Assignment_2 Documentation TING Chun Yin 2204502 https://github.com/tingchunyin/da2_regression

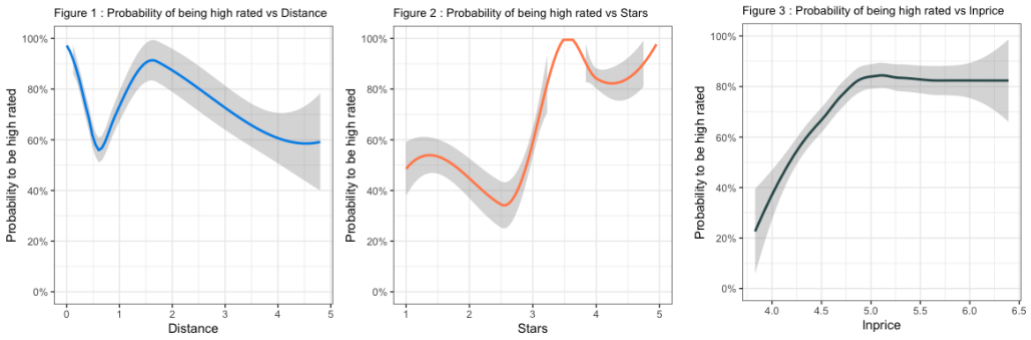
Overview:

We would like to know the relationship between Distance/Stars/Price and customer hotel ratings. Hotels-Europe data is used in this assignment. The hotels price and feature tables are merged by using left join.

Data filtering and transformation:

Records where "hotel_id", "stars" and "distance" are null are removed from the data frame. Hotel user rating are used as the dependent variable (y) and added a binary variable column "high_rated", which "1" indicates hotels which have ratings higher than or equal to 4, otherwise "0" is being used. **Seville** is the chosen city for this assignment, and only "hotels" as accommodation type which are cheaper than or equal to USD$600 per night are included.

Visualization:



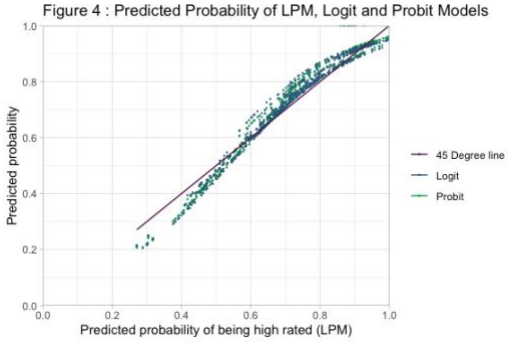
With the "loess regression" in Figure 1, looking at the local minimums and maximums, I think it would be a good idea to **introduce knots at 0.6 and 1.6 miles mark**, where the regression line changes its direction. In Figure 2, it is observed that being high rated (y) is probably positively correlated depending on stars of the hotel (x). In Figure 3, it is observed that being high rated (y) is probably positively correlated depending on the log price of the hotel (x), but the probability become stable when log price is above 5.0. And in Figure 4, it is observed that both the Logit and Probit are close to each other and is also close to the LPM line as the s-shaped curve is lying close to the 45 degree line.

Analytical Result:

For **stars**, in the Regression Model Summary table, we can see that hotels with both **1-4 stars or above 4 stars** have a similar probability of 9.6%-9.8% to have a higher rating (y) depending on one unit increase of stars (x), so it can possibly be concluded that stars has a positive correlation of 9.7% with the rating, but the correlation is not likely to differ too much while comparing between hotels with 1-4 stars and above 4 stars.

For **Distance**, in the **first 0.6 mile**, with every unit increase in distance, the probability of the hotel being high rated is 69.7% lower. Opposingly, for distance further than **0.6 to 1.6 mile**, with every unit increase in distance, the probability of the hotel being high rated is 27.8% higher. For distance **further than 1.6 mile**, the probability of the hotel being high rated is 10.2% lower.

For **Price**, the LPM model suggest that when there is 1 unit increase in price, the probability of the hotel to be high rated will increase by 8.1%, while the Logit and Probit model suggest the probability is 7.5% for the same increment in price.



To have an overall view on the difference between the Logit/Probit and the LPM, from Figure 4, it is observed that the Logit and Probit look similar with each other. At the same time, both Logit and Probit are close to the LPM regression as showed by the s-shaped curve lying close to 45 degree line.

Appendix

data summary

Data Summary

	mean	SD	min	max	median	p95	N
high_rated	0.74	0.44	0.00	1.00	1.00	1.00	895
distance	0.87	0.99	0.00	4.80	0.60	3.70	895
stars	3.33	1.00	1.00	5.00	4.00	5.00	895
lnprice	4.94	0.57	3.83	6.39	4.88	6.02	895

Regression Model Summary

Regression Model Summary of Hotels in Seville

	1. LPM	2. logit coeffs	3. logit marg	4. probit	5. probit marg
Constant	0.277*	−1.035	−1.035	−0.666	−0.666
	(0.139)	(0.874)	(0.874)	(0.510)	(0.510)
4 stars or lower	0.098**	0.526**	0.083**	0.326**	0.087**
higher than 4 stars	0.096	15.455	0.266**	4.291	0.265**
disatnce >= 0.6	−0.697**	−4.658**	−0.730**	−2.693**	−0.715**
disatnce >0.6, <=1.6	0.278**	1.529**	0.240**	0.896**	0.238**
disatnce >1.6	−0.102**	−0.503**	−0.079**	−0.309**	−0.082**
lnprice	0.081**	0.480**	0.075**	0.282**	0.075**
Num.Obs.	895	895	895	895	895
RMSE	0.40	0.40	0.40	0.40	0.40

* p < 0.05, ** p < 0.01