Data Analysis : Assignment 2

Ghazal Ayobi and Shah Ali Gardezi

Introduction

The Question of this case study is how hotels distance from center is related to highly rated hotels. For this assignment we use Hotels-Europe data. This data set contains two tables Features and Price. We joined the two tables using left join.

Data Transformation

As a process of filtering and data transformation, we use hotel user rating as the dependent variables and transformed it to a binary variable called *highly_rated* which equals to one if *rating* is more than *4*, *0* otherwise. We selected ***Paris*** City and considered ***Hotels*** as accommodation type. Moreover, we excluded hotels with less than *USD600* per night, removed null and duplicated values from the data set.

Analysis

In the first place in order to understand what functional form to include in the regression. We examined lowess regression with highly rated hotels to decide in what functional form to include stars and distance. The figures in the appendix shows that distance and highly rated hotels have negative relationship. distance increases by one mile and. between 1.2 miles and 3 miles highly rated hotels and distance does not indicate in correlation. after 3 miles, highly rated declines. as the graph indicates after 3 stars the relationship between highly rated and stars changes. Price and highly rated has interesting relationship one unit change in price till USD 250 has positive relationship, after it declines. Thus, we included distance as a piece-wise linear spline with knots at 1.2 and 3 miles and stars with a knot at 3 stars. For additional variables we included log of price as a piece-wise linear spline with knot at 5.5. Other variables are *price* which was transformed to log of Price and *weekend*.

- Filtered the data for **Paris**
- Selected Hotels and Apartments for accommodation type
- Price is less than 600
- Removed null values from *stars*
- Removed duplicates
- Removed Null values from *rating*, *stars*, and *distance*
- Created log of price, *lnprice*

Table 1: Summary Statistics

	Mean	SD	Min	Max	Median	P95	N
highly_rated	0.56	0.50	0.00	1.00	1.00	1.00	11397
distance	1.62	0.78	0.10	4.20	1.60	2.90	11397
stars	3.22	0.78	1.00	5.00	3.00	4.00	11397

- Created *highly_rated* if rating >= 4

```
cm <- c('(Intercept)' = 'Constant')
summary1 <- msummary(list(lpm, logit, logit_marg, probit, probit_marg),
         fmt="%.3f",
         gof_omit = 'DF|Deviance|Log.Lik.|F|R2 Adj.|AIC|BIC|R2|PseudoR2',
         stars=c('*' = .05, '**' = .01),
         coef_rename = cm,
         coef_omit = 'as.factor(country)*'

)
```

```
# adding pseudo R2 (not work for mfx)
glance_custom.glm <- function(x) data.frame(`PseudoR2` = pR2(x)["McFadden"])
cm <- c('(Intercept)' = 'Constant')

summary2 <- msummary(list(lpm, logit, probit),
         fmt="%.3f",
         gof_omit = 'DF|Deviance|Log.Lik.|F|R2 Adj.|AIC|BIC',
         stars=c('*' = .05, '**' = .01),
         coef_rename = cm
)
```

```
## fitting null model for pseudo-r2
## fitting null model for pseudo-r2
```

```
#distance
g1 <- ggplot(data = data, aes(x=distance, y=highly_rated)) +
  geom_smooth(method="loess", color="3a5e8cFF") +
  scale_y_continuous(expand = c(0.01,0.01),limits = c(0,1), breaks = seq(0,1,0.2), labels = scales::per
  labs(x = "Distance",y = "Probability of Highly Rated") +
  theme_bw() +
  ggtitle("Probability of Highly Rated vs Distance") +
  theme(plot.title = element_text(size = 12), axis.title = element_text(size=8) )
```

```
g2 <- ggplot(data = data) +
  geom_point(aes(x=pred_lpm, y=pred_probit, color="Probit"), size=0.5,  shape=16) +
  geom_point(aes(x=pred_lpm, y=pred_logit,  color="Logit"), size=0.5,  shape=16) +
  geom_line(aes(x=pred_lpm, y=pred_lpm,    color="45 Degree line"), size=0.5) +
  labs(x = "Predicted probability of Highly Rated (LPM)", y="Predicted probability")+
  scale_y_continuous(expand = c(0.00,0.0), limits = c(0,1), breaks = seq(0,1,0.2)) +
  scale_x_continuous(expand = c(0.00,0.0), limits = c(0,1), breaks = seq(0,1,0.2)) +
  scale_color_manual(name = "", values=c("#541352FF", "#3a5e8cFF","#10a53dFF")) +
  theme_bw() +
theme(legend.position=c(0.55,0.08),
        legend.direction = "horizontal",
        legend.text = element_text(size = 7))
```