Nils Baker Case Study

Chaman Preet Kaur and Chris Atterbury

October 15, 2015

Contents

1 Executive Summary

This section is to be written last and contains a short (approx 250 words) summary of the case study as a whole.

2 Background

Summarizes the case study prompt. Contains any basic term definitions and prepares reader for discussion. Should propose basic question so that the Models section can dive right in.

3 Models

3.1 Initial Considerations

The data that we have received to pursue Nils Baker's hypothesis includes four columns. Initially, they are *ID*, *Total.Households.in.Area*, *Households.with.Account*, and *Inside.Outside.Footprint*. First, we drop the last two rows of the data set because they are empty and change the column names to *ID*, *Total.Households*, *Accounts*, and *Footprint*, respectively. Next, further preprocessing of the data is performed so that we have numeric data (see **Additional Codeblock 1** in the *Appendix* for the code used to do this).
The first column, *ID*, is identical to the row number and, thus, will not be considered in this case study. Each row itself is a separate Metropolitan Statistical Area (MSA). For our purposes, we need only understand

that each row contains a diferent geographic region. The next two columns, *Total.Households* and *Accounts*, contain the number of total households and the number of those households that have a checking account with the bank. The last column, *Footprint*, was originally coded with either "Inside" or "Outside." A region was considered "Inside" if there was both a physical bank location and an ATM in the region. This was recoded as 1 for easier analysis of the data. A region was considered "Outside" if there was not a physical bank location and only an ATM in the region. This was recoded as 0.

Finally, we will add a column called *Accts.Hsehld* that takes the *Accounts* column and divides it by the *Total.Households* column. We may want to choose this as the response variable since this allows us to get closer to evaluating Nils Baker's hypothesis without as much interpretation of the model.

With the data in a usable form, we can now view the correlation matrix to get an initial feel for the relationships amongst the variables.

```
##                          ID Total.Households    Accounts  Footprint
## ID                1.0000000      -0.67957529 -0.66181262  0.3030236
## Total.Households -0.6795753       1.00000000  0.91121152 -0.3004534
## Accounts         -0.6618126       0.91121152  1.00000000 -0.2171381
## Footprint         0.3030236      -0.30045335 -0.21713807  1.0000000
## Accts.Hsehld      0.1732239      -0.08685308  0.07169036  0.1523331
##                  Accts.Hsehld
## ID                 0.17322388
## Total.Households  -0.08685308
## Accounts           0.07169036
## Footprint          0.15233308
## Accts.Hsehld       1.00000000
```

We can ignore the values involving `ID` since this is basically just a running counter of which number region we see. Its correlation with `Total.Households` is based on the fact that the data set's observations are organized by the number of households in the region with the largest first. We see that `Accounts` has a strong correlation

3.2 Procedure

The goal in this case study, as stated by Nils Baker, is to ascertain whether or not the presence of a physical bank in a region increases the likelihood of a given household possessing a checking account. We will start with Simple Linear Regression (SLR) models, before considering more complex Multiple Linear Regression (MLR) models.

3.3 SLR Models

Details how we arrived at the models that we tried and how they worked. Include details about both the processes and the model to which they led. Our thought processes should be detailed so that there is no question how we got to our models. This needs to do all of the leg work so that the Conclusions section can focus on the actual meaning of the model. *This section should include subsections for each model with two hashes and then further subsectioniong using three hashes for each part of the model discussion.*

4 Conclusions

The goal is that everything is built up to this point so that little we can just plow right into the meaning of the model. Other general conclusions can be included.

