Data Analysis - Assignment 2

Nawal Zehra Hasan

Introduction

For this assignment we were given the Hotels Europe Dataset **Hotel Price dataset** with different features about the hotels that can be found here **Hotel Features dataset**. I wanted to understand how hotel ratings in ***Amsterdam*** are associated with different features of the hotel such as distance and stars as the explanatory variables.

Data Analysis

The table summary reveals that of the total of 667 observations 414 are highly rate(>4). The descriptive statistic tell us that of of the total of 667 observations 62% of the hotels are highly-rated. These highly rated hotels are on average 1.58 miles away from the city center with 3.52 stars on average. We can also see that the average price of a highly rated hotel in Amsterdam was approximately 267 Euros.

Interpretations and Analysis

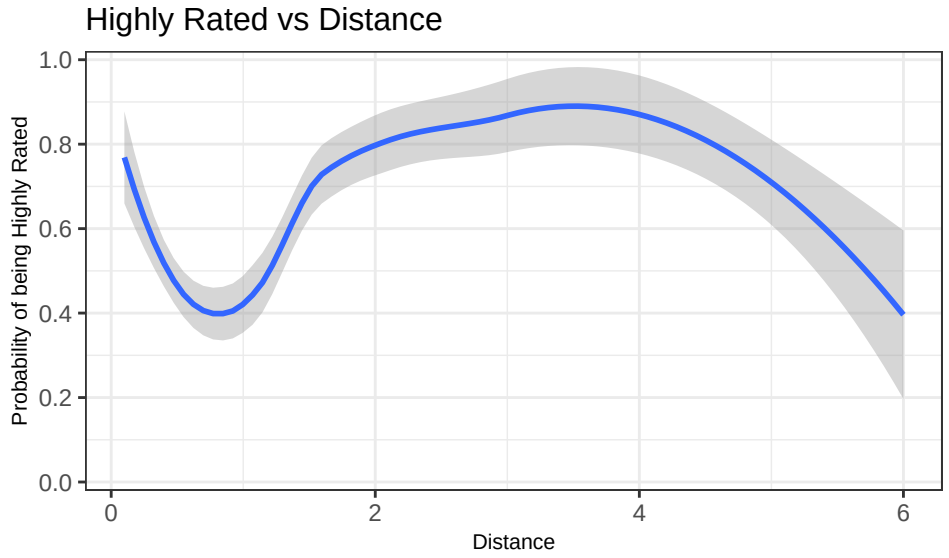
I ran an LPM model where we regressed highly rated distance, stars, and the log price. By checking the lowess regression I added splines to distance at 0.75 and 3.5. The coefficients of LPM gave us some significant coefficients. However, when we looked at the predicted probabilities of the model, there were values of above 1, which cannot be considered as in case of probabilities. Hence, we decided to run logit and probit models to limit our predicted models between 0 and 1. As expected, the predicted probabilities were between 0 and 1. Since we cannot interpret the coefficients of these 2 models we estimate the probit and logit regressions to calculate the corresponding marginal effects. These enable interpretation of the resulting coefficients, similar to an LPM model.For distance in LPM model, for hotels in the distance to city center of less than 0.75 miles, if a hotel is one mile farther away, I expect them to be rated 14.4 % less likely to be highly rated on average. For the distance between 0.75 miles and 3.5 miles, a hotel one mile farther is expected to be 24.5% more likely to be highly rated. For hotels with a distancew of greater than 3.5 miles from the city center, every one mile farther the hotel is, the hotel is on average 24.9% less likely to be highly rated. For stars in LMP model, if a hotel has one more star we expect it to be 25.9% more likely to be highly rated on average. This is quite a significant percentage. This show that every added star contributes to the ratings of the hotels. With respect to the log(price) variable, the coefficients are significant at 99.9% across the models. As per the LPM, the probability of a hotel being highly rated is 18.7% as the price is higher by 1%. The probit and logit marginal models give out similar coefficients, hence their interpretation is same.The results show that hotels that are more than one mile away from the city center higher likelihood of being rated better, possibly because the hotels within the one mile range of the city center tend to be more expense, hence they have fewer guests, resulting in incomplete information regarding those hotels since number of guests are less compared to in hotels that are farther away. Also, higher priced hotel is better rated on average. This can be due better customer service, better facilities as the price charged for such hotels is fairly high. When it comes to stars, it is quite surprising to see the contribution one added star has to ratings of a hotel. Therefore, we can conclude with much certainty that there tends to be a positive association between number of stars and the probability of being highly rated. Nonetheless, we should be aware of the problems with generalizing such results with respect to external validity.

