Assignment-2

- Shahana Ayobi - Mohammad Ahmed

2022-12-02

Github: https://github.com/ayobishahana/Data-Analysis-2/tree/main/Assignment-2

Introduction

In this study, we are interested to examine how highly rated hotels are related to distance and stars in Paris. We combined the `hotels-price` and `hotels-features` datasets using left join for our analysis.

Data Cleaning

To continue with our analysis, we first filtered the `city_actual` variable to Paris and `accommmodation type` to Hotels. As instructed, we created a new binary variable called `highly_rated` which takes the value 1 for hotels with higher or equal to 4 rating and takes 0 for lower ratings. We also created another binary variable for `stars` called `top_stars` which takes the value 1 for hotels with higher or equal to 4 stars and takes 0 for lower star values. This would enable us to interpret the coefficients clearer. We also removed missing values to reduce bias in the analysis. For `distance`, we decided to use splines at 1.2 and 2.5 miles after looking at the kinks in graph shown in `Figure 1` allowing slope to differ for different values of distance.

Analysis

The summary table indicates that the mean of highly rated hotels is 0.57, showing that the dataset contains more highly rated hotels. `Table 1` LPM, Logit, and Probit regression models. The LPM results show that hotels with top stars are more likely to be highly rated by 42.9 percentage points(pp). Additionally, for hotels with less than 1.2 miles distance from the center, a one mile increase in distance decreases the probability of being highly rated by 11.3 pp, and for hotels with higher than 2.5 miles distance, this probability decreases by 11.5 pp. Looking at Logit and Probit models in `Table 1`, the signs and significance of coefficients are similar to LPM. To be able to interpret, we use marginal effects for the two models shown in `Table 2`. For instance, the probability of a hotel being highly rated increases by 43 pp (both Logit and Probit models) for the hotels with top stars. The different splines for variable distance also have quite similar marginal effects for Logit and Probit models with the LPM coefficients. `Figure 2` depicts the results of three models, with logit and probit predicted probabilities on the y axis and LPM predicted probabilities on the x axis. The S-shaped curve close to the 45 degree line indicates that logit and probit are very similar to each other and very close to LPM. To compare the logit and probit models, we calculated the Pseudo R2 and discovered that they both have the same Pseudo R2.

.

.

Table 1: Summary Statistics

	Mean	SD	Min	Max	Median	P95	N
highly_rated	0.57	0.49	0.00	1.00	1.00	1.00	12035
distance	1.61	0.78	0.10	4.20	1.50	2.90	12035
stars	3.25	0.79	1.00	5.00	3.00	5.00	12035

Table 2: The Probability of Highly Rated Hotels- LPM, Logit, and Probit models

	(1)LPM	(2) Logit	(5) Probit
Constant	0.540**	0.305**	0.160*
	(0.021)	(0.107)	(0.064)
Top Stars	0.429**	2.081**	1.249**
	(0.009)	(0.050)	(0.028)
Distance Spline <1.2 miles	−0.113**	−0.595**	−0.334**
	(0.021)	(0.108)	(0.065)
Distance Spline 1.2-2.5 miles	0.017	0.088	0.033
	(0.011)	(0.053)	(0.032)
Distance Spline 2.5< miles	−0.115**	−0.561**	−0.326**
	(0.025)	(0.126)	(0.075)
Num.Obs.	12 035	12 035	12 035
R2	0.180		
R2 Adj.	0.180		
RMSE	0.45	0.45	0.45

* p < 0.05, ** p < 0.01

Table 3: The Probability of Highly Rated Hotels- Logit, and Probit Marginal Effects

