HW2

Zsiros, Gabriella

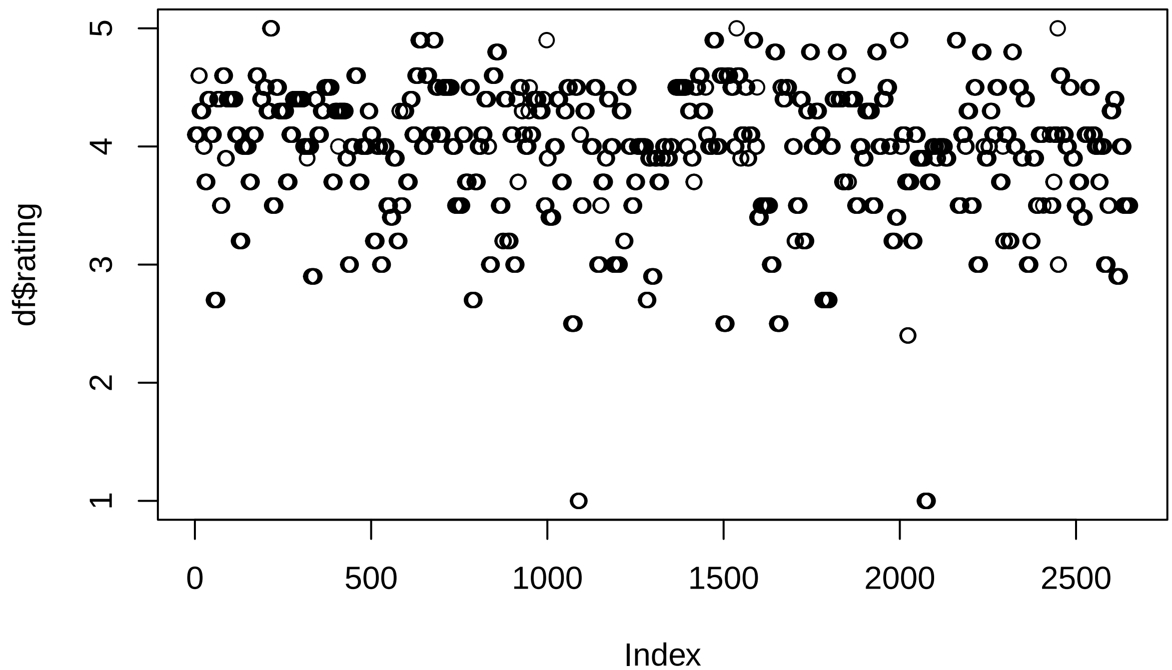
2022-12-10

Data selection

After loading the two data tables, we determine the target city, in this case Budapest.Joining the datatables we get the following structure:

```
##  [1] "hotel_id"           "city"               "distance"
##  [4] "stars"              "rating"             "country"
##  [7] "city_actual"        "rating_reviewcount" "center1label"
## [10] "center2label"       "neighbourhood"      "ratingta"
## [13] "ratingta_count"     "distance_alter"     "accommodation_type"
## [16] "price"              "offer"              "offer_cat"
## [19] "year"               "month"              "weekend"
## [22] "holiday"            "nnights"            "scarce_room"
```

Creating a quick overview with plotting, we can see that some extreme values showing up around rating = 1. By rational thinking this seems to be distorted, since people tend to give radically negative reviews after some bad experience that may not always be in proportion with the actual overall impression and is very subjective.

