DA2 Assignment 2

For this assignment, we will analyze the probability prediction for whether a hotel is highly rated, regressed on some independent variables using the Linear Probability Model, Logit Model and Probit Model.

Data Cleaning & Filtering

- Filter for 'accommodation_type' as Hotel and 'city' is Paris (n = 16,876).
- Filter for non-na 'rating' and 'stars' as we will use these fields as covariates (n = 15,154).
- Filter for 'rating_reviewcount' >= 100 as we want hotels with lots of reviews (n = 8,851).
- Take the natural log of 'distance' and 'price' as we want to analyze the effect of distance and price increase, not the actual values of distance and price.
- Create the binary variable 'highly_rated' as we will use this as our left-hand side variable.

The regression formula

We will build our probability models by regressing the binary variable 'highly_rated' on various independent variables as follows:

$$highly\_rated = \beta_0 + \beta_1 \ln(distance) + \beta_2 stars + \beta_3 \ln(price) + \beta_4 weekend + \beta_5 offer + e$$

Analysis Explanation

Coefficients Estimation

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-1.3807	0.045	-30.991	0.000	-1.468	-1.293
ln_distance	-0.0082	0.006	-1.329	0.184	-0.020	0.004
stars	0.2345	0.006	38.956	0.000	0.223	0.246
ln_price	0.2028	0.009	21.990	0.000	0.185	0.221
weekend	0.1069	0.011	10.006	0.000	0.086	0.128
offer	0.1088	0.010	10.584	0.000	0.089	0.129

fig. 1: Linear Probability Model coefficient estimation

	dy/dx	std err	z	P>|z|	[0.025	0.975]
ln_distance	-0.0076	0.006	-1.207	0.227	-0.020	0.005
stars	0.2421	0.006	38.771	0.000	0.230	0.254
ln_price	0.2134	0.010	22.231	0.000	0.195	0.232
weekend	0.1069	0.011	10.181	0.000	0.086	0.128
offer	0.1089	0.010	11.167	0.000	0.090	0.128

fig. 2: Logit Model marginal difference coefficient estimation

	dy/dx	std err	z	P>|z|	[0.025	0.975]
ln_distance	-0.0087	0.006	-1.399	0.162	-0.021	0.003
stars	0.2404	0.006	39.598	0.000	0.228	0.252
ln_price	0.2082	0.009	22.183	0.000	0.190	0.227
weekend	0.1075	0.010	10.266	0.000	0.087	0.128
offer	0.1062	0.010	10.793	0.000	0.087	0.125

fig. 3: Probit Model marginal difference coefficient estimation

From the three figures (fig. 1, fig. 2, fig. 3), we can interpret the following information (with each statement made while keeping the other variables constant):

- **ln_distance**: For a 1% increase in distance, the probability of the hotel being highly rated decreases by approximately 0.76%-0.82%. The coefficient is not statistically significant, suggesting that distance might not have a strong correlation with the probability of the hotel being highly rated.
- **stars**: For each additional star of the hotel, the probability of the hotel being highly rated increases by approximately 23.45%-24.21%. This is significant at 1%, suggesting a strong positive correlation with the probability of the hotel being highly rated.
    - Higher star hotels offer better quality and hence receive better reviews?
- **ln_price**: For a 1% increase in price, the probability of the hotel being highly rated increases by approximately 20.28%-21.34%. This is significant at 1%, suggesting a strong positive correlation with the probability of the hotel being highly rated.
    - This might suggest the same behavior as the **stars** variable above since more expensive hotels offer better quality?
- **weekend**: If the hotel offer rooms on weekend, the probability of the hotel being highly rated increases by 10.69%-10.75% compared to that on weekdays. This is significant at 1%, suggesting a strong positive correlation with the probability of the hotel being highly rated.
    - Hotels' weekend experience is better than weekday experience?
- **offer**: If the hotel has a promotional offer, the probability of the hotel being highly rated increases by approximately 10.62%-10.89%. This is significant at 1%, suggesting a strong positive correlation with the probability of the hotel being highly rated.
    - Promotional offers give incentives to customers to provide better reviews?

Predicted Probability Estimation

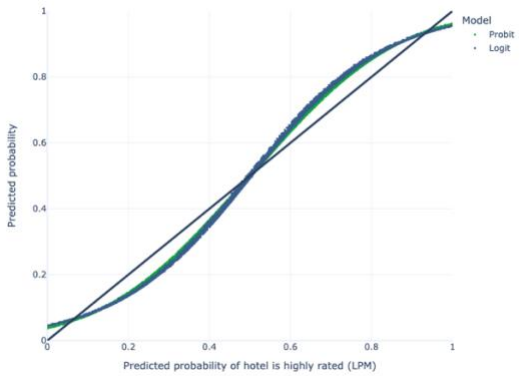

fig. 4: LPM, Logit & Probit predicted probabilities

	LPM	Logit	Probit
R-squared	0.287	0.308	0.307
Brier-score	0.175	0.170	0.170
Pseudo R-squared	NaN	0.254	0.253
Log-loss	-0.585	-0.510	-0.510

fig. 5: All model fitness scores

	lpm_pred	logit_pred	probit_pred
count	8851.000000	8851.000000	8851.000000
mean	0.571348	0.571348	0.570444
std	0.265185	0.273221	0.270408
min	-0.345381	0.005751	0.001334
25%	0.413347	0.370626	0.377961
50%	0.550004	0.571845	0.565997
75%	0.763716	0.830895	0.821913
max	1.464411	0.997302	0.999676

fig. 6: Description of all model predicted probabilities

From fig. 4, the Logit and Probit models have non-linearity but are pretty close to the LPM models.
From fig.5 and fig.6, we can make some statements as follows:

- The Logit and Probit models give a narrower predicted probabilities range (0.001 - 0.999) compared to the LPM's range (-0.345 - 1.464) thanks to the non-linearity of the Logit and Probit.
- The Brier-score for Logit and Probit models are also a bit better than that of the LPM (0.170 compared to 0.175). The smaller the Brier-score, the better calibrated the model.
- - Although the differences are not significant, it is still better to use the Logit and Probit models for probability prediction than the Linear Probability Model.

Summary

- From our LPM, Logit and Probit models, the probability of a hotel is highly rated has strong correlations to multiple covariates:
    - stars: higher stars hotels have a better chance of being highly rated
    - price: more expensive hotels have a better chance of being highly rated
    - weekend: hotels offering rooms during the weekend have a better chance of being highly rated
    - offer: hotels having running promotional offers have a better chance of being highly rated
- Interestingly, distance to the city center does not significantly contribute to the probability of a hotel is highly rated. Some might expect that the further the hotel is away from the city center, the lower the probability that the hotel is highly rated. We do not see that pattern here in our sample.