Nils Baker Case Study Report

Jeff Baker Leighton Dong

September 30, 2014

Executive Summary

Nils Baker posed the question, "Does the presence of a physical bank branch create demand for checking accounts?" The intuitive hypothesis was that having a physical branch in the area should lead to increased demand for accounts. In order to test this hypothesis, we examined data on 120 metropolitan areas, which included A) the total number of households, B) households with accounts, and C) whether a physical bank branch resided in these areas or not.

We began with a preliminary examination of each of the variables individually, to ensure the data were clean and realistic. Next, we generated three linear models to determine if any model would answer the question. A series of diagnostic tests were performed on each model. From the results, it was clear that linear regression was not the best method for answering Mr. Baker's question, and was thus abandoned.

As a final attempt to address the question, an ANOVA test was conducted. This entailed splitting the data into two groups (areas with bank branches, and areas without) and comparing the average proportions of households that have bank accounts. We found the averages to be statistically equal between groups–having a bank did not appear to make a difference in bank account demand. Thus, from this data, we conclude that the presence of a local bank branch does not create demand for checking accounts. If the executive team would like to further investigate this topic, some data recommendations are provided to support any future analysis.

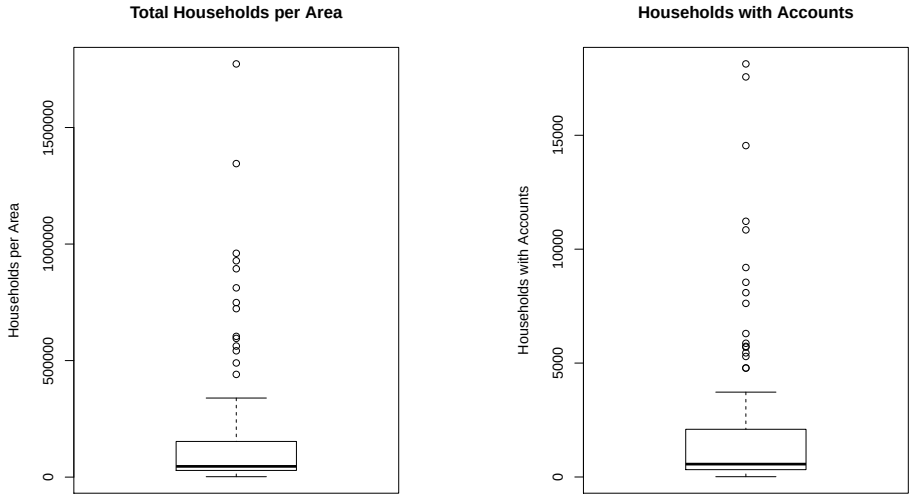
Data Analysis

Basic Diagnostics

We began with a straightforward, univariate data exploration exercise, examining each of the variables within the data set independently. For this study, we defined total households in the area as X_1, and whether or not a bank was inside the Metropolitan Statistical Area (MSA) as dummy variable X_2; $X_2 = 1$ if a bank was inside the MSA, and $X_2 = 0$ if a bank was outside the MSA (Appendix Table 2).

We noted a large range in the number of households within these MSAs; this is noteworthy, since the presence of a bank branch in a smaller town may or may not have a more pronounced impact on account sign-ups than in a bigger MSA with potentially more bank choices. Thus, the data appeared diverse enough to justify further pursuit of a potential statistical solution.

After plotting the data, we discovered our first finding: the distribution for each variable was clearly skewed (Figure 1).



(a) Total Households per Area (X_1) (b) Households with Accounts (Y)

Figure 1

A scatter plot of the variables revealed (at first glance) that a linear relationship exists, and a linear model might be appropriate in answering the question posed by Mr. Baker (Appendix Figure 5). The correlation coefficient ($r = 0.91$) corroborates this finding.