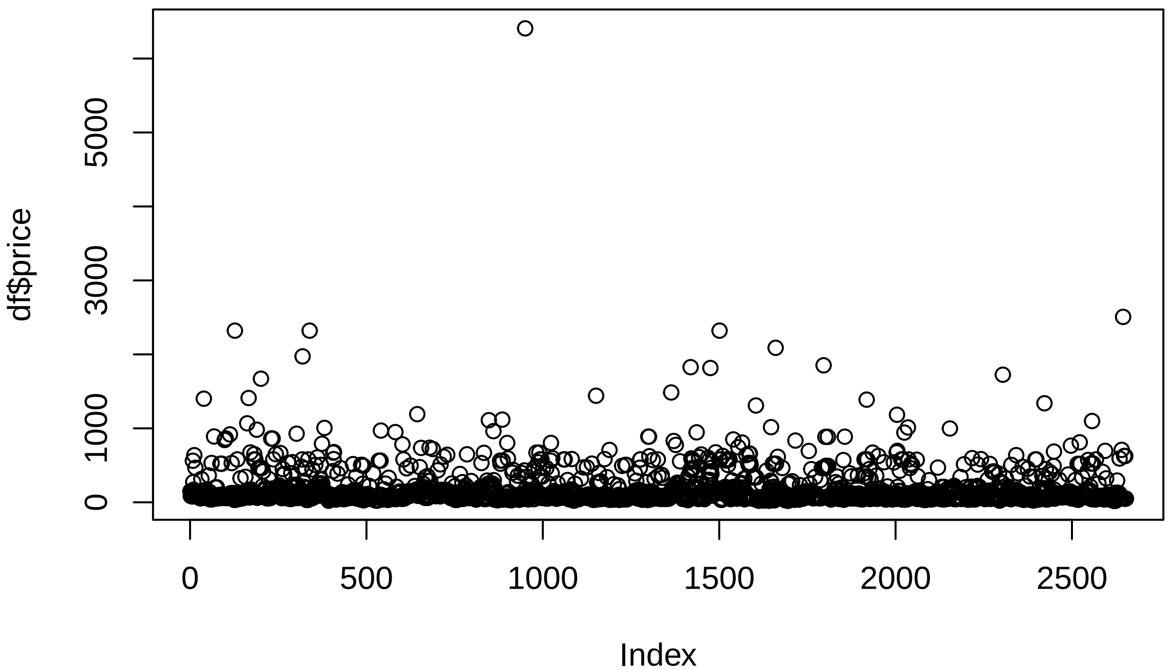



```
##    df$rating   n
## 1        1.0  14
## 2        2.4   3
## 3        2.5  29
## 4        2.7  45
## 5        2.9  29
## 6        3.0 103
## 7        3.2  99
## 8        3.4  51
## 9        3.5 216
## 10       3.7 185
## 11       3.9 164
## 12       4.0 402
## 13       4.1 273
## 14       4.3 221
## 15       4.4 235
## 16       4.5 268
## 17       4.6 114
## 18       4.8  62
## 19       4.9  51
## 20       5.0   7
## 21        NA  82
```

Similar overview on price table shows extreme values above price >6000.

Table 1: Statistical overview

	Unique (#)	Missing (%)	Mean	SD	Min	Median	Max	
distance	34	0	1.0	0.8	0.0	0.8	5.9	
rating	19	0	4.0	0.5	2.4	4.0	5.0	
price	423	0	165.7	220.0	19.0	89.0	2507.0	



After filtering out the extremes, we can check a summary of the final dataset. (Table 1: Statistical overview)

After that, we introduce a binary variable on high rating (highly_rated), categorizing rating greater or equal than 4 as 1, and 0 if the rating is below.

```
##   hotel_id stars rating price highly_rated
## 1     3078     4    4.1   150            1
## 2     3078     4    4.1    86            1
## 3     3078     4    4.1    99            1
## 4     3078     4    4.1   150            1
## 5     3078     4    4.1   130            1
## 6     3078     4    4.1    79            1
```

```
##      hotel_id stars rating price highly_rated
## 2551     3417     3    3.5   626            0
## 2552     3417     3    3.5    50            0
## 2553     3417     3    3.5    50            0
```

```
## 2554     3417     3    3.5   626             0
## 2555     3417     3    3.5    51             0
## 2556     3417     3    3.5    50             0
```

To make sure that the analyzed variables are not each other's linear expression, we rule out collinearity between the two independent variables. Correlation:

```
## [1] -0.1120202
```

We can see that there is an inverse correlation but they are not completely collinear.

