Case Study 1: Nils Baker

MSAN 601

Justin Midiri, Sheri Nguyen

October 5, 2016

1 Executive Summary

Nils Baker, vice president of a regional bank in the U.S., is interested in whether or not the physical presence of a bank increases demand for his particular banks checking account service. Through our analysis, we found that the presence of a physical bank branch does not necessarily imply a larger demand for checking accounts. We did, however, find that knowing the number of households in a particular area can help estimate the expected demand for checking accounts. We found a strong linear relationship between the total number of households in a particular area and the total number of checking accounts in that area. If the total number of households increases by 1% we may see close to a 1% increase in expected households with checking accounts. Although the total number of households in a particular area explains 83% of the variation in the expected demand for checking accounts, we feel that the model can benefit from adding more variables. A stronger model may include the number of accounts held by a particular household, account holder income and other significant variables.

2 Introduction

The following is a case study to examine if the presence of a physical bank branch impacts the demand for checking accounts for a large regional retail bank in the United States, using the Nils Baker dataset.

3 Detailed Processes

3.1 Sample and Testing Criteria

The sample dataset consists of 120 Metropolitan Statistical Areas (MSAs). The variables in this case study are defined as follows:

Independent Variables:

- **TotalHH** : The total number of households that reside in a particular MSA
- **InOutFootprint** : Whether there is a physical branch in that particular MSA
  - If there is a physical location, that location is said to have an "inside footprint" and is assigned a value of "1"
  - Otherwise it is said to have an "outside footprint" and assigned a value of "0"

Response Variable:

- **HHAccounts** : Number of households with a checking account through Nils' Baker's bank

Please note that all MSAs contained the bank's ATMs regardless if there was a physical branch or not. All tests in this study were compared to significance level of $\alpha = .05$. We declared all p-values less than α to be statistically significant.

3.1.1 Preliminary Examination

We began our investigation by looking at how our variables relate to one another through the use of pairwise correlation matrices and tables (Appendix 4.1.1-4.1.2). This enabled us to see the relationship between the independent variables ($TotalHH$ and $InOutFootprint$) and the response variable ($HHAccounts$) as well as any relationships among the independent variables with each other.

First, we checked how the data is distributed for each type of footprint. We made a faceted graph (Appendix 4.1.3) and observed that the distribution for both inside and outside was very similar for MSAs with lesser households; however, we noticed that as the number of households increased, we had much fewer data points for inside footprints. This may be due to a data gathering error or may suggest that an inside footprint doesn't necessarily mean that there exists a correlation between demand for checking acounts and the type of footprint. Taking this into consideration, we examined the relationship between $TotalHH$ and $HHAccounts$.

Through our pairwise correlation matrices, we found there existed a strong correlation between $TotalHH$ and $HHAccounts$. To confirm this, we analyzed the histograms of each variable and found that they were both right-skewed. However, this made sense and indicated a strong relationship between the two (Appendix 4.1.4-4.1.5). Since the data is not normally distributed, we needed to transform our data. After transforming the data, we noticed that the residuals were better distributed (Appendix 4.1.6-4.1.7). For the rest of the study, we used $log(HHAccounts)$ and $log(TotalHH)$. We also reran our faceted graph to see if the distribution of our datapoints looked less skewed. As expected, the results of the log transforms were much better (Appendix 4.1.8) and we no longer saw the gaps in data for the inside footprint.

3.1.2 Building the Model

We decided to use the backward stepwise technique, by including all variables provided to us. Our first model (Appendix 4.2.1) produced the following regression line:

$$log(\widehat{HHAccounts}) = -4.35645 + 0.98521 * log(TotalHH) + 0.18519 * InOutFootprint \quad (1)$$

The coefficient for $InOutFootprint$ is not statistically significant at our defined $\alpha = .05$. Before we decided to exclude $InOutFootprint$ from our model, we checked the residual plots (Appendix 4.2.2) and they looked good, despite a Breusch Pagan test suggesting heteroskedasticity with a p-value of 0.03636. We suspected the heteroskedasticity was due to endogeneity, or omitted variables bias. We ran a Ramsey RESET test to see if we may be perhaps missing any interaction terms, and obtained a p-value of 0.3714, meaning we are not missing any terms.

