Data Analysis 2 - Assignment 2

Introduction

This is a probability analysis on hotel ratings in Rome in 2018, that aims to uncover the relationship between hotel ratings being high or not, with regards to key attributes. The goal is examine how high rating is related to the other hotel features, and predict the probability of a hotel being rated higher than 4.0.

Executive summary

If our goal is to achieve a higher rating for our hotel with the least minimum resources (cost), we should select a location around 2.8 km away from the city center, and have as many stars as possible (assuming that gaining one star has an equal cost despite location, and real estate is more costly towards the center)

Potential improvements to the model

- include price, or overpriced / underpriced binary variable (from linear regression)
- include interactions between stars and distance, to see in which distance segment is it more beneficial to spend more money on one extra star

Data

The dataset used is the hotels europe dataset, accessable here: https://osf.io/r6uqb/, and the city in scope is Rome.

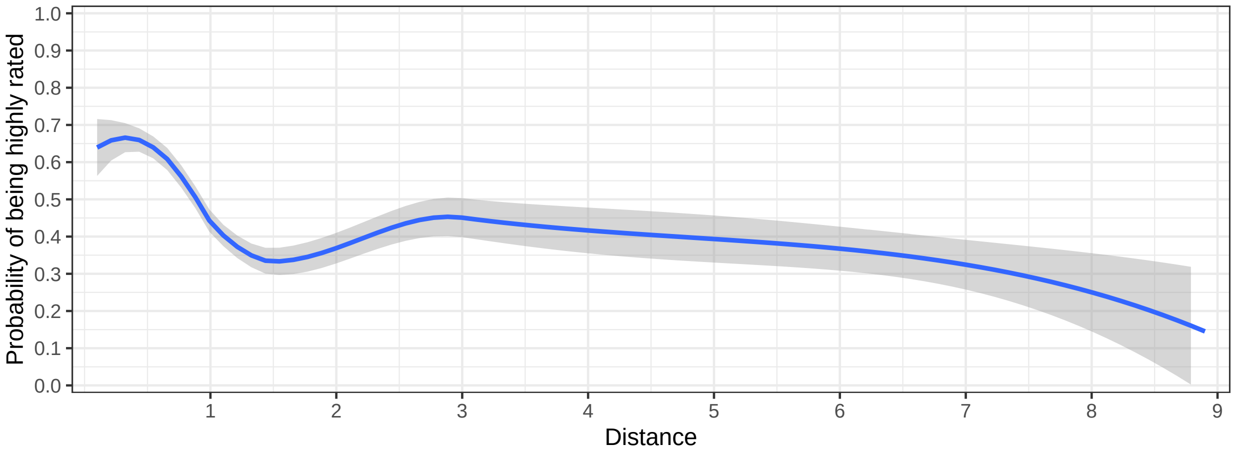
I have narrowed down the dataset to 2972 observations: hotels in 2018, below price of 650 USD, maximum 10 km from the center, with at least 10 ratings. I selected a single year to prevent time inconsistencies, e.g. new hotel built or old one renovated. Forbetter interpretation, I excluded c.20 hotels with 3.5 starts. The price and distance was restricted after examining the distributions (out of scope)

My left-hand side variable is a binomial variable, highly_rated, which is equal to 1 if rating is above 4.0, and is 0 otherwise.