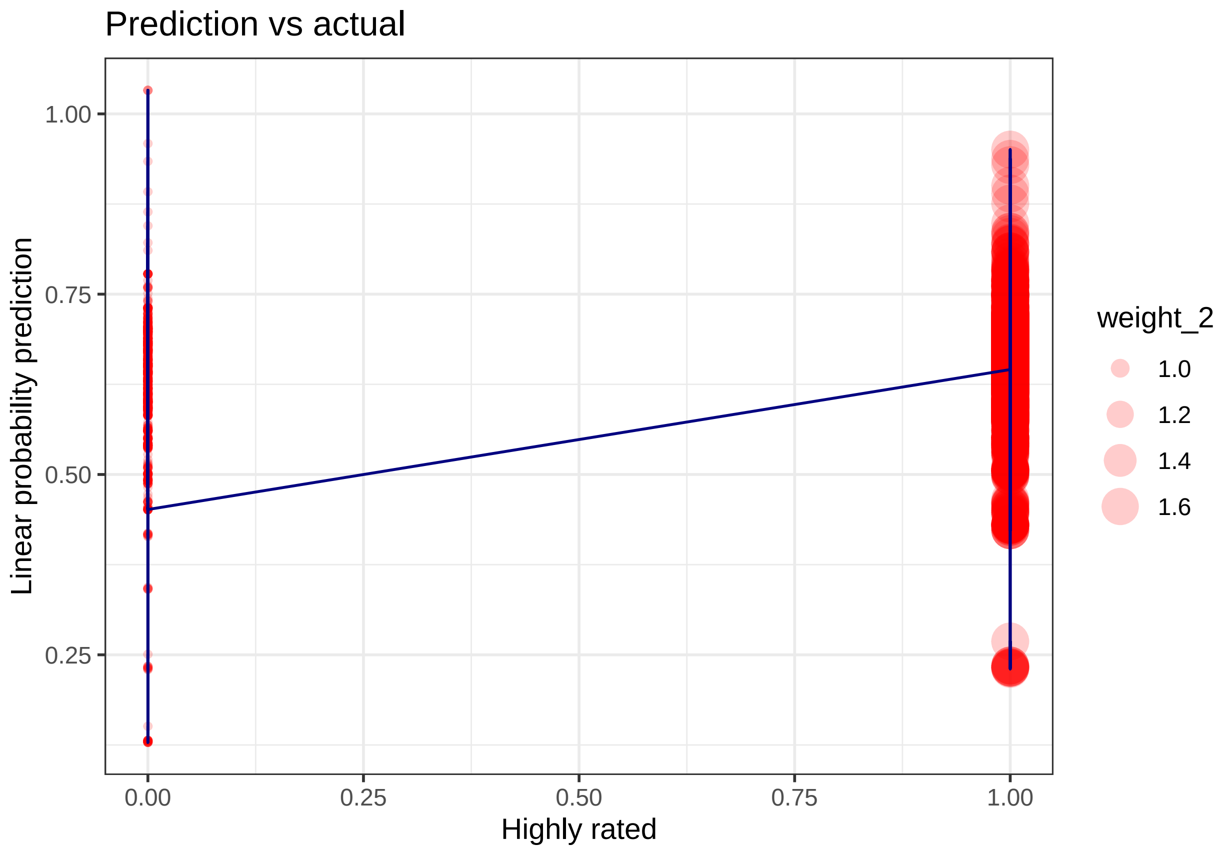
Linear probability model

```
## OLS estimation, Dep. Var.: highly_rated
## Observations: 2,556
## Standard-errors: Heteroskedasticity-robust
##              Estimate Std. Error  t value  Pr(>|t|)
## (Intercept)  0.713262   0.017164 41.55488 < 2.2e-16 ***
## distance    -0.099674   0.010442 -9.54532 < 2.2e-16 ***
## price        0.000146   0.000054  2.70792 0.0068158 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 0.471364   Adj. R2: 0.036652
```