	(3) logit Marg	(4) Probit Marg
Top Stars	0.430**	0.430**
	0.430**	1.249**
	2.081**	0.430**
	2.081**	1.249**
	(0.008)	(0.008)
	(0.008)	(0.028)
	(0.050)	(0.008)
	(0.050)	(0.028)
Distance Spline <1.2 miles	−0.119**	−0.111**
	−0.119**	−0.334**
	−0.595**	−0.111**
	−0.595**	−0.334**
	(0.022)	(0.022)
	(0.022)	(0.065)
	(0.108)	(0.022)
	(0.108)	(0.065)
Distance Spline 1.2-2.5 miles	0.018	0.011
	0.018	0.033
	0.088	0.011
	0.088	0.033
	(0.011)	(0.011)
	(0.011)	(0.032)
	(0.053)	(0.011)
	(0.053)	(0.032)
Distance Spline 2.5< miles	−0.112**	−0.108**
	−0.112**	−0.326**
	−0.561**	−0.108**
	−0.561**	−0.326**
	(0.024)	(0.023)
	(0.024)	(0.075)
	(0.126)	(0.023)
	(0.126)	(0.075)
Constant	0.305**	0.160*
	(0.107)	(0.064)
Num.Obs.	12 035	12 035

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

Table 4: Logit, Probit with Pseudo R2

	(2) Logit	(3) Probit
Constant	0.305**	0.160*
	(0.107)	(0.064)
Top Stars	2.081**	1.249**
	(0.050)	(0.028)
Distance Spline <1.2 miles	−0.595**	−0.334**
	(0.108)	(0.065)
Distance Spline 1.2-2.5 miles	0.088	0.033
	(0.053)	(0.032)
Distance Spline 2.5< miles	−0.561**	−0.326**
	(0.126)	(0.075)
Num.Obs.	12 035	12 035
RMSE	0.45	0.45
PseudoR2	0.144	0.144

