Data Analysis 2 : Assignment 2

Ali Hasnain Khan Sial (2101874) & Rauhan Nazir (2003231)

12/3/2021

Overview:

The goal of this assignment is to establish if there is a correlation between a binary variable, "highly_rated" and other explanatory variables, namely stars, distance, and log of price. The binary variable is designed to take a value of 1 if the rating is greater than 4 and 0 for all other instances. We have used 5 different regression models to try and explain this, however certain models perform better than others as they don't have limitations that other models do. We have explained this in detail later. Two different datasets were downloaded using OSF hotels-europe containing information about hotel features and prices. A single data table was later acquired by joining them on hotel_id.

Data Filtering & Adding Lsplines

We filtered the original data table and focused only on the hotels in Barcelona, having prices less than USD 600 in the month of November for the year 2017. We excluded the weekend and the missing values as well. Table 1 shows the summary of the final data that was used for regressions. Loess for each of the explanatory variables (with highly_rated), allowed us to identify the values we required to add lsplines on, as shown in Exhibit 2. Firstly, for stars, there was no need for it since there was no significant change in the general pattern, while for distance it was added at 1 and 2.5 and for log of price at 4.75

Estimated Models:

The 5 models that we used are Linear Probability Model (LPM), Logit, Probit, Logit Marginal Difference and Probit Marginal Difference. LPM model is the most basic one with a major limitation. There is no guarantee that the probability won't exceed 1, as proven in our case (1.19550). To overcome this limitation, we incorporated Logit and Probit models. They ensure that the probability is always between 0 and 1, as shown in the S curve in Exhibit 4, however, they only allow us to establish the direction of correlation, not the magnitude. Logit and Probit Marginal difference models allow us to overcome this. Both these models are the ones with the most meaningful interpretations and serve our purpose best.

Summary & Interpretations:

Exhibit 3 shows the results for all the regressions that were run. Both the Logit Marginal Difference and the Probit Marginal Difference models yield on average yield quite identical probabilities, and it was no different for our results. For instance, for the hotels having a log of price less than 4.5, the probability that they are going to be highly rated is 57.1% and 58.3% for Logit Marginal Difference and the Probit Marginal Difference model respectively, being significant at a confidence interval of 99% for both. Likewise, for hotels having a distance of greater than 2.5 miles from the city center, the probability that the hotel is highly rated is 48.3% and 47.6% for Logit Marginal Difference and Probit marginal difference model respectively (While it is significant for Probit at a confidence interval of 95%). Similar interpretations can be made for other explanatory variables as well.

Exhibit 1

Table 1: Data Summary Table

	mean	SD	Min	Max	Median	P95	N
highly_rated	0.73	0.44	0.00	1.00	1.00	1.00	345
distance	1.18	0.80	0.10	4.60	1.00	2.80	345
stars	3.50	0.97	1.00	5.00	4.00	5.00	345
lnprice	4.66	0.42	3.91	6.17	4.60	5.54	345

