# 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 with `Total.Households`, which is reasonable since we expect that regions with more households will have more accounts, just based on volume. Also, this explains the negative correlation between `Accounts` and `ID`. The numbers indicate what we would expect for `Footprint`; the correlations with the households and accounts do not show anything. That is, the locations of physical banks were not chosen based on the number of households or accounts in a region. This correlation matrix ends up showing us that trying to predict `Accts.Hsehld` may be difficult, but hopefully combining a few features will help. Please see **Additional Figure 1** in the *Appendix* for the pairs plot that shows these relationships graphically.

## 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

### 3.3.1 Predicting Number of Accounts

We will start by evaluating SLR models for `Accounts`. Even though these models may be more difficult to interpret in terms of Nils Baker's hypothesis, if we find an especially good model, then we may make an exception. The summaries of these three models are below.

```
##
## Call:
## lm(formula = Accounts ~ Total.Households, data = d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4645.8  -358.6  -223.9    77.7  7823.8
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      2.260e+02  1.447e+02   1.562    0.121
## Total.Households 1.086e-02  4.521e-04  24.029   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1366 on 118 degrees of freedom
## Multiple R-squared:  0.8303, Adjusted R-squared:  0.8289
## F-statistic: 577.4 on 1 and 118 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = Accounts ~ Footprint, data = d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2540.2 -1778.4  -819.1   -40.1 16943.4
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2627.2      395.4   6.645 9.86e-10 ***
## Footprint    -1437.6      594.9  -2.416   0.0172 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3236 on 118 degrees of freedom
## Multiple R-squared:  0.04715,    Adjusted R-squared:  0.03907
## F-statistic: 5.839 on 1 and 118 DF,  p-value: 0.01721
```

```
##
## Call:
## lm(formula = Accounts ~ Accts.Hsehld, data = d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2671.2 -1594.0 -1433.4   143.6 16047.8
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1742.8      439.6   3.965 0.000126 ***
## Accts.Hsehld  17527.1    22448.6   0.781 0.436504
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3307 on 118 degrees of freedom
## Multiple R-squared:  0.00514,    Adjusted R-squared:  -0.003292
## F-statistic: 0.6096 on 1 and 118 DF,  p-value: 0.4365
```

Considering the results of the correlation matrix above, none of these results are particularly surprising. The total number of households in a region proves to be a good predictor of the number of accounts in the region. The coefficient of determination, denoted $R^2$, is high, so we know that a significant amount of the variation in `Accounts` is explained by `Total.Households`. Additionally, the p-value for the coefficient and the model are significant. However, this model only serves to highlight the drawbacks of choosing `Accounts` as the response variable since we can only determine that regions with more households have more accounts. We can say nothing about how likely a given household is to have an account.

The other two models, which use `Footprint` and `Accts.Hsehld`, are exceptionally poor. `Footprint` is the feature that most interests us since it tells us if a physical bank is in the region. While we would have been surprised to see the presence of a physical bank be a good predictor of the gross number of accounts, seeing that it is not fits in with what we would expect. Now we will look at SLR models predicting the number of accounts per household in a region.

### 3.3.2 Predicting Accounts per Household

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**

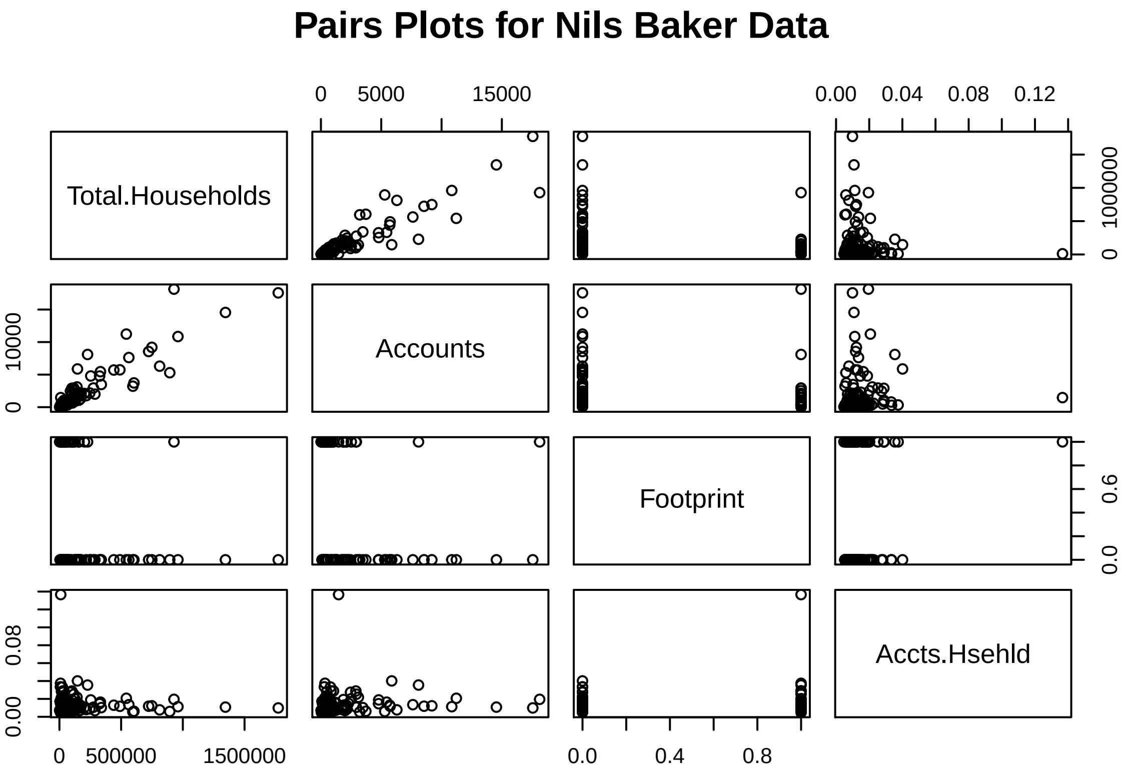
```r
d <- read.csv("41330723.csv", header = TRUE, stringsAsFactors = FALSE)
d <- d[1:120, ] # last two rows contain no data
names(d) <- c("ID", "Total.Households", "Accounts", "Footprint")

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