Coefficient shows that hotels in the same price category are 9% less likely to be highly rated (4 or more) with the distance increasing. Coefficient for price has a relatively higher SE, and lower significance than the distance variable.

```
## OLS estimation, Dep. Var.: highly_rated
## Observations: 2,556
## Standard-errors: Heteroskedasticity-robust
##              Estimate Std. Error  t value  Pr(>|t|)
## (Intercept)  0.741748   0.013645  54.3597 < 2.2e-16 ***
## distance    -0.103966   0.010324 -10.0701 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 0.472446   Adj. R2: 0.032604
```

LPM plot prediction vs actual



Non-linear probability

Logit

Logit model:

```
## GLM estimation, family = binomial(link = "logit"), Dep. Var.: highly_rated
## Observations: 2,556
## Standard-errors: IID
##              Estimate Std. Error  t value   Pr(>|t|)
## (Intercept)  0.889225   0.079062 11.24720  < 2.2e-16 ***
## distance    -0.436499   0.053476 -8.16247 3.2824e-16 ***
## price        0.000770   0.000227  3.38321 7.1644e-04 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Log-Likelihood: -1,624.3   Adj. Pseudo R2: 0.027507
##            BIC:  3,272.2     Squared Cor.: 0.039642
```

Probit

Probit model:

Table 2: Logit / Probit comparison

	min	P25	Median	Mean	P75	Max
pred2	0.16	0.60	0.65	0.64	0.69	0.93
pred3	0.15	0.60	0.66	0.64	0.69	0.92

```
## GLM estimation, family = binomial(link = "probit"), Dep. Var.: highly_rated
## Observations: 2,556
## Standard-errors: IID
##              Estimate Std. Error  t value  Pr(>|t|)
## (Intercept)  0.564091   0.047342 11.91516 < 2.2e-16 ***
## distance    -0.270738   0.032044 -8.44892 < 2.2e-16 ***
## price        0.000393   0.000129  3.04962 0.0022913 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Log-Likelihood: -1,625.0   Adj. Pseudo R2: 0.027115
##            BIC:  3,273.5     Squared Cor.: 0.038705
```