Exhibit 2

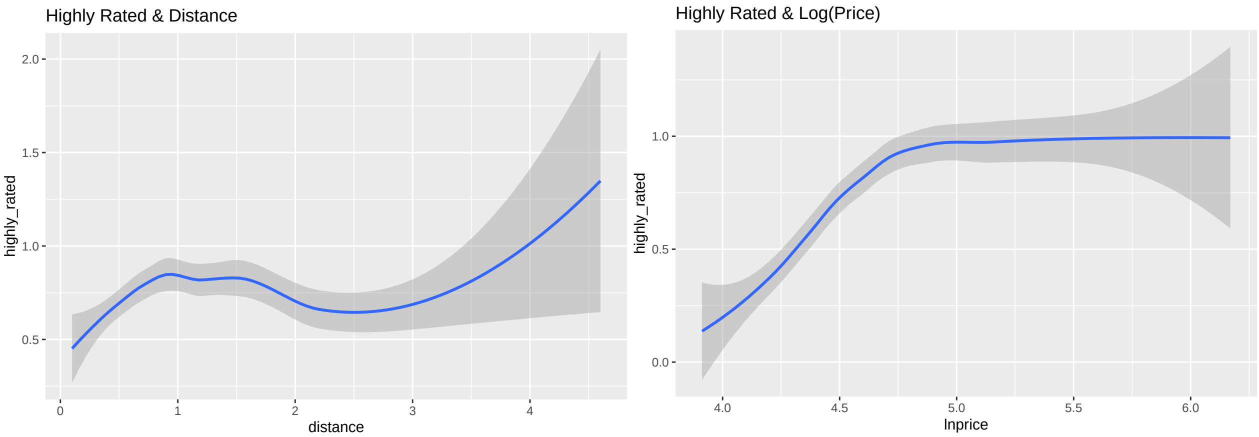


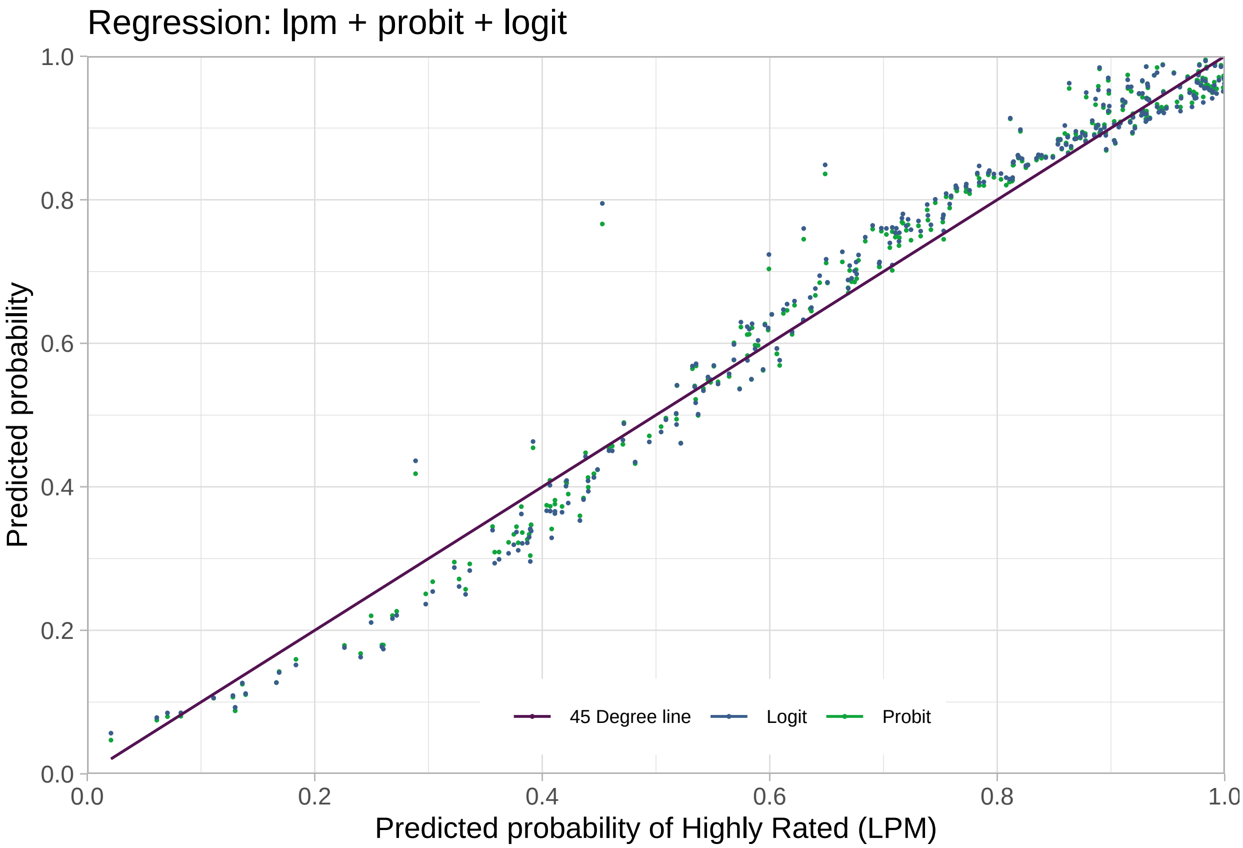
Exhibit 3

Table 2: Regression Model Summary

	lpm	logit	logit_marg	probit	probit_marg
Intercept	−3.362**	−20.946**		−12.271**	
	(0.482)	(3.651)		(2.035)	
stars	0.072*	0.413*	0.053	0.255*	0.058*
	(0.030)	(0.206)	(0.028)	(0.120)	(0.026)
distance (<1)	0.168	1.211	0.157	0.693	0.156
	(0.089)	(0.631)	(0.085)	(0.366)	(0.081)
distance (>=1, <2.5)	−0.085	−0.725	−0.094	−0.429	−0.097
	(0.052)	(0.392)	(0.053)	(0.225)	(0.050)
distance (>=2.5)	0.258*	3.736	0.483	2.110*	0.476*
	(0.129)	(1.914)	(0.257)	(1.067)	(0.237)
log(price) (<4.5)	0.825**	4.420**	0.571**	2.585**	0.583**
	(0.118)	(0.869)	(0.143)	(0.489)	(0.096)
log(price) (>=4.5)	−0.062	1.278	0.165	0.475	0.107
	(0.093)	(1.443)	(0.186)	(0.646)	(0.146)
Num.Obs.	345	345	345	345	345

* p < 0.05, ** p < 0.01

Exhibit 4


Regression: lpm + probit + logit
Predicted probability
1.0
0.8
0.6
0.4
0.2
0.0
45 Degree line
Logit
Probit
0.0
0.2
0.4
0.6
0.8
1.0
Predicted probability of Highly Rated (LPM)