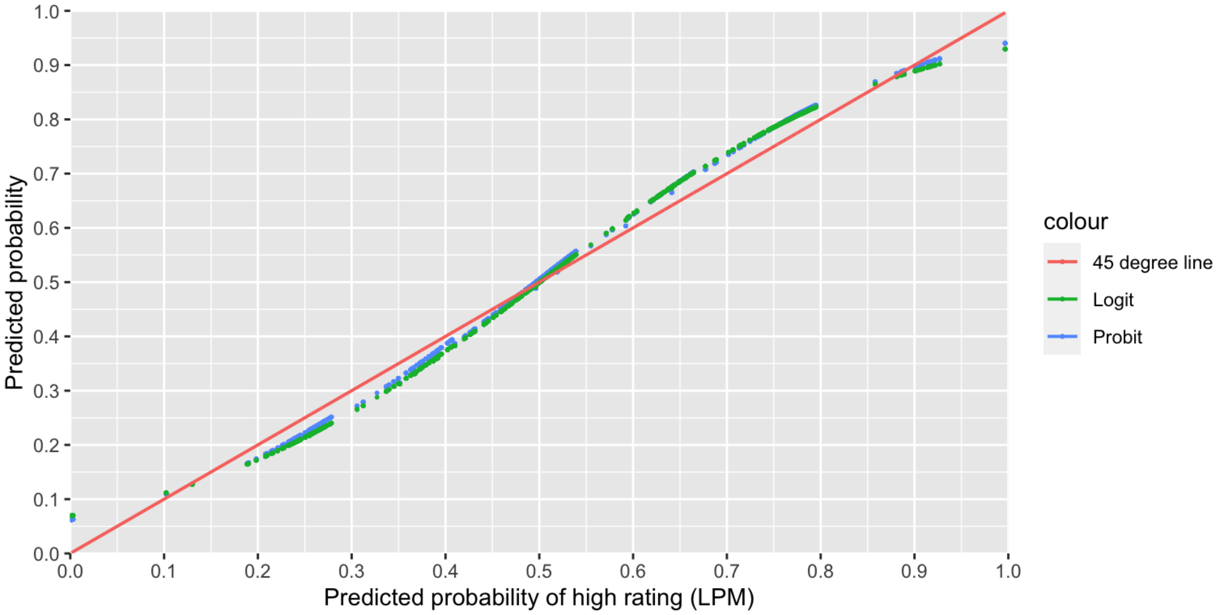
DA2 - Assignment 2

Introduction

In this analysis I investigate whether there is a relationship between hotels' (high) rating and their stars and distances to the city center. I used the hotel-europe dataset (https://osf.io/p6tyr/) with the location set to Berlin.

I created a binary variable based on the hotel ratings, if a hotel is rated 4 or above, the highly_rated column takes a value of 1, in other cases 0. Based on this, the mean value of this variable for the dataset is 0.5976124. That is also the probability of a hotel being highly rated in Berlin.

As the next step, I calculated the predictions of high rating (of hotels) using the linear probability, logit and probit models and plotted the results in the following graph.



The 45 degree line is the linear probability model, the two S-curves are the logit and probit models. Based on the eye-test, the results look similar, but let's take a look at the coefficient results in the next chart.

	LPM	logit coeffs	logit marginals	probit coeffs	probit marginals
Constant	−0.225**	−3.770**		−2.237**	
	(0.033)	(0.183)		(0.105)	
distance	−0.016**	−0.086**	−0.016**	−0.055**	−0.017**
	(0.004)	(0.017)	(0.004)	(0.010)	(0.004)
stars	0.257**	1.339**	0.255**	0.802**	0.255**
	(0.008)	(0.053)	(0.014)	(0.030)	(0.007)
Num.Obs.	4200	4200	4200	4200	4200
Std.Errors	Heteroskedasticity-robust	IID		IID	

* p < 0.05, ** p < 0.01

In case of logit and probit models, instead of the raw coefficients, we interpret the marginal differences, which have the same interpretation as the coefficients in case of linear probability models. Based on this the three models produce very similar results. If we look at the logit marginals we can see, that if the distance is greater, we can expect the probability of high rating to be lower by 1.65%. And if a hotel posesses more stars, the probability of it being highly reated is greater by 25.55%.