Marginal differences logit & probit

```
## Call:
## logitmfx(formula = highly_rated ~ distance + price, data = df,
##     atmean = FALSE, robust = T)
##
## Marginal Effects:
##                dF/dx   Std. Err.       z     P>|z|
## distance -9.6963e-02  1.2493e-02 -7.7611 8.421e-15 ***
## price     1.7096e-04  7.3685e-05  2.3202   0.02033 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Call:
## probitmfx(formula = highly_rated ~ distance + price, data = df,
##     atmean = FALSE, robust = T)
##
## Marginal Effects:
##                dF/dx   Std. Err.       z   P>|z|
## distance -9.8402e-02  1.1068e-02 -8.8910 < 2e-16 ***
## price     1.4297e-04  6.9219e-05  2.0655 0.03888 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Coefficient are similar not only to each other, with 9.69% and 9.84% probability of high rating decrease in the same price category when the distance form the city center is decreasing.

Comparing logit and probit

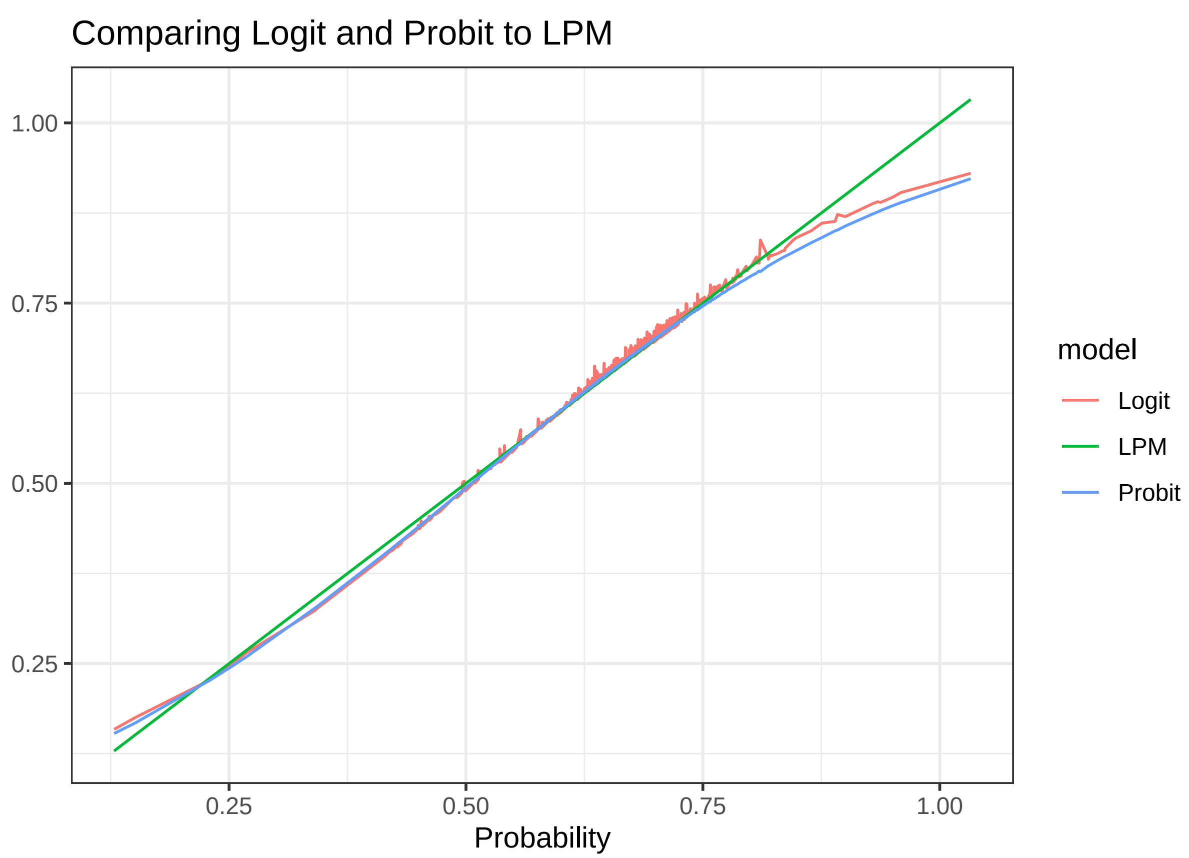
Logit and probit has very similar statistical characteristics. Since the difference is very little, either of these two fits the purpose and can been chosen an nonlinear probability model. (Table 2:Logit / Probit comparison)

Summary

```
##                                      lpm               logit              probit
## Dependent Var.:          highly_rated        highly_rated        highly_rated
##
## Constant          0.7133*** (0.0172)  0.8892*** (0.0791)  0.5641*** (0.0473)
## distance         -0.0997*** (0.0104) -0.4365*** (0.0535) -0.2707*** (0.0320)
## price               0.0001** (5.4e-5)  0.0008*** (0.0002)   0.0004** (0.0001)
## _______________ ___________________ ___________________ ___________________
## Family                            OLS               Logit              Probit
## S.E. type         Heteroskedast.-rob.                 IID                 IID
## Observations                    2,556               2,556               2,556
## Squared Cor.                  0.03741             0.03964             0.03871
## Pseudo R2                     0.02779             0.02870             0.02831
## BIC                           3,432.3             3,272.2             3,273.5
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since logit and probit are nonlinear, coefficient is harnder ot interpret, as the slope of the function changes depending on the x. however, the marginal difference of them is similar to the linear probability model. Distance v ariable has a high significance in all of the models. Pseudo R2is the highest of the logit model.

Plotting comparison of the three models



Lower and higher probabilities are different in LPM compared to logit & probit, but are hardly distinguishable in the mid range. Since logit and probit are nonlinear, coefficient is harder to interpret, as the slope of the function changes depending on the x.