Data Analysis : Assignment 2

Ghazal Ayobi and Shah Ali Gardezi

Introduction

The Question of this case study is how hotels stars is related to highly rated hotels. For this assignment we use Hotels-Europe data. This data set contains two tables **Features** and **Price**. We joined the two tables using left join.

Data Transformation

As a process of filtering and data transformation, we use hotel user rating as the dependent variables and transformed it to a binary variable called *highly_rated* which equals to one if *rating* is more than 4, 0 otherwise. We transformed stars to a binary variables called *top_stars* which equals to one if *stars* is more than 4, and 0 otherwise. We examined *Lowess* regression with highly rated hotels and distance. Looking at kinks from Figure 1, we decided to put two knots at 1.2 and 3 miles. Other control variables log of Price, *lnprince*, and *weekend* (binary variable). We selected ***Paris*** City and considered ***Hotels*** as accommodation type. Moreover, we excluded hotels with less than *USD 600* per night, and we removed null and duplicated values from the data set. We are interested to estimate the probability of highly rated hotels on hotels having top stars(4, 5 stars) and other explanatory variables such as distance, log of price and weekend.

Anaalysis

The summary table shows us that mean of highly_rated lies above 0.5 indicating the presence of more highly rated hotels in the dataset. Table 2 shows six regression models: lpm0 lpm, logit, marginal logit, probit and marginal probit. Model 1, *lpm0* indicates that Top stars hotels are 43.1 percentage points are more likely to be highly rated. The 95% confidence interval around the slope parameter is [0.413, 0.449] which implies that we can be 95% confident that top stars in the hotels-europe are highly rated. As the distance from city center increases by one unit within 0 - 1.2 mile the probability of highly rated hotels decreases by 10.8%. Interestingly, the increase of one unit distance between the 1.2-3 mile distance has no effect on the probability of highly rated hotel. However, this probability tends to decreases by 24.1% by the increase of distance by one unit beyond 3 miles.

By looking at the logit and probit estimates for our model, the probability of highly rated to top stars, distance, price and weekend are same as linear model. By looking at the column 3 and 4, the Logit Coefficients are almost five times the size of corresponding logit marginal differences. Furthermore, in the column 5 and 6, probit coefficient is almost three times the size of corresponding probit marginal differences. It is interesting to observe that the two marginal differences, logit and probit, are the same with LPM coefficients in column 2. Thus, we will be interpreting the coefficients of marginals differences of both logit and probit models. Figure 2 helps visualize the findings of three models with predicted probabilities of logit and probit on y axis and predicted probability of LPM in the x axis. 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. To generalize the result we can say that top stars hotels have a around 30% higher probability to be highly rated with other variables (distance, price and weekend) being the same. In order to compare logit and probit models we calculated the Pseudo R2 and found that both models have the same Pseudo R2.

Table 1: Summary Statistics

	Mean	SD	Min	Max	Median	P95	N
highly_rated	0.56	0.50	0.00	1.00	1.00	1.00	11397
distance	1.62	0.78	0.10	4.20	1.60	2.90	11397
stars	3.22	0.78	1.00	5.00	3.00	4.00	11397

Table 2: The Probability of Highly rated hotels and top stars Hotels- LPM, Logit, and Probit models

	(1) LPM0	(2)LPM	(3) logit coeffs	(4) logit Marg	(5) Probit	(6) Probit Marg
Constant	0.527**	−0.984**	−7.516**		−4.521**	
	(0.022)	(0.058)	(0.320)		(0.188)	
top_stars	0.431**	0.315**	1.598**	0.325**	0.962**	0.326**
	(0.009)	(0.010)	(0.054)	(0.011)	(0.031)	(0.010)
lspline(distance, c(1.2, 3))1	−0.108**	−0.040	−0.218	−0.041	−0.135*	−0.043*
	(0.021)	(0.021)	(0.112)	(0.021)	(0.067)	(0.021)
lspline(distance, c(1.2, 3))2	0.008	0.022*	0.101*	0.019*	0.055*	0.017*
	(0.009)	(0.009)	(0.045)	(0.009)	(0.027)	(0.009)
lspline(distance, c(1.2, 3))3	−0.241**	−0.243**	−1.272**	−0.240**	−0.742**	−0.234**
	(0.050)	(0.049)	(0.274)	(0.045)	(0.159)	(0.042)
lnprice		0.274**	1.394**	0.263**	0.839**	0.264**
		(0.010)	(0.054)	(0.012)	(0.032)	(0.010)
weekend		0.093**	0.533**	0.100**	0.316**	0.099**
		(0.010)	(0.051)	(0.010)	(0.030)	(0.010)
Num.Obs.	11 397	11 397	11 397	11 397	11 397	11 397

