DA2 - Assignment 2

Ersan Kucukoglu

04 December 2021

Introduction

In this assignment, I used the hotels-europe dataset and filtered on a city which is Barcelona. The accommodation type was fixed as 'hotel' and missing values for our variables of interest were filtered out. The filtered data contains 352 observations.I used the hotel user rating to create a binary variable: highly_rated=1 if rating is greater than and equal to 4, 0 otherwise. I examined how high rating is related to the other hotel features in the data. I estimate linear probability, logit, and probit models with distance and stars as explanatory variables.

Descriptive Statistics

Based on descriptive statistics for the Barcelona sample data. The table above indicates that 73% of the hotels in Barcelona are highly rated. The distribution of the stars is left skewed with an average of 3.49. For the distance, it shows right skewed distribution with an average of 1.18 km.

It can be observed from the graph (Relationship between high rating and stars, distance features)the binary highly_rated variable is related to distance and stars. I used lowess curves to plot distance and stars against the highly rated variable.

Regressions and summary of the models

Linear probability model: The fact that the hotel is 1 km from the city center shows that it has a 2 percent smaller probability of being highly rated, keeping all other variables constant.This is not very significant.On average for the hotels in the stars range from 1 to 3, for one more star, it is 23.6% more likely to get a high rate. When the stars from 3 to 5, it is 18.6% more likely to get a high rate, and results are significant at a 1% significance level. For the logit and probit models' mean marginal differences, in terms of distance, we can observe from the summary of models table that they are very similar to the LMP model. On average, for one more star, the logit marginal coefficient for stars, from 1 to 3, indicates that a hotel's likelihood of being highly rated is 15,4% greater, and the probit marginal coefficient it is being highly rated is 16,4 % greater. The results are significant at a 1% significance level. For 3 stars to 5 stars, logit marginal shows that it is the likelihood of being highly rated is 20.3% greater, and the probit marginal coefficient it is being highly rated is 20,1 % greater. The results are significant at 1% significance level. The linear probability model predictions and the logit and probit models have similar predictions. It can be observed that the logit and probit have identical brier scores and almost the same values for pseudo R2 and log-loss. The predicted probability for the three models is shown in the model comparison graph.

Compared the LPM, logit and probit models, the baseline is the predictions of the LPM that correspond to the 45-degree line. The predicted probabilities from the logit and probit are very close to each other. The range of predicted values of the logit model is [0.162, 0.985], while for the probit it is [0.159,0.986].

Appendix

Table 1: Descriptive Statistics Table

	mean	Median	Min	Max	P95	SD	N
highly_rated	0.73	1.00	0.00	1.00	1.00	0.45	352
distance	1.18	1.00	0.10	4.60	2.80	0.80	352
stars	3.49	4.00	1.00	5.00	5.00	0.98	352

