Data Analysis Assignnment2

Jo Kudo: 2303175

Goal of this assignment

The purpose of this research is to investigate the potential relationship between the highly rated status of hotels and a couple of key factors: the number of stars a hotel has, its distance from a central point. The highly rated status is a binary variable, indicating 'True' if a hotel's rating is four or higher, and 'False' if otherwise. To assess the impact of price more accurately, we will analyze the logarithm of the hotel prices, which allows us to focus on relative price differences rather than absolute values. Our analysis will include linear probability models, logit models, and probit models, all of which were covered in our coursework.

Data filtering

The dataset has been meticulously filtered and cleaned as follows:

- This study is confined to hotels located within the central area of Athens.
- Only weekdays in December since 2017 are considered to maintain consistency.

Analysis Interpretation

For our analysis, we employed three distinct statistical models. The details of the Linear Probability Model are exhibited in Figure1. This model suggests that a hotel's likelihood of being highly rated is primarily influenced by its number of stars. Specifically, each additional star increases the probability of a hotel being highly rated by approximately 21.2%. However, the distance correlates with a 12.3% drop in the probability of a high rating.

Also, logit and probit model are seen. The logit assumes a logistic distribution error, while the probit assumes a normal distributed errors. The results are shown by Figure2, and 3

OLS Regression Results

Dep. Variable:	highly_rated	R-squared:	0.213
Model:	OLS	Adj. R-squared:	0.208
Method:	Least Squares	F-statistic:	42.78
Date:	Wed, 06 Dec 2023	Prob (F-statistic):	3.60e-17
Time:	23:34:41	Log-Likelihood:	-191.79
No. Observations:	319	AIC:	389.6
Df Residuals:	316	BIC:	400.9
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.0220	0.083	-0.266	0.791	-0.185	0.141
stars	0.2116	0.023	9.086	0.000	0.166	0.257
distance	-0.1234	0.052	-2.386	0.018	-0.225	-0.022

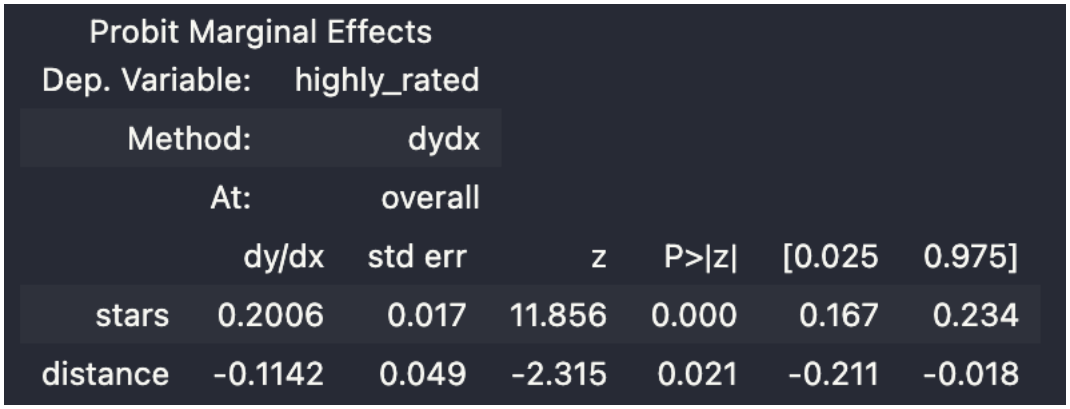
Omnibus:	108.370	Durbin-Watson:	0.932
Prob(Omnibus):	0.000	Jarque-Bera (JB):	16.675
Skew:	-0.038	Prob(JB):	0.000239
Kurtosis:	1.883	Cond. No.	12.4

Figure1: OLS Regression in terms of stars and distance, considering rating.

Logit Marginal Effects

Dep. Variable:	highly_rated					
Method:	dydx					
At:	overall					
	dy/dx	std err	z	P>|z|	[0.025	0.975]
stars	0.1994	0.017	11.907	0.000	0.167	0.232
distance	-0.1183	0.051	-2.337	0.019	-0.218	-0.019

Figure2: Logit Marginal Effects



Probit Marginal Effects

Dep. Variable:	highly_rated					
Method:	dydx					
At:	overall					
	dy/dx	std err	z	P>|z|	[0.025	0.975]
stars	0.2006	0.017	11.856	0.000	0.167	0.234
distance	-0.1142	0.049	-2.315	0.021	-0.211	-0.018

Figure3: Probit Marginal Effects