Máté Schieszler
Péter Szilvási

Data Analysis 2- Assignment 2

GitHub: https://github.com/mateschieszler/DA2_2023

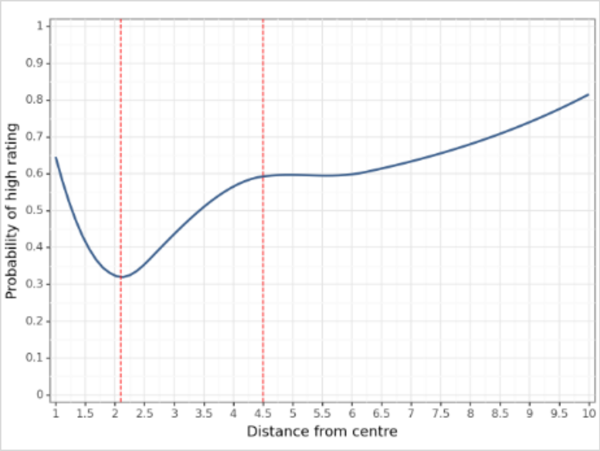
Chosen city: Milan

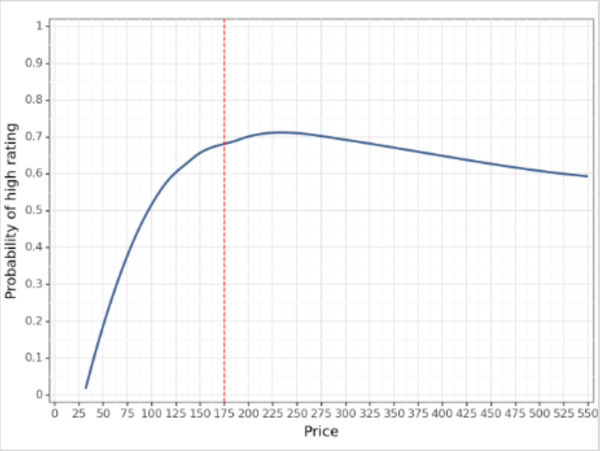
1. Data Cleaning and Filtering:

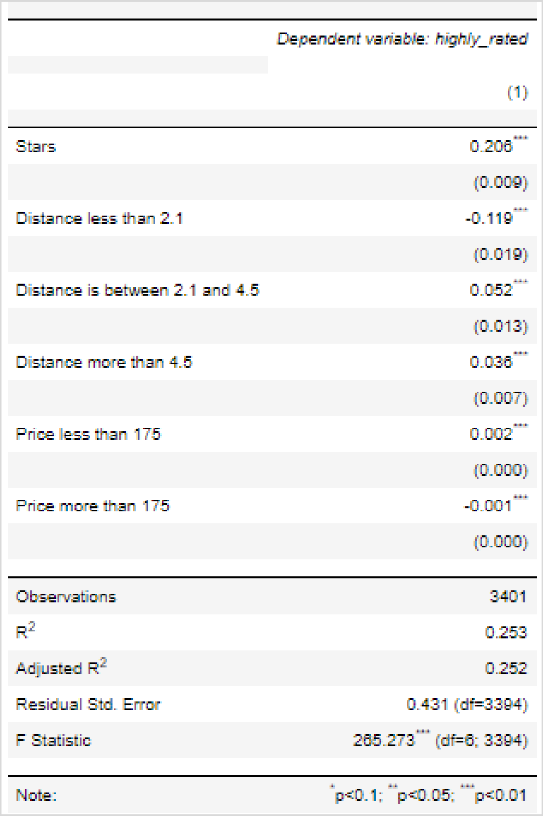
We dropped null values for rating column as it would lead to a `highly_rated` = 0. We also dropped null values for stars column as we have enough number of observations with 537 accommodations. For prices, to filter out extreme values we dropped values that were beyond the 550 EUR/night border (95th percentile = 542.95; max = 8617). While for distance we dropped extremes beyond 10kms as we are only looking at Milan, not including smaller cities around it.

2. LPM MODEL & PREDICTION

In our Linear Probability Model, we examined how high rating is related to number of stars, the hotel's distance from the city centre, and its price. After examining lowess regressions for price and distance we decided to include the explanatory variables as piecewise linear splines with knots at 2.1 and 4.5 for distance and at 175 euros/night for price.







	Dependent variable: highly_rated
	(1)
Stars	0.206^{***}
	(0.009)
Distance less than 2.1	-0.119^{***}
	(0.019)
Distance is between 2.1 and 4.5	0.052^{***}
	(0.013)
Distance more than 4.5	0.036^{***}
	(0.007)
Price less than 175	0.002^{***}
	(0.000)
Price more than 175	-0.001^{***}
	(0.000)
Observations	3401
R^2	0.253
Adjusted R^2	0.252
Residual Std. Error	0.431 (df=3394)
F Statistic	265.273^{***} (df=6; 3394)
Note:	$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

This model shows us that in Milan, accommodations with similar features with a higher star rating are significantly 20.6pp more likely to be highly rated, when controlling for distance from the city center and price category. The probability of a high rating decreases initially with greater distance from the center (-11.9pp) but then reverses (5.2pp for distances between 2.1 and 4.5 kms and 3.6pp after 4.5 kms), indicating a nuanced location-rating relationship. Additionally, while the likelihood of a high rating initially increases with nightly price (2pp/10eur), it starts to decline beyond a certain point by 1pp/10 EUR. Notably, the top 1% of predicted highly rated accommodations in Milan has an average price of 250 EUR/night, emphasizing a non-linear and statistically significant association between pricing and high ratings. All the coefficients are all statistically significant at 1%. The predicted values from this model are not bound between 0 and 1, in fact they differ more than marginally (min: -0.1997, max: 1.0943)