Examine how high rating is related to stars and distance

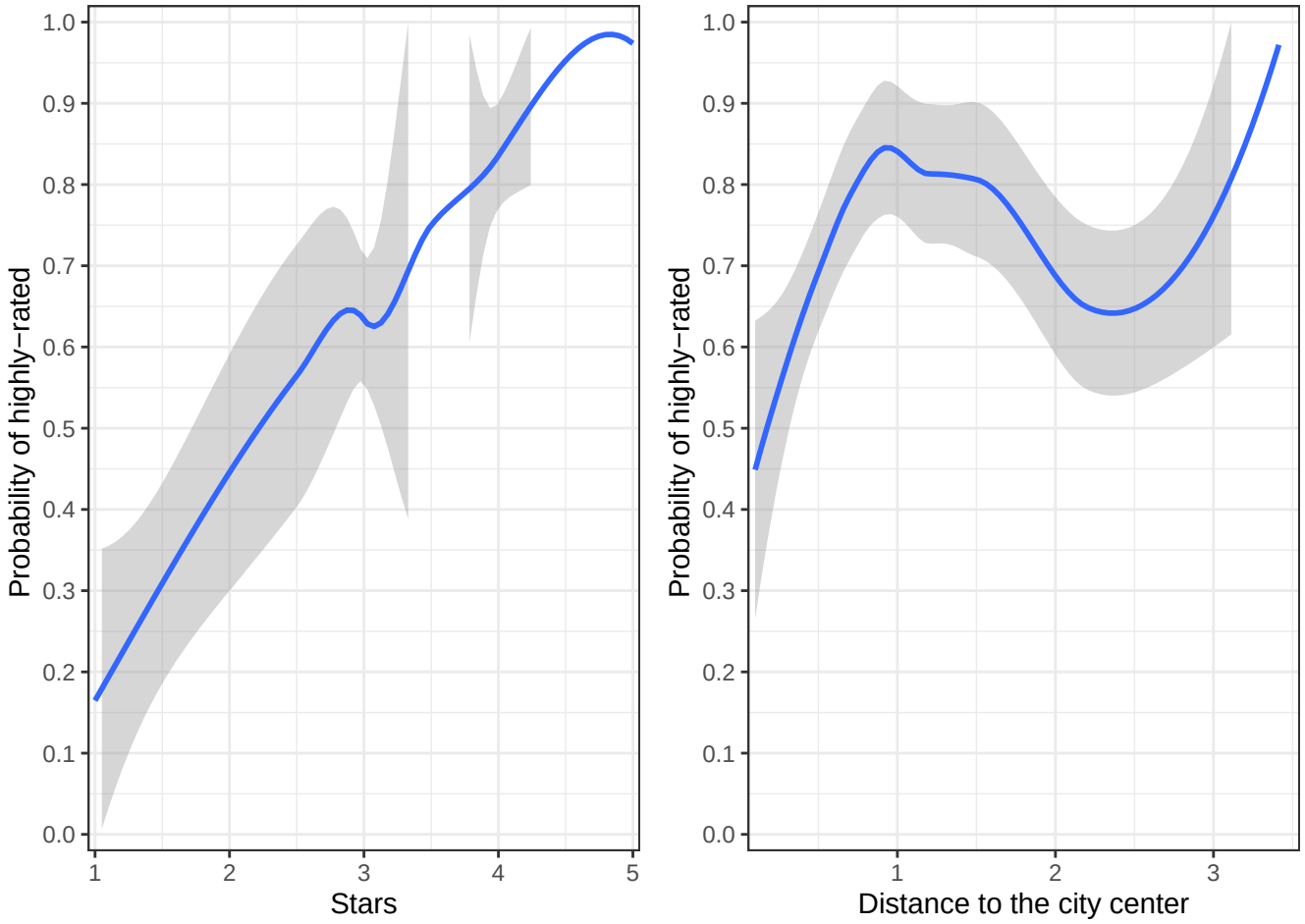
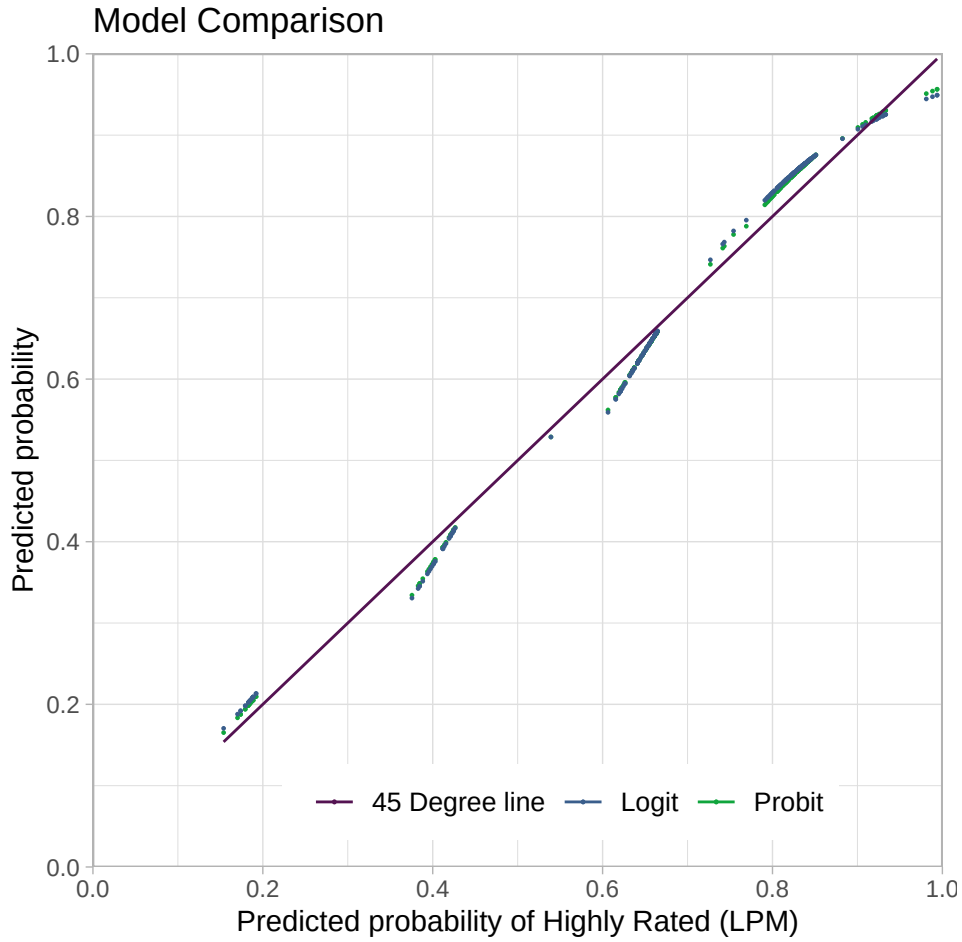


Table 2: Regression Models

	LPM	Logit	Logit_marginal	Probit	Probit_marginal
Constant	−0.042	−2.274**		−1.408**	
	(0.128)	(0.734)		(0.433)	
distance	−0.018	−0.132	−0.021	−0.079	−0.021
	(0.028)	(0.170)	(0.027)	(0.099)	(0.026)
lspline(stars, c(3))1	0.236**	0.982**	0.154**	0.608**	0.164**
	(0.048)	(0.269)	(0.048)	(0.159)	(0.038)
lspline(stars, c(3))2	0.186**	1.292**	0.203**	0.746**	0.201**
	(0.032)	(0.262)	(0.049)	(0.145)	(0.034)
Num.Obs.	352	352	352	352	352
Std.Errors	Heteroskedasticity-robust	IID		IID	

* p < 0.05, ** p < 0.01



*Goodness of fit

lpm	**logit**	**probit**
highly_rated	highly_rated	highly_rated
-0.0424 (0.1276)	-2.274** (0.7343)	-1.408** (0.4328)
-0.0182 (0.0279)	-0.1317 (0.1695)	-0.0786 (0.0985)
0.2363*** (0.0481)	0.9822*** (0.2692)	0.6079*** (0.1593)
0.1864*** (0.0322)	1.292*** (0.2618)	0.7456*** (0.1452)
OLS	Logit	Probit
Heteroskedas.-rob.	IID	IID
0.20339	–	–
0.15800	0.15804	0.15796
0.18637	0.17633	0.17750
NaN	-0.48263	-0.48195