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

Figure 1 : Probability of Highly Rated vs Distance

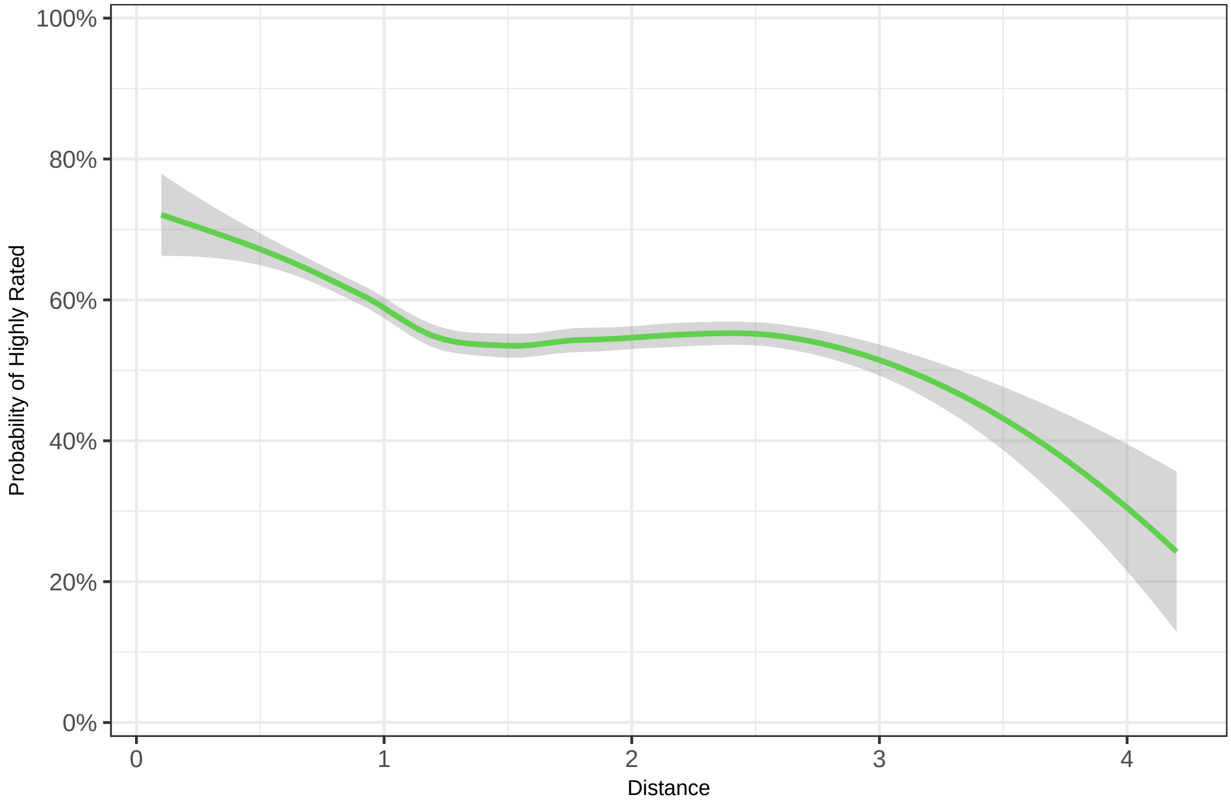
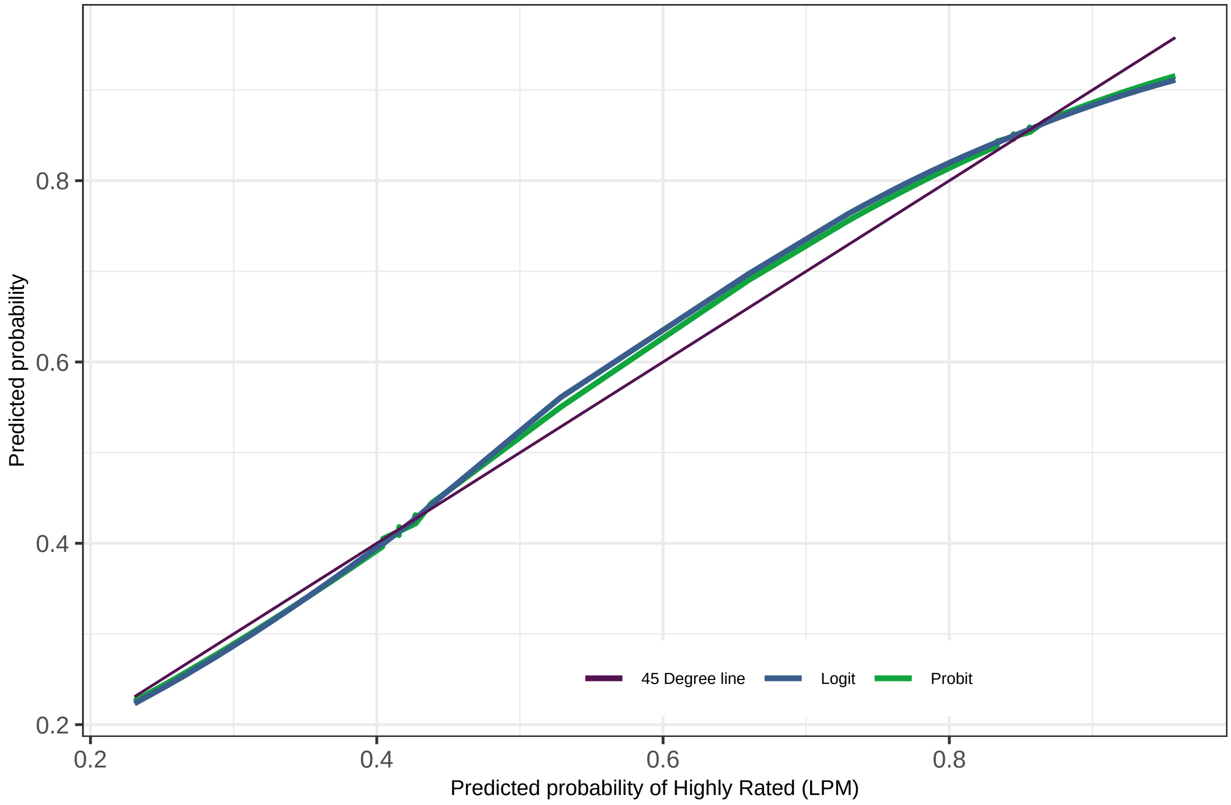


Figure 2 : Predicted Probability of LPM, Logit and Probit Models


Predicted probability
Predicted probability of Highly Rated (LPM)
45 Degree line
Logit
Probit