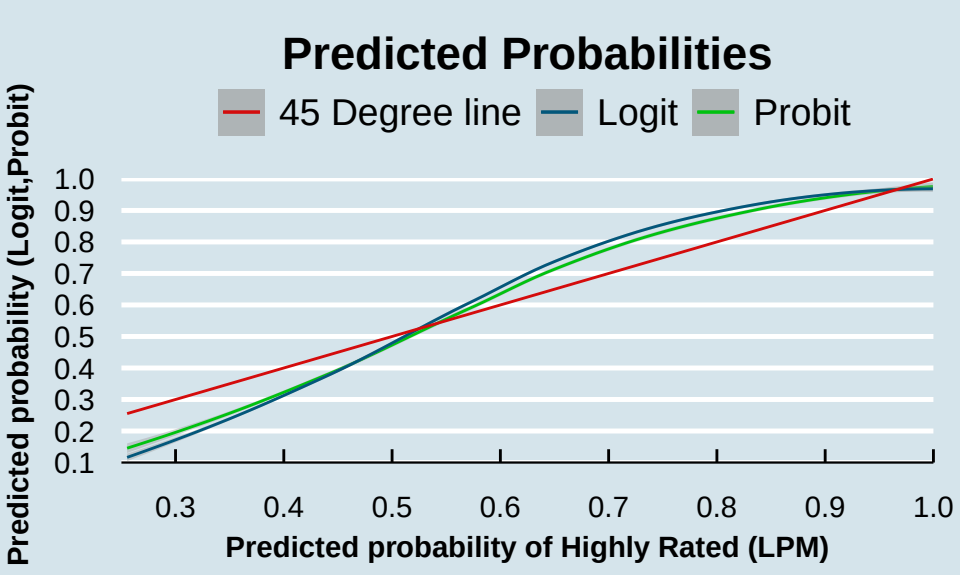
Table 1: Summary Statistics

	Mean	SD	Min	Max	Median	P95	N
highly_rated	0.62	0.49	0.00	1.00	1.00	1.00	667
distance	1.58	1.41	0.10	6.00	1.10	4.90	667
stars	3.52	0.97	1.00	5.00	3.50	5.00	667
price	267.40	157.29	77.00	1259.00	222.00	551.00	667
log_price	5.46	0.49	4.34	7.14	5.40	6.31	667

	LPM	logit	logit Marg	Probit	Probit Marg
Intercept	−1.384**	−14.889**		−8.341**	
	(0.188)	(1.895)		(1.038)	
distance (< 0.75)	−0.144	−0.552	−0.065	−0.273	−0.058
	(0.078)	(0.719)	(0.070)	(0.405)	(0.085)
distance (>=0.75, <3.5)	0.247**	1.559**	0.183**	0.807**	0.171**
	(0.025)	(0.225)	(0.039)	(0.119)	(0.022)
distance (>=3.5)	−0.249**	−1.668**	−0.195**	−0.862**	−0.182**
	(0.043)	(0.293)	(0.047)	(0.161)	(0.032)
stars	0.259**	2.119**	0.248**	1.199**	0.254**
	(0.012)	(0.206)	(0.036)	(0.108)	(0.016)
log_price	0.187**	1.390**	0.163**	0.771**	0.163**
	(0.035)	(0.319)	(0.042)	(0.178)	(0.036)
Num.Obs.	667	667	667	667	667

* p < 0.05, ** p < 0.01




Predicted Probabilities
45 Degree line
Logit
Probit
Predicted probability (Logit,Probit)
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Predicted probability of Highly Rated (LPM)