Partners: Corinne & Myitzu

Major decisions

We chose the city of Amsterdam as it had many observations (2,058). We added a column to the dataframe called 'highly_rated' which was simply the 'rating' column when hotels were rated 4 or higher. We then made the 'highly_rated' variable a dummy variable (1 if highly_rated (rating of 4 or larger), 0 if not (rating less than 4)).

LPM: We performed a linear probability model regression (LPM) in order to see how unit changes in *distance* and *stars* would affect the probability of the dependent variable being 1 (a hotel being highly rated).

-constant/intercept coefficient: disregarding explanatory variables, at zero, probability of hotels being highly rated is -.4371

-distance coefficient: keeping stars constant, a one unit increase in distance means a .0273 increase in the probability of hotels being highly rated

-stars coefficient :keeping distance constant, a one unit increase in stars means a .2876 increase in the probability of hotels being highly rated

-p-values for both x-variables were 0 (or rounded to zero) meaning that the variables are statistically significant (there is most likely a relationship between the x-variables and probability of hotels being highly rated)

However, LPM assumes linearity and just in case the data turns out to be better fit by a nonlinear probability model, we should also consider using logit and probit regressions to model our data.

Logit: We produced a table with logit regression results, however for logit regressions, the coefficients are hard to interpret, so we must use **marginal effects** in order to get coefficients similarly interpretable to a LPM. We then create a logit marginal effects table.

Results:

-distance coefficient: keeping stars constant, a one unit increase in distance means a .0210 increase in the probability of hotels being highly rated

-stars coefficient :keeping distance constant, a one unit increase in stars means a .2742 increase in the probability of hotels being highly rated

-p-values for both x-variables were 0 (or rounded to zero) meaning that the variables are statistically significant (there is most likely a relationship between the x-variables and probability of hotels being highly rated)

Probit: We produced a table with probit regression results, however similarly to logit, the coefficients are hard to interpret. Therefore, we created a probit marginal effects table.

Results:

-distance coefficient: keeping stars constant, a one unit increase in distance means a .0211 increase in the probability of hotels being highly rated

-stars coefficient : (same coefficient and interpretation as 'stars' for logit).

-p-values: (same p-values for both x-variables and interpretations for logit)

Predicted Probabilities:

The predicted probabilities for all three models (LPM, Logit, and Probit) were all similarly in the 0.5 region indicating that ambiguity in predicting the probability of hotels being highly rated. However, looking at the fit of the predicted probabilities (using R-squared and Brier-score), the Logit model proved to provide the best fit for the predicted probability of hotels being highly_rated. (Probit model was almost exactly the same)

Overall Summary:

-Distance and Stars seem to have a relationship with the dependent variable, hotels being highly rated (we would reject the null hypothesis) and the Logit Model proves to provide the best fit for understanding the marginal effects.