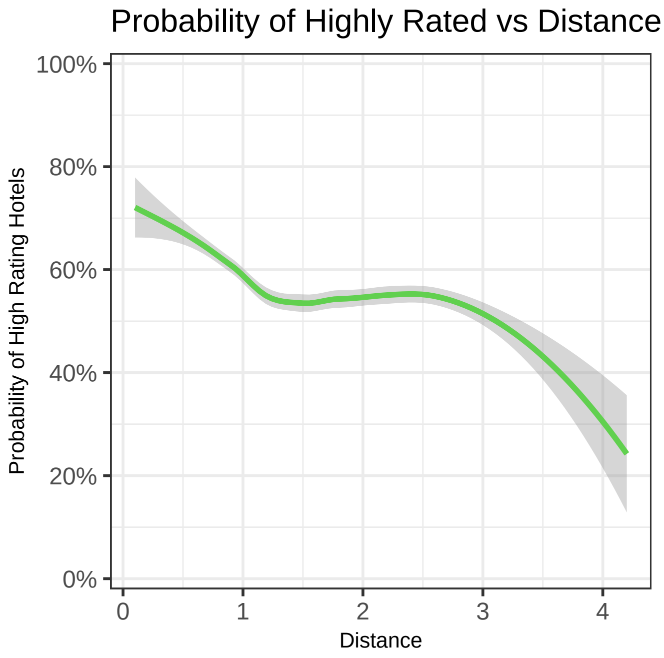
	Mean	SD	Min	Max	Median	P95	N
high_rating	0.57	0.49	0.00	1.00	1.00	1.00	12035
distance	1.61	0.78	0.10	4.20	1.50	2.90	12035
stars	3.25	0.79	1.00	5.00	3.00	5.00	12035

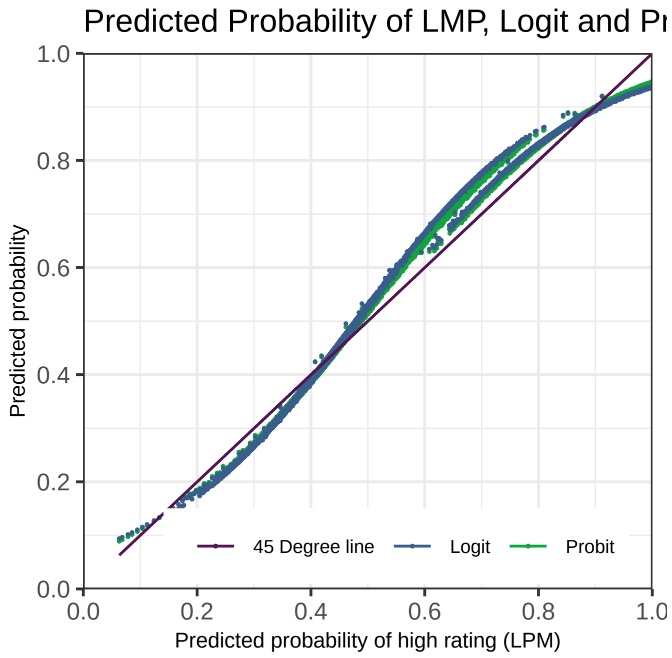
Table 2: Probability of high rating hotels and more stars - LMP, Logit, and Probit models

	LMP0	LMP	logit coeffs	logit Marg	Probit	Probit Marg
Constant	0.534**	−0.483**	−5.473**		−3.226**	
	(0.021)	(0.044)	(0.260)		(0.152)	
more_stars	0.429**	0.334**	1.664**	0.338**	1.006**	0.340**
	(0.009)	(0.009)	(0.052)	(0.010)	(0.030)	(0.010)
lspline(distance, c(1.2, 3))1	−0.104**	−0.051*	−0.242*	−0.046*	−0.144*	−0.045*
	(0.021)	(0.020)	(0.108)	(0.021)	(0.065)	(0.021)
lspline(distance, c(1.2, 3))2	0.002	0.015	0.069	0.013	0.031	0.010
	(0.009)	(0.008)	(0.044)	(0.008)	(0.026)	(0.008)
lspline(distance, c(1.2, 3))3	−0.229**	−0.227**	−1.186**	−0.224**	−0.684**	−0.216**
	(0.049)	(0.048)	(0.266)	(0.043)	(0.155)	(0.041)
log_price		0.190**	1.070**	0.202**	0.629**	0.198**
		(0.007)	(0.044)	(0.010)	(0.026)	(0.008)
Num.Obs.	12 035	12 035	12 035	12 035	12 035	12 035

* $p < 0.05$, ** $p < 0.01$

for of the independent variables. To generalize the result, it shows that hotels with top stars other things (distance, price) the same are highly rated. To sum, top stars hotels have a 43 percent points higher chance to be highly rated.


Probability of Highly Rated vs Distance
Probability of High Rating Hotels
Distance


Predicted Probability of LMP, Logit and Pr
Predicted probability
Predicted probability of high rating (LPM)
45 Degree line
Logit
Probit