Goodness of fit

Based on the Brier-score, the LPM model provides the best prediction with a value of 0.19. The whole table can be found in the appendix.

Conclusion

We can conclude, that not surprisingly the farther the hotel from the city center, the lower the possibility of a high rating. Even less surprising, that the higher the number of a stars, the higher the possibility of a high rating. It should be noted, that probably in Berlin, where there is no definite city center, like in Budapest for example, the so called distance variable might not have a strong relationship with the probability of high rating, since the people, who visit Berlin, usually have different preferences in terms of choosing a location for accomodation.

Appendix

Summary table on the ‘highly_rated’ variable

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##  0.0000  0.0000  1.0000  0.5976  1.0000  1.0000      30
```

Linear probability model

```
## OLS estimation, Dep. Var.: highly_rated
## Observations: 4,200 
## Standard-errors: Heteroskedasticity-robust 
##              Estimate Std. Error  t value   Pr(>|t|)    
## (Intercept) -0.225110   0.032887 -6.84495 8.7583e-12 ***
## distance    -0.016338   0.004005 -4.07891 4.6087e-05 ***
## stars        0.257430   0.007587 33.92948  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 0.438341   Adj. R2: 0.199499
```

Logit model

```
## GLM estimation, family = binomial(link = "logit"), Dep. Var.: highly_rated
## Observations: 4,200 
## Standard-errors: IID 
##              Estimate Std. Error   t value   Pr(>|t|)    
## (Intercept) -3.769987   0.182621 -20.64374  < 2.2e-16 ***
## distance    -0.086283   0.017200  -5.01657 5.2601e-07 ***
## stars        1.338895   0.052640  25.43515  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Log-Likelihood: -2,360.5   Adj. Pseudo R2: 0.164552
##            BIC:  4,746.1     Squared Cor.: 0.199382
```

Probit model

```
## GLM estimation, family = binomial(link = "probit"), Dep. Var.: highly_rated
## Observations: 4,200
## Standard-errors: IID
##              Estimate Std. Error   t value   Pr(>|t|)
## (Intercept) -2.236548   0.105372 -21.22532  < 2.2e-16 ***
## distance    -0.054578   0.010477  -5.20958 1.8927e-07 ***
## stars        0.802246   0.029677  27.03225  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Log-Likelihood: -2,356.6   Adj. Pseudo R2: 0.165951
##            BIC:  4,738.2     Squared Cor.: 0.199438
```

Logit marginal differences

```
## Call:
## logitmfx(formula = model_formula, data = data, atmean = FALSE,
##     robust = T)
##
## Marginal Effects:
##              dF/dx Std. Err.       z     P>|z|
## distance -0.016465  0.003582 -4.5966 4.294e-06 ***
## stars     0.255495  0.013605 18.7791 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Probit marginal differences

```
## Call:
## probitmfx(formula = model_formula, data = data, atmean = FALSE,
##     robust = T)
##
## Marginal Effects:
##               dF/dx  Std. Err.       z     P>|z|
## distance -0.0173201  0.0035900 -4.8246 1.403e-06 ***
## stars     0.2545875  0.0073607 34.5872 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Brier-score

	lpm <chr>	**logit** <chr>	**probit** <chr>
Dependent Var.:	highly_rated	highly_rated	highly_rated
(Intercept)	-0.2251*** (0.0329)	-3.770*** (0.1826)	-2.237*** (0.1054)
distance	-0.0163*** (0.0040)	-0.0863*** (0.0172)	-0.0546*** (0.0105)
stars	0.2574*** (0.0076)	1.339*** (0.0526)	0.8022*** (0.0297)
_______________	_________________	_________________	_________________
Family	OLS	Logit	Probit

	lpm	**logit**	**probit**
	<chr>	<chr>	<chr>
S.E. type	Heteroskedast.-rob.	IID	IID
Brier score	0.19214	0.19227	0.19228

9 rows