5 Appendix

5.1 Additional Codeblock

Additional Codeblock 1

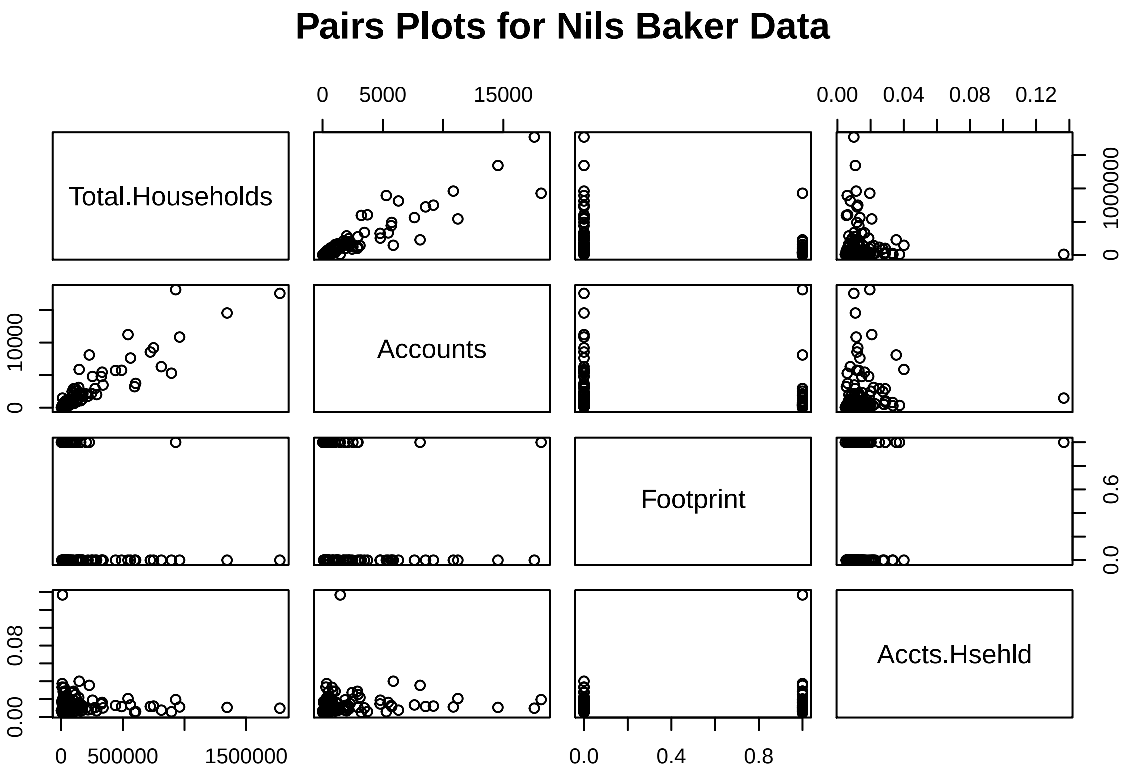
```
d <- read.csv("41330723.csv", header = TRUE, stringsAsFactors = FALSE)
d <- d[1:120, ] # last two rows contain no data
names(d) <- c("ID", "Total.Households", "Accounts", "Footprint")

for (i in 2:3) {
  d[[i]] <- gsub(",", "", d[[i]])
  d[[i]] <- as.numeric(d[[i]])
}
d[["Footprint"]][d[["Footprint"]] == "Outside"] <- 0
d[["Footprint"]][d[["Footprint"]] == "Inside"] <- 1
d[["Footprint"]] <- as.numeric(d[["Footprint"]])
d[["ID"]] <- as.numeric(d[["ID"]])
d[["Accts.Hsehld"]] <- d[["Accounts"]] / d[["Total.Households"]]
```

```
head(d)
```

```
##   ID Total.Households Accounts Footprint Accts.Hsehld
## 1  1          1772960    17563         0  0.009906033
## 2  2          1345209    14547         0  0.010813933
## 3  3           960434    10847         0  0.011293853
## 4  4           928274    18133         1  0.019534103
## 5  5           893995     5291         0  0.005918378
## 6  6           812137     6297         0  0.007753618
```

5.2 Additional Figures

Additional Figure 1


Pairs Plots for Nils Baker Data
Total.Households
Accounts
Footprint
Accts.Hsehld
0
5000
15000
0.00
0.04
0.08
0.12
0
1000000
0
10000
0.0
0.6
0.00
0.08
0
500000
1500000
0.0
0.4
0.8