DA2 - Hotels Europe Analysis

Adam Kovacs, Nam Son Nguyen

04 December 2021

Introduction

In this project, we analyze the **hotels-europe dataset** to investigate how high rating is related to the other hotel features in the data. We estimate linear probability, logit, and probit models with distance and stars as explanatory variables.

Our descriptive table indicates that 43% of the hotels in London are highly rates. Regarding the other two variables, the distribution of distance from center is right-skewed with an average of 2.60 kilometers, while we can infer a roughly normal distribution (mean ~ median) for stars with an average of 3.45.

Main text

From the linear probability model: Being 1 km farther away from the city center tend to have a 2 percentage points greater probability of being high rated, keeping all other variables constant. Having 1 more star is associated with a 33 percentage points higher likelihood of being high rated, all else being equal. These is statistically significant even at 5% and 1% significance level respectively. As for the type of accommodation, we chose hotels as our base category, because it has the most observations, making standard errors are lower. In the LPM, the apartment and bed and breakfast are the two categories that are statistically significant at 1% level, and both have a lower probability by 17 and 16 percentage points on average compared to hotels with the same attributes to be highly rated, holding all other variables constant. The guest house is another accommodation type that is significant at 5% level, it has on average a 12% less likelihood of being highly rate compared to a hotel with same attributes. Next, let us turn to the average marginal differences given by the logit and probit models. We can see that with regards to distance, they are exactly the same as the LMP model. As for stars, having 1 more tend to result in only 30 percentage points higher likelihood of high rating (compared to 33 estimated by the LPM), keeping everything else constant. Among the accommodation types, in these models only Apartment is significant at 1% level, which has a 14 percentage points lower likelihood of having higher rating than hotels that have the same attributes (compared to 17 in the LPM). The bed and breakfast is also significant in 5% level in both models, having a marginal effect of -15 and -14 percentage points in the logit and probit models respectively (compared to the -16 estimated by the LPM). Finally, the guest house is only significant at 5% by the logit model, such a type of accommodation with the same attributes otherwise as a hotel has on average a 12 percentage point lower likelihood of being highly rated.

We evaluated several goodness-of-fit metrics. According to the Brier score, the best accuracy of fitted values was achieved by the Logit model (lowest mean squared error between predicted and actual) followed by the Probit, and LPM. Regarding how much our prediction deviates from the actual values, log-loss tells us that the LPM model is the best performing one with the lowest log of average correction (-0.484) and Probit with the worst (-0.489). Considering all, they perform relatively poor in terms of goodness of fit, and we may have to consider more confounders to include in our future investigation.

The predicted probabilities range from -0.37 to 1.11. Probabilities below 0 and above 1 do not make economic sense. To get an idea about these extreme values, we looked at the average values of the covariates in the bottom and top 1% of the distribution. At the bottom, the average of stars is 1.88, the average distance from the center is 2.07 km and the mode of accomodation type is Bed and breakfast. As for the top 1%, the average of stars is 3.47, the average distance from the center is interestingly more at 2.6 km and the mode of accomodation type is Hotel.

We compare our 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. Compared to the LPM, they are practically the same around 0.5, but give higher estimates for higher values, and lower for lower values except for the tails where it is in the other way around. The range of predicted values of the logit model is [0.007, 0.967], while for the probit it is [0.002,0.976]. These are narrower than that of the LPM.

Conclusion

In London, accommodations farther away from the city center and having more stars are both associated with higher likelihood of being high rated by around 2 and 30-33 percentage points respectively. We also used accommodation type as a further explanatory variable and found that hotels have the highest probability of being highly rated (other types either lower or same). Among the models, based on the Brier score, the logit model gave the most precise estimates.