Preliminary analysis

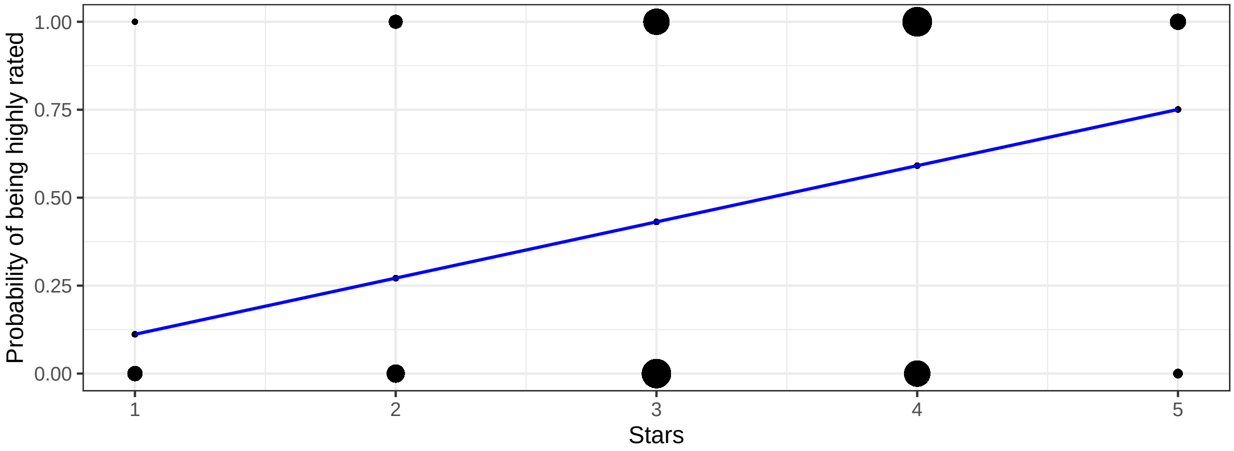
I examined the selected explanatory variables (distance and number of stars) to decide how to include them in the final models,

Distance: Given the shape of the loess function, should be included as linear spline with knots at 1.5 and 2.8



Stars: The simple lpm model of *highly_rated* ~ *stars* shows us the probability of being highly rated per stars (the ratio of highly rated hotels within total number per the stars category)

From this we learn that the more stars, the higher the rating (about linear)



Probability models

I have created five models:

- 3 LPM models, one with distance as linear variable, one with distance as lspline with knots at 1.5 and 2.8, and one with distance as lspline and stars as well
- A logit model with distance as lspline with knots at 1.5 and 2.8, and stars
- A probit model with distance as lspline with knots at 1.5 and 2.8, and stars

Results

- all results are statistically signifant at p=5% at least
- lpm1 shows that on avg. if distance is 1 km higher, rating is lower by 0.036
- lpm2 shows the coefficient is negative for <1.5 km and >2.8 km, but positive for 1.5-2.8 km segment. Meaning that in 1.5-2.8 km range, hotels that are further away by 1 km have a higher rating by 0.140 on avg.

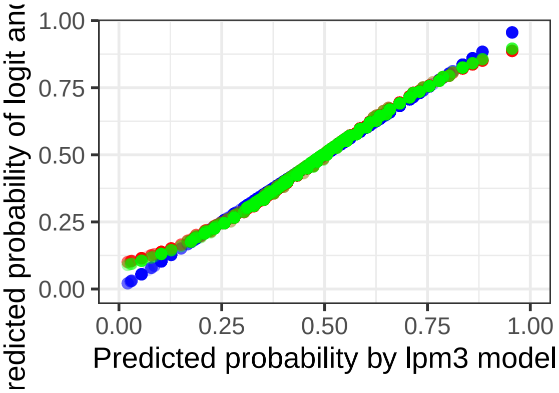
- In lmp3, introducing stars as a variable decreases the coefficient for distance in the 1st and (from -0.301 to -0.241) and in the 2nd segment (from 0.14 to 0.072), but not in the 3rd segment. From this we can conclude, that if d<1.5 km then 1 km higher distance from the center has a more significant impact that 1 more stars, but in the 1.5-2.8 km segment stars are more important, and in the d>2.8 km segment stars and distance are equally important.
- the logit and probit models have almost same coefficients as the lmp3 model (se figure below)

	lpm1	lpm2	lpm3	logit_marg	probit_marg
distance	-0.036 ***				
	(0.005)				
stars			0.153 ***	0.156 ***	0.156 ***
			(0.010)	(0.013)	(0.010)
dist<1.5		-0.301 ***	-0.241 ***	-0.234 ***	-0.232 ***
		(0.026)	(0.026)	(0.028)	(0.025)
1.5<dist<2.8		0.140 ***	0.072 *	0.071 *	0.070 *
		(0.030)	(0.029)	(0.029)	(0.029)
dist>2.8		-0.037 **	-0.036 **	-0.034 **	-0.035 **
		(0.012)	(0.011)	(0.011)	(0.011)
N	2972	2972	2972	0	0

*** p < 0.001; ** p < 0.01; * p < 0.05.

Predictions

- predictions as expected, close to lmp model, with divergence in the lower and higer range


redicted probability of logit anc
Predicted probability by lpm3 model