Data Analysis-Assignment2

Khawaja Hassan & Maryam Khan

Overview:

The purpose of this assignment is to analyze how highly-rated variable was associated with other hotel features.We created a binary variable for highly rated based on user rating >= 4 as 1 and lesser than that as 0. Moreover, we based our regression model on hotels in Lisbon and used independent variables like stars, distance & log prices. Lastly the data set used in this project is Hotels Europe pricing and feature from OSF project repository.

Data filtering & Lspline

To make sure our analysis is more associative to the variables we filtered and cleaned the data by removing any observation where star and distance were null. To further narrow down our analysis we filtered our data on year 2018 and months greater than 4. To start of with our regression model first we run simple linear regression to estimate if we needed to incorporate spline in our independent variable.

Estimated Models:

Initially when we run our LPM our maximum and minimum probability was within the given range of 0-1. However, we know that it is possible if we add more observation it might exceed the given range. Therefore, we decided to run logit & probit regression along with average marginal effect to comprehend not only the likelihood but also the magnitude of the coefficients.

Interpretations:

Our analysis consists of 5 models shown in Table 2. In our analysis there are two explanatory variables (stars 1-4) and log(price) which lie in the 99% confidence interval therefore we shall be using their magnitudes to understand the association with the dependent variable. The LPM model shows that for one unit increase in the star rating between 1-4 stars range the probability of hotel being highly rated increases by 20.9%. In the same star range however, Logit and Probit average marginal difference show probability of hotel being highly rated increases by 13% instead. Similarly, the LPM model also shows that for 1% increase in price, the probability of highly rated hotel increases by 24.8%. In comparison the Logit and Probit AME suggest an increase in probability of highly rated hotel by increases 22% for same increase in price. For all other variables, because we could not predict probability of hotels being highly rated with 1% significance level, we will not be using their coefficients to deduce the change in magnitude of the dependent variable. The sign of those coefficients' values, however correctly suggest the direction of the change (with the exception of Stars >=4).To summarize, we infer that Logit and Probit are very similar with each other and very close to LPM as shown by the S-shaped curve lying close to 45 degree line.

Table 1: Data Summary Table

	mean	SD	Min	Max	Median	P95	N
highly_rated	0.78	0.41	0.00	1.00	1.00	1.00	493
distance	1.18	1.05	0.00	5.00	1.00	3.98	493
stars	3.64	0.89	1.00	5.00	4.00	5.00	493
lnprice	5.12	0.45	3.81	6.36	5.13	5.82	493

Exhibit 1

Exhibit 2

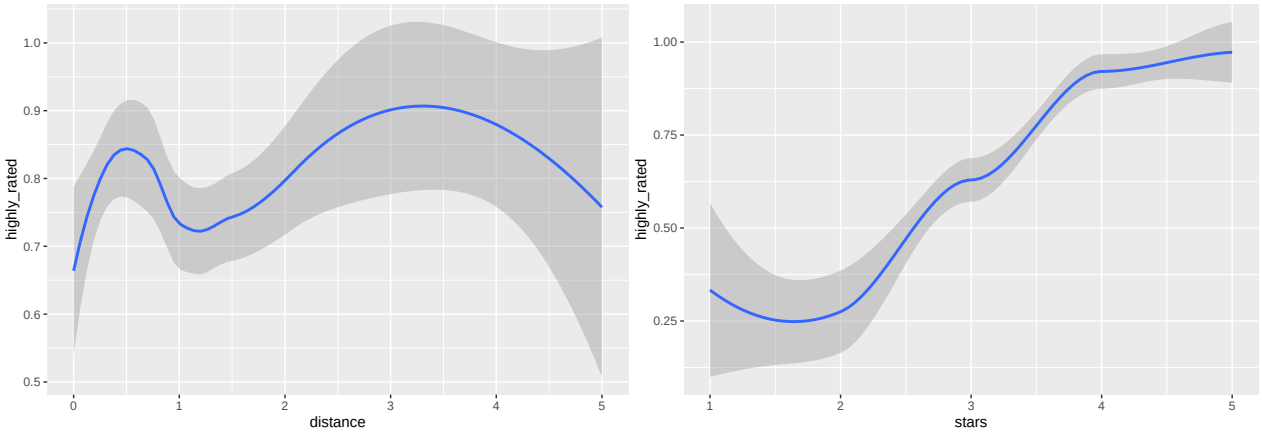

Table 2: Regression Model Summary

	lpm	logit	logit_marg	probit	probit_marg
Intercept	−1.245**	−12.122**		−7.078**	
	(0.220)	(1.963)		(1.089)	
stars (1-4)	0.209**	1.094**	0.129**	0.633**	0.130**
	(0.027)	(0.216)	(0.031)	(0.120)	(0.023)
stars (>=4)	−0.017	0.679	0.072	0.279	0.054
	(0.050)	(0.769)	(0.073)	(0.353)	(0.063)
distance (>=0.5)	−0.053	−0.089	−0.011	−0.013	−0.003
	(0.176)	(1.503)	(0.177)	(0.844)	(0.174)
distance (>0.5,<=1)	0.177	1.064	0.125	0.508	0.105
	(0.134)	(1.140)	(0.135)	(0.644)	(0.132)
distance (>1, =<3)	0.009	0.236	0.028	0.158	0.032
	(0.046)	(0.425)	(0.050)	(0.235)	(0.048)
distance (>3)	−0.064	−0.603	−0.071	−0.397	−0.082
	(0.070)	(0.628)	(0.074)	(0.349)	(0.072)
log(price)	0.248**	1.884**	0.222**	1.109**	0.228**
	(0.048)	(0.409)	(0.056)	(0.230)	(0.045)
Num.Obs.	493	493	493	493	493

* p < 0.05, ** p < 0.01

Exhibit 3

Exhibit 4