We, then, proceeded to look at the model without $InOutFootprint$ (Appendix 4.2.3) and derived the following regression:

$$log(\widehat{HHAccounts}) = -4.00299 + 0.96075 * log(TotalHH) \quad (2)$$

We had previously ran the residual plots for this regression (Appendix 4.1.6) and the distribution of the residuals looked good. When we ran the Breusch Pagan test for this model, we got a p-value of 0.02112. For the sake of interpretability, we did not feel that it was necessary to perform any transformations, such as Box-Cox, on these variables to attempt to fix the heteroskedasticity and concluded again that this problem may be contributed to omitted variable bias. Which intuitively made sense since we only had one regressor and believed that there is undoubtedly more than one factor in determining the demand for checking accounts.

3.2 Conclusion

We inferred that our best model was model (2) for the reasons that follow:

- Easy interpretability
- High explanation power of variation in expected $HHAccounts$ (${R_a}^2 = 0.8348$)
- Statistically significant coefficient for all variables

We acknowledged that there was heteroskedasticity in this model, however, after running countless transformations, we decided to attribute the issue to endogenity, or omitted variable bias. Otherwise, we believe that this is the best model to explain the expected variation in $HHAccounts$ given the data we received. With this model, we may interpret that a 1% increase in TotalHH will result in an expected increase in HHAccounts by 0.9608% (Appendix 4.2.4). Due to the fact that adding $InOutFootprint$ to our models showed that $InOutFootprint$ is not statistically significant at the $\alpha = .05$ level, we concluded that the type of footprint does not positively or negatively impact the expected demand for checking accounts. Therefore to answer Nils Baker's inquiry, we say there was no data that confirmed a physical presence of a bank branch creates demand for a checking account.

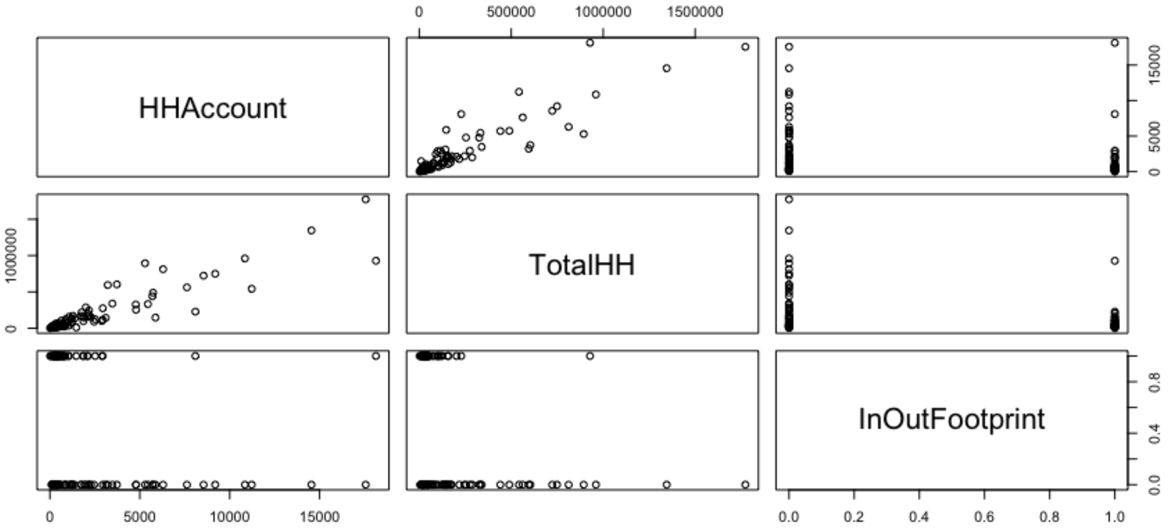
For further analysis, we would like to suggest obtaining more variables, perhaps account holder income, employment status, and family size. We feel that account holder income would be a very influential predictor. We consider that there may be individuals/families who may be suffering from economic hardship and may not have enough money to maintain an account. This leads to another predictor variable that we felt could also be a significant predictor for the data, employment status. Much like account holder income, employment status of an individual may also contribute to the ability of an individual to maintain an account. If an individual is unemployed, it is likely that they will not have an adequate source of income to maintain an account. Another variable, family size, could also have an effect on the number of accounts a particular household might have. We felt that larger families with more children may have multiple bank accounts with different purposes (i.e.

retirement, education, etc.). By including additional significant variables to our model, we may be able to reduce the hetoskedasticity from our model and account for more of the variation in expected $HHAccounts$.