* $p < 0.05$, ** $p < 0.01$

Figure 1 : Probability of Highly Rated vs Distanc

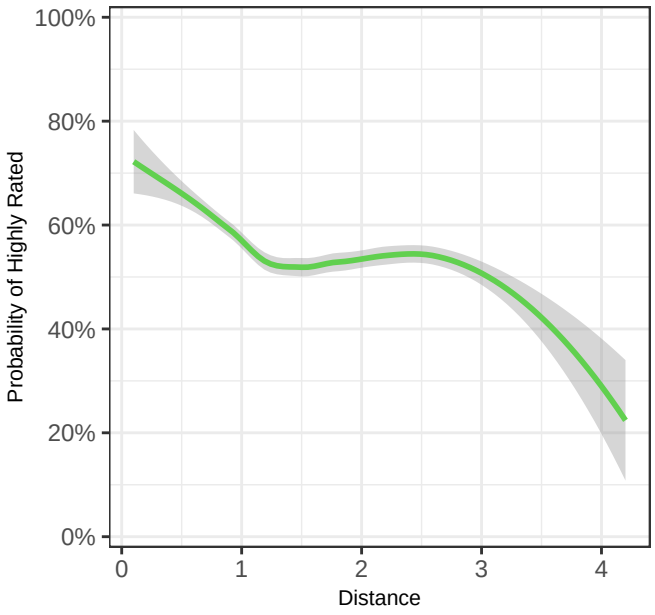
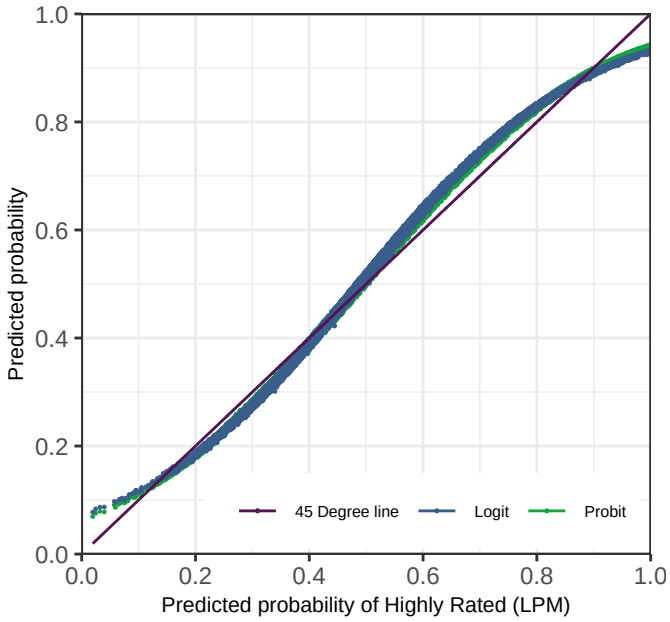


Figure 2 : Predicted Probability of LPM, Logit and



Appendix

- Filtered the data for **Paris**
- Selected Hotels and Apartments for accommodation type
- Filtered for Price is less than 600
- Removed duplicates
- Removed Null values from *rating*, *stars*, and *distance*
- Created log of price, *lnprice*
- Created *highly_rated* if rating >= 4
- Created *top_stars* if stars >= 4