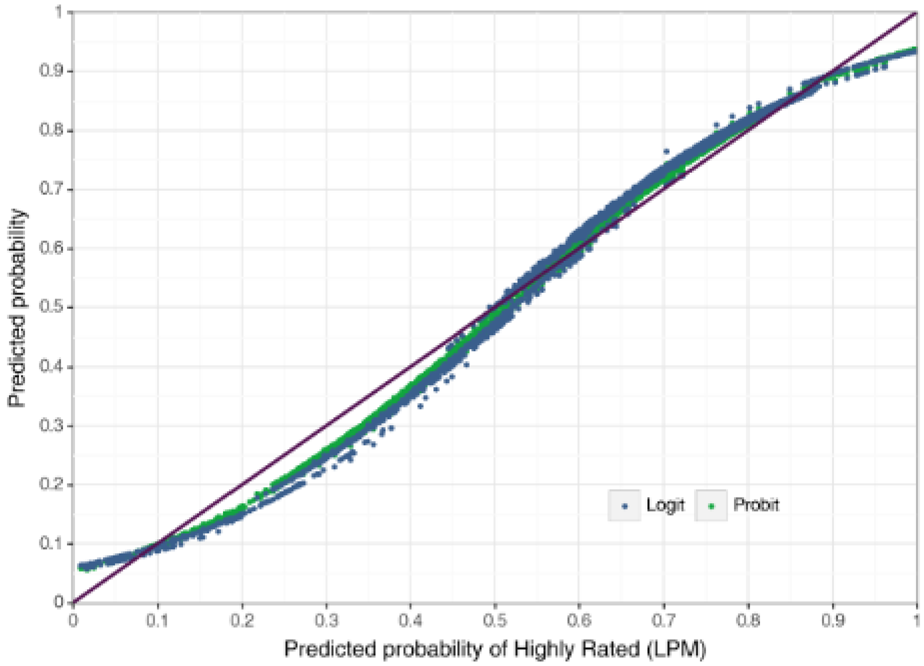
3. Logit & Probit Models

When comparing logit and probit models to lpm the marginal differences change only slightly compared to lpm coefficients while all of them remain significant. Our logit and probit models are almost indistinguishable.

Logit Marginal Effects	**dy/dx**	**std err**	**z**	**P>|z|**	**[0.005**	**0.995]**
Stars	0.2115	0.009	23.318	0.000	0.188	0.235
Distance less than 2.1	-0.1098	0.019	-5.885	0.000	-0.158	-0.062
Distance between 2.1 and 4.5	0.0457	0.011	4.032	0.000	0.017	0.075
Distance more than 4.5	0.0325	0.007	4.388	0.000	0.013	0.052
Price less than 175	0.0019	0.000	7.972	0.000	0.001	0.003
Price more than 175	-0.0005	0.000	-4.427	0.000	-0.001	-0.000

Probit Marginal Effects	**dy/dx**	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
Stars	0.2052	0.009	23.151	0.000	0.188	0.223
Distance less than 2.1	-0.1087	0.019	-5.793	0.000	-0.145	-0.072
Distance between 2.1 and 4.5	0.0449	0.012	3.884	0.000	0.022	0.068
Distance more than 4.5	0.0343	0.007	4.579	0.000	0.020	0.049
Price less than 175	0.0020	0.000	8.211	0.000	0.002	0.002
Price more than 175	-0.0005	0.000	-4.685	0.000	-0.001	-0.000

However, they produce better predicted values as at tails, probit and logit models are less extreme and bound between 0 and 1. To find which of the three models we should use for4 our conclusion we examine the goodness of fit in the next part.



4. Goodness of fit

The logit and probit models exhibit comparable predictive performance, with the logit model showing a slightly better fit. This improvement is attributed to the logit model's ability to capture more distinct probability model means and medians within the y=0 and y=1 groups. Our Brier-score also suggests that logit (0.181) and probit (0.182), both provide a better fit than our LPM (0.186). The log-loss values of each model show the similar ordering: logit (-0.545) and probit (-0.547) are better than LMP (-0.756).

	LPM	Logit	Probit
R-squared	0.253	0.272	0.269
Brier-score	0.186	0.181	0.182
Pseudo R-squared	NaN	0.210	0.207
Log-loss	-0.756	-0.545	-0.547

median	**pred_lpm**	**pred_logit**	**pred_probit**
highly_rated			
0	0.413	0.366	0.381
1	0.677	0.710	0.701
mean	**pred_lpm**	**pred_logit**	**pred_probit**
highly_rated			
0	0.403	0.396	0.400
1	0.656	0.662	0.659

5. Conclusion

Using the best fitting Logit model, we can draw the following conclusions:

The number of stars has a positive impact on the likelihood of a hotel in Milan being highly rated, with a one-unit increase associated with a 21.15% higher probability. In contrast distance in before 2.1 km has a negative impact, as hotels located less than 2.1 km away exhibit a decrease of 10.98% in the probability of being highly rated with one unit increase in distance, while those situated between 2.1 km and 4.5 km experience an increase of 4.57%, and hotels beyond 4.5 km see a rise of 3.25%. Regarding pricing, hotels priced below 175 EUR/night demonstrate an increase of 0.19% in the probability of being highly rated, while those above 175 EUR/night experience decrease of 0.05%. All of these findings are statistically significant at 99% CI.