4 Appendix

4.1 Preliminary Observations

4.1.1

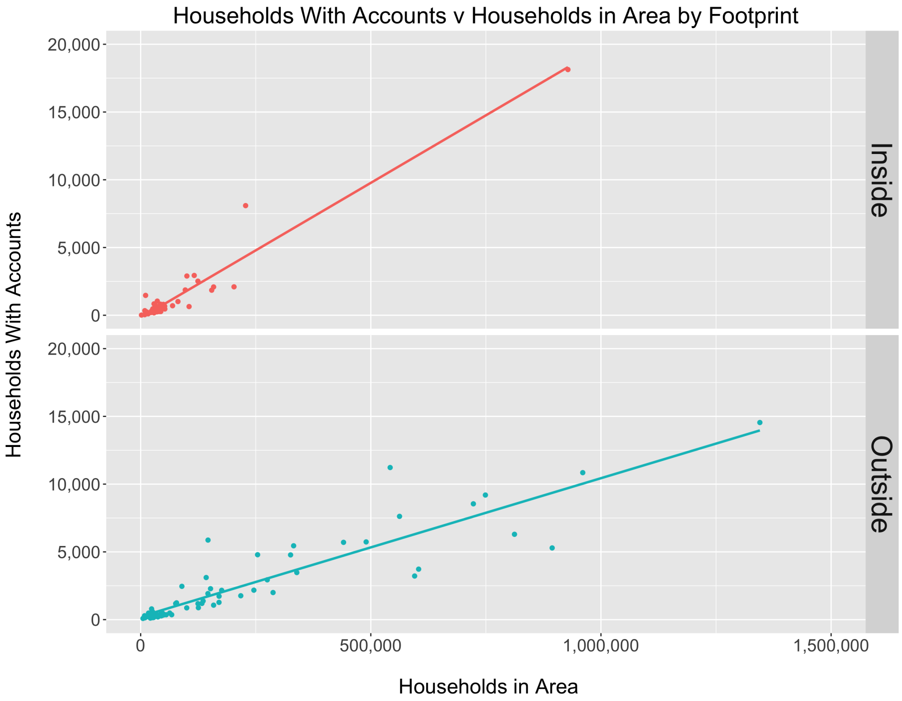


4.1.2

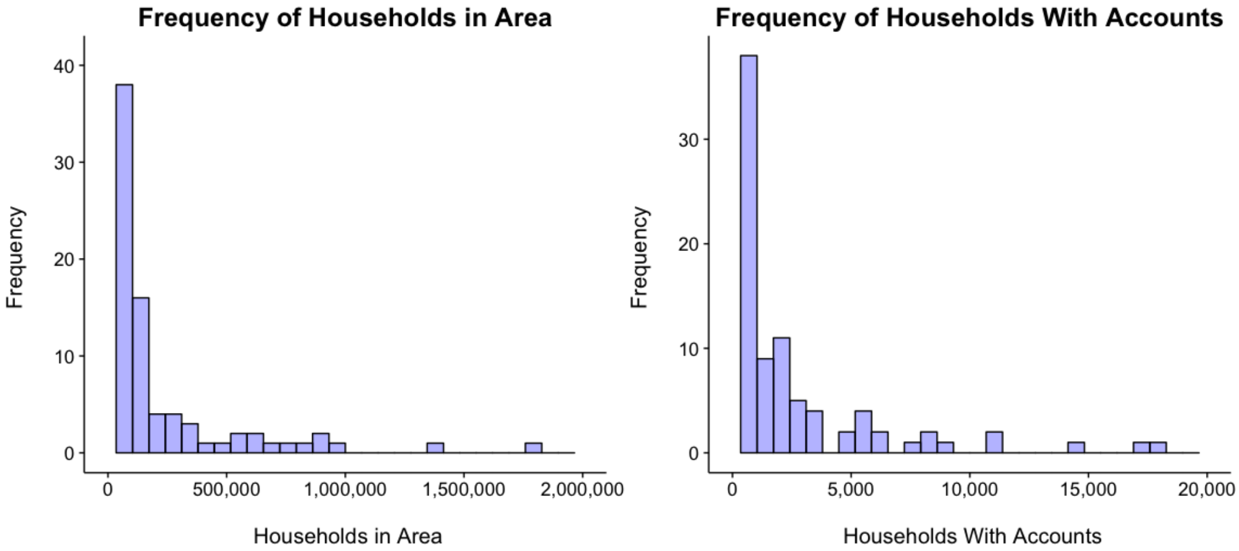
	HHAccount	TotalHH	InOutFootprint
HHAccounts	1.0000000	0.9112115	-0.2171381
TotalHH	0.9112115	1.0000000	-0.3004534
InOutFootprint	-0.2171381	-0.3004534	1.0000000

Table 1: Correlation Matrix

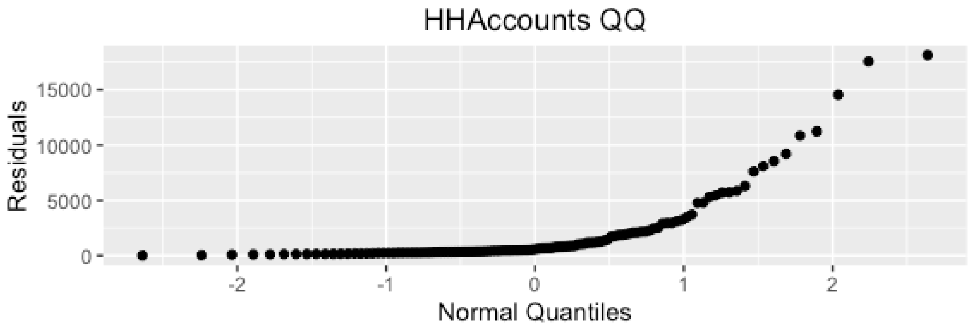
4.1.3

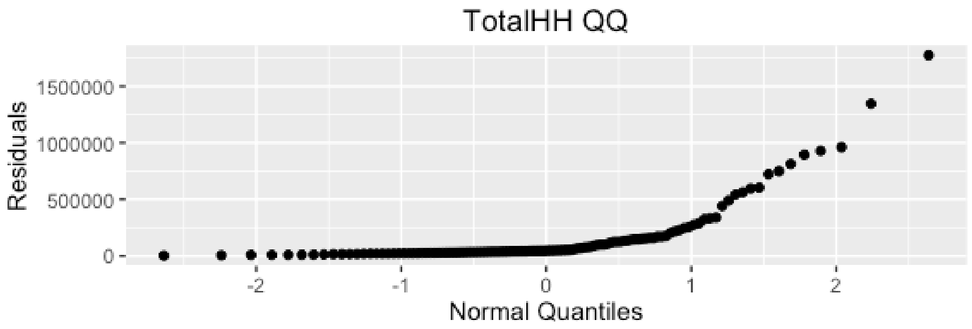

Households With Accounts v Households in Area by Footprint
Households With Accounts
20,000
15,000
10,000
5,000
0
Inside
Outside
0
500,000
1,000,000
1,500,000
Households in Area

4.1.4

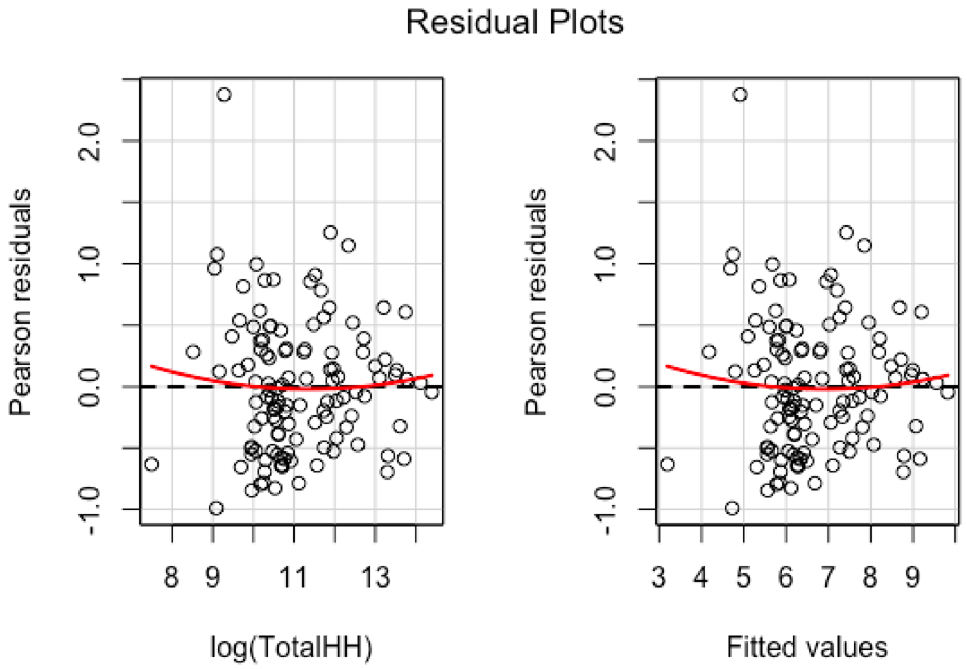

Frequency of Households in Area
Frequency
40
30
20
10
0
0
500,000
1,000,000
1,500,000
2,000,000
Households in Area
Frequency of Households With Accounts
Frequency
30
20
10
0
0
5,000
10,000
15,000
20,000
Households With Accounts

4.1.5

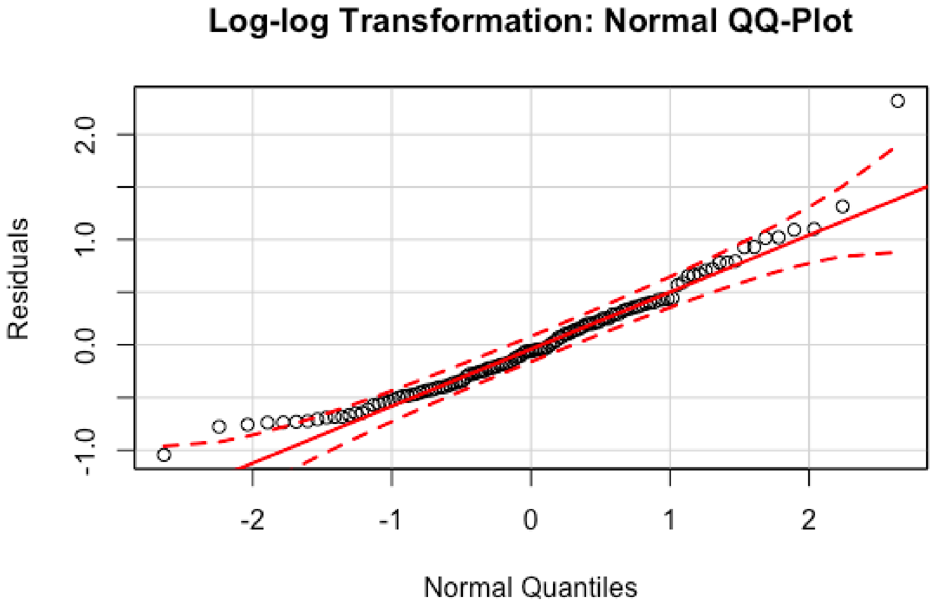

HHAccounts QQ
Residuals
15000
10000
5000
0
-2
-1
0
1
2
Normal Quantiles


TotalHH QQ
Residuals
1500000
1000000
500000
0
-2
-1
0
1
2
Normal Quantiles

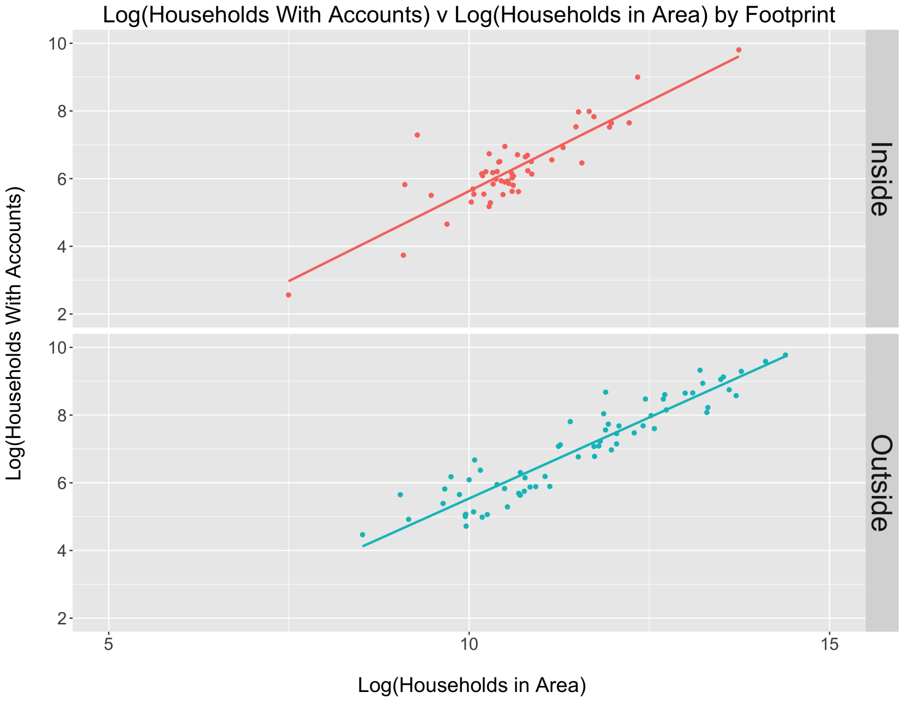
4.1.6


Residual Plots
Pearson residuals
2.0
1.0
0.0
-1.0
8
9
11
13
log(TotalHH)
Pearson residuals
2.0
1.0
0.0
-1.0
3
4
5
6
7
8
9
Fitted values

4.1.7


Log-log Transformation: Normal QQ-Plot
Residuals
2.0
1.0
0.0
-1.0
-2
-1
0
1
2
Normal Quantiles

4.1.8


Log(Households With Accounts) v Log(Households in Area) by Footprint
Log(Households With Accounts)
10
8
6
4
2
Inside
Outside
5
10
15
Log(Households in Area)

4.2 Building the Model

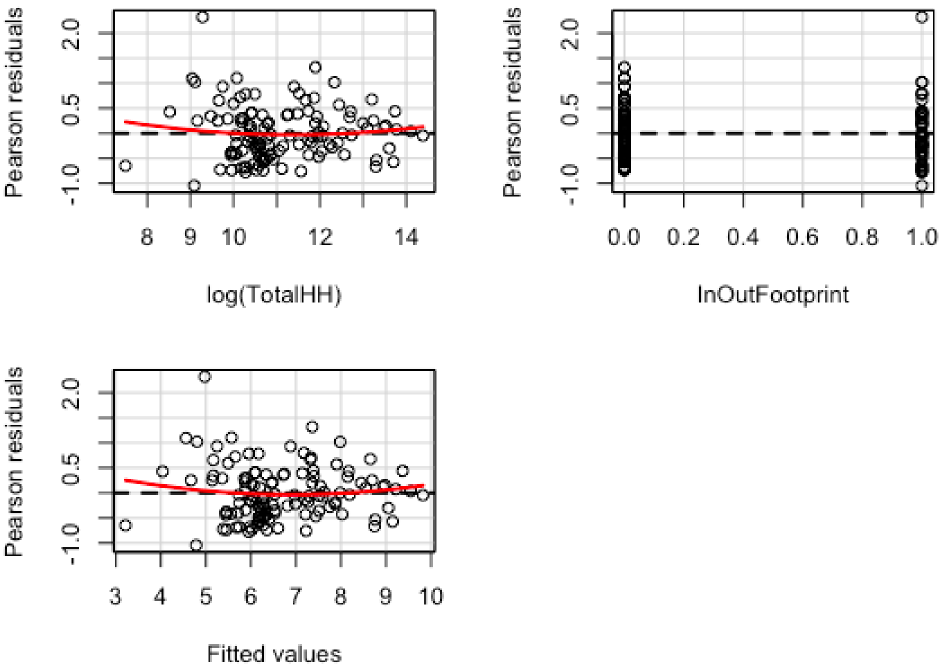
4.2.1

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-4.3564	0.4791	-9.09	0.0000
log(TotalHH)	0.9852	0.0413	23.87	0.0000
InOutFootprint	0.1852	0.1063	1.74	0.0840

Residual standard error: 0.5436 on 117 degrees of freedom
Multiple R-squared: 0.8403, Adjusted R-squared: 0.8376
F-statistic: 307.8 on 2 and 117 DF, p-value: < 2.2e-16

Table 2: log(HHAccounts) ∼ log(TotalHH) + InOutFootprint

4.2.2 Residual Plots for log(HHAccounts) ∼ log(TotalHH) + InOutFootprint



4.2.3

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-4.0030	0.4378	-9.14	0.0000
log(TotalHH)	0.9608	0.0391	24.54	0.0000

Residual standard error: 0.5483 on 118 degrees of freedom
Multiple R-squared: 0.8362, Adjusted R-squared: 0.8348
F-statistic: 602.2 on 1 and 118 DF, p-value: < 2.2e-16

Table 3: log(HHAccounts) ∼ log(TotalHH)

4.2.4 Interpreting Our Final Model

$$\begin{aligned} log(HHAccounts) &= -4.0030 + 0.9608 * log(TotalHH) \\ \frac{\partial}{\partial HHAccounts} &= \frac{\partial TotalHH}{totalHH} * 0.9608 \\ 100 * \frac{\partial}{\partial HHAccounts} &= \frac{\partial TotalHH}{totalHH} * 0.9608 * 100 \\ \%\Delta HHAccounts &= \%\Delta TotalHH * 0.9608 \end{aligned} \tag{3}$$

Therefore, a 1% increase in TotalHH, will result in a HHAccounts incrase by 0.9608%.