Linear Regression Models

1. Linear Regression Model of Data

After examining these variables, we then generated a linear model representing the entire data set, regressing households with accounts (Y) on X_1 and dummy variable X_2. The interaction term between X_1 and X_2 was also included (Appendix Table 3).

At the $\alpha = 0.05$ level of significance, the resulting model was statistically significant (p-value $= 2.2 \times 10^{-16}$), explaining 89% of the variation in Y ($R^2 = 0.8923$); additionally, b_1 and b_3 (the interaction term coefficient) were highly significant (p-values of 2×10^{-16} and 1.48×10^{-12}, respectively). We also note the large, negative coefficient of the dummy variable (b_2) in this model; this is inconsequential, since this value simply moves the intercept of the model. We also noted b_2 to be statistically insignificant (p-value $= 0.063$) as we continued validating the model through other diagnostic measures.

Next, plotting the fitted values against residuals revealed clear heteroskedasticity in the data (Appendix Figure 7). Additionally–and perhaps more importantly–a sequence plot of the residuals shows a distinct pattern within the plot: a strong indicator that the model is likely invalid for this case (Figure 2). A Breusch-Pagan test for heteroskedasticity confirmed this suspicion: $\chi^2_{BP} = 11.9965$, with a p-value$_{BP}$ of 7.395×10^{-3}. Since the p-value$_{BP}$ < p-value$_{crit}$ at our chosen 5% level of significance, we reject the null hypothesis that the variance is constant.

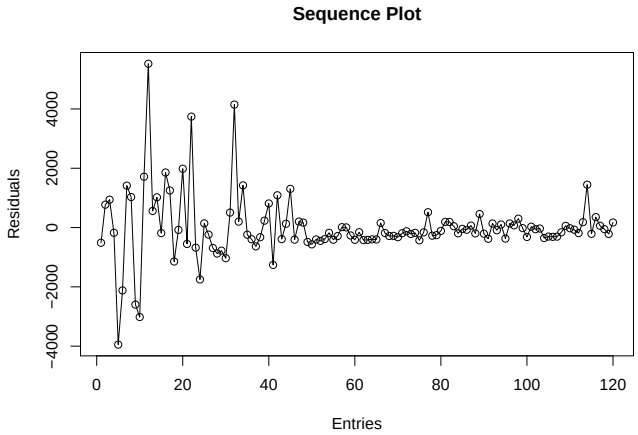


Figure 2: Residuals Sequence Plot

Additionally, a check on normality of the data was performed by inspecting the Q-Q plot of the MLR: the heavy-tailed symmetrical pattern indicates the data are not normally distributed (Appendix Figure 6). A Shapiro-Wilk normality test on the residuals returned a p-value of 2.749×10^{-12}; this is much smaller than our 5% significance level, so we reject the null hypothesis that the data are normally distributed. This further supported the invalidity of the model in it's current form.

2. Transformed Linear Model

As a second approach to improving this model, we conducted a Box-Cox transformation on our dependent variable (households with checking accounts), using a bisection search function to find the best lambda. In this case, we minimized SSE at $\lambda = 0.23$. A second round of diagnostics tests were performed to check the validity of this updated model: $\chi^2_{BP} = 12.7367$, with a p-value$_{BP}$ of 5.242×10^{-3}. Since the p-value$_{BP}$ is less than the 5% significance level, we again reject the null hypothesis that the variance is constant, despite the transformation. However, performing a Shapiro-Wilk normality test on the transformed data resulted in a p-value of 0.3779. Since this value is greater than our 5% level of significance, we actually reject the null hypothesis that the residuals are non-normal in this updated model.

The transformation corrected the non-normality found in the previous model's Shapiro-Wilk test, as expected; the new Q-Q plot displayed a much tighter group of points along the Q-Q line, indicating normality in the residuals (Appendix Figure 10). However, the residual plot revealed a distinct curved pattern, confirming heteroskedasticity within this updated model as well (Figure 3). The sequence plot also displayed a distinct pattern in the