Data Analysis 2: Assignment 2

Muhammad Arbash Malik - 2202071

Introduction:

The goal of this assignment is to regress a binary variable, "highly_rated" with other variables. The binary variable is designed to take a value of 1 if the rating is greater than 4 and 0 for all other instances. I 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. Explanations for the models will be later in the report. 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 using outer condition.

1. Data Manipulation

I filtered the original data table and focused only on the hotels in Barcelona, having prices less than EUR 550 as well as less than distance of 5 km from the city center. I also excluded any missing values from the dataset as well. Table 1 shows the summary of the final data that was used for regressions. Columns for log of price and log of distance were also created to check which variable would be better in regressions. I chose **`lnprice`** in addition to **`stars`** and **`distance`** because of better distributions. Loess graph was made for each of the explanatory variables with **`highly_rated`**, allowed us to identify the values we required to add splines 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 2.25 and for log of price at 4.75.

2. Models Descriptions:

The 5 models that I used are Linear Probability Model (LPM), Logit Model, Probit Model, Logit Marginal Effects and Probit Marginal Effects. The Linear Probability Model (LPM) model is the one with a major limitation. There is no restriction on LPM so that it doesn't generate predicted probability greater than 1, as proven in our case (1.104). To overcome this limitation, I incorporated Logit and Probit models. They ensure that the probability is always between 0 and 1, as shown in the S curve in Exhibit 3. Both the Logit and Probit model predictions, when plotted, are close to the 45 Degree line, except at the tails and less extreme than LPM. We can also see both Logit and Probit are almost indistinguishable from Exhibit 3. 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.

3. Summary & Interpretations:

Exhibit 3 shows the results for all the regressions that were run (LPM on X axis, Logit and Probit on Y Axis.). Both probit and logit have predicted probabilities (of highly rated) that range between 0.018 and 0.97 (narrower than for LPM, range -0.12 to 1.1, thanks to the nonlinear form). You can also see that LPM estimates, and Logit & Probit Marginal Effects are very similar to each other. For example, if we take the variable **`stars`** the estimates are 16%, 14%, 14.5% for LPM, Logit marginal, and Probit marginal respectively. Both the Logit Marginal Effects and the Probit Marginal Effects models yield on average yield quite identical probabilities, and in my case, this was the situation as well.

For instance, hotels having a distance less than 1 km from the city center, the probability that the hotel is highly rated is 22.1% and 21.5% for Logit Marginal Effects and Probit Marginal Effects model respectively, being significant at a confidence interval of 99% for both). Hotels having a distance greater than 2.5 km from the city center, the probability that the hotel is highly rated is 8.7% and 9.2% for Logit Marginal Effects and Probit Marginal Effects model respectively, being significant at a confidence interval of 99% for both).

Hotels having a log of price less than 4.5, the probability that they are going to be highly rated is 32.3% and 33.6% for Logit Marginal Effects and the Probit Marginal Effects model respectively, being significant at a confidence interval of 99% for both.

Exhibit 1: Table 1 – Describe result for chosen variables:

	Observations	Mean	Standard Deviation	Minimum	25th Percentile	Median	75th Percentile	Maximum
distance	2846.0	1.31	0.90	0.10	0.60	1.10	1.80	4.70
stars	2846.0	3.56	0.94	1.00	3.00	4.00	4.00	5.00
lnprice	2846.0	5.10	0.54	3.87	4.64	5.08	5.49	6.31

Exhibit 2: Loess graphs for splines:

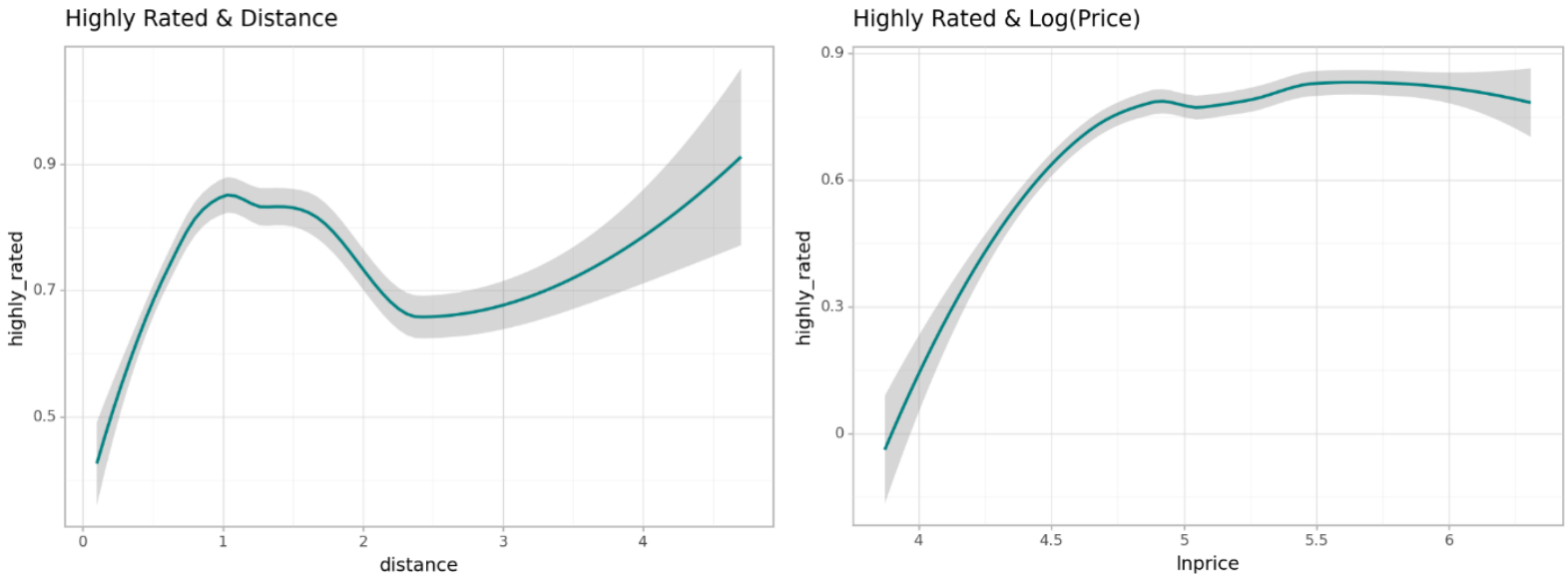

Exhibit 3: Logit and Probit graph:

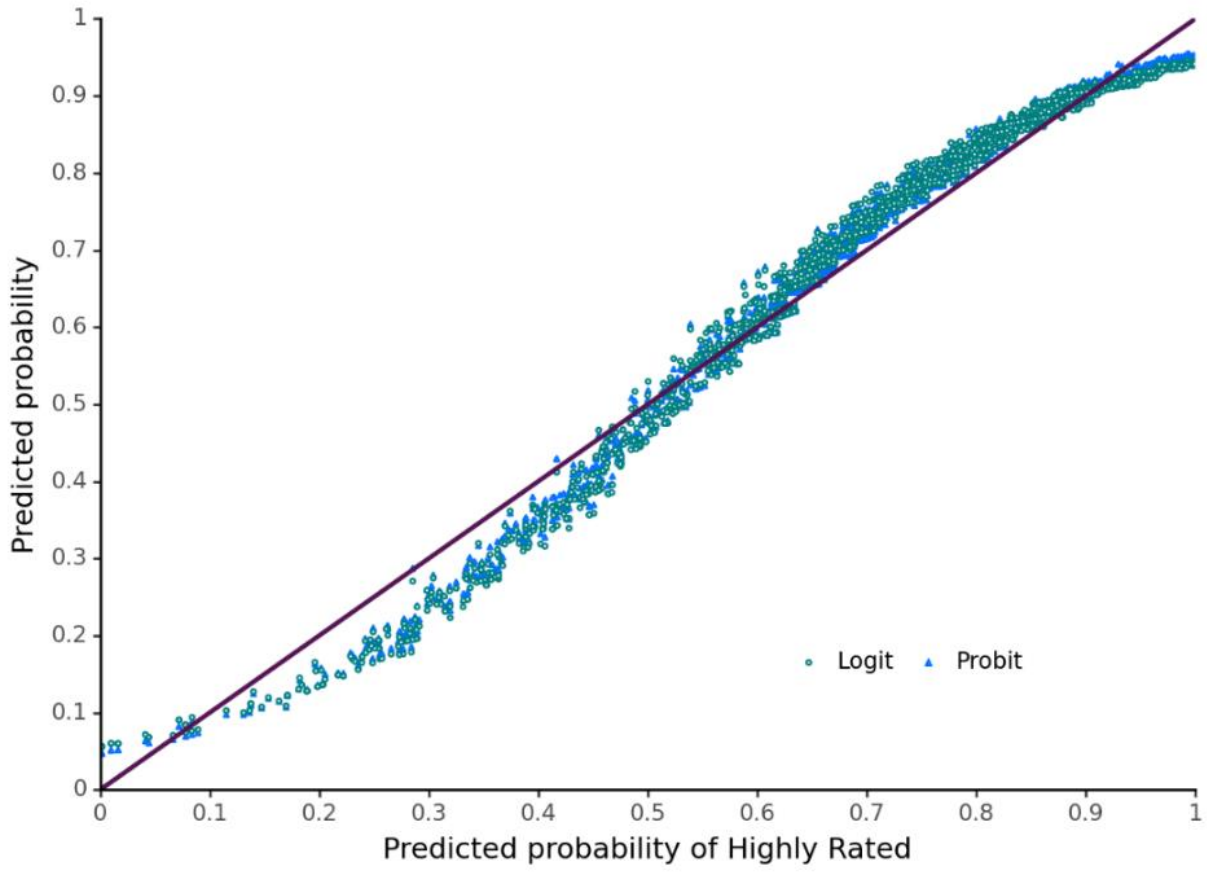

Exhibit 4: Table 2 – LPM, Logit and Probit Summary

	LPM (0)	Logit (1)	Probit (2)
Intercept	-2.07***	-13.13***	-7.85***
	(0.24)	(1.51)	(0.88)
R-squared	0.22		
R-squared Adj.	0.22		
lspline(distance, [1, 2.25])[0]	0.24***	1.49***	0.84***
	(0.04)	(0.23)	(0.13)
lspline(distance, [1, 2.25])[1]	-0.14***	-0.96***	-0.55***
	(0.02)	(0.14)	(0.08)
lspline(distance, [1, 2.25])[2]	0.09***	0.59***	0.36***
	(0.03)	(0.17)	(0.10)
lspline(lnprice, [4.75])[0]	0.45***	2.18***	1.31***
	(0.05)	(0.33)	(0.19)
lspline(lnprice, [4.75])[1]	-0.04*	-0.08	-0.04
	(0.02)	(0.13)	(0.08)
stars	0.16***	0.95***	0.57***
	(0.01)	(0.06)	(0.03)
Observations	2846	2846	2846

Standard errors in parentheses.
* p<.1, ** p<.05, ***p<.01

Exhibit 4: Table 3 – Logit Marginal Difference and Probit Marginal Effects

Variables	LMD dy/dx	LMD std err	LMD P>|z|	PMD dy/dx	PMD std err	PMD P>|z|
lspline(distance, [1, 2.25])[0]	0.2211	0.033	0.000	0.2159	0.033	0.000
lspline(distance, [1, 2.25])[1]	-0.1430	0.021	0.000	-0.1410	0.021	0.000
lspline(distance, [1, 2.25])[2]	0.0870	0.026	0.001	0.0927	0.026	0.000
lspline(lnprice, [4.75])[0]	0.3231	0.048	0.000	0.3367	0.049	0.000
lspline(lnprice, [4.75])[1]	-0.0116	0.020	0.555	-0.0116	0.019	0.552
stars	0.1416	0.007	0.000	0.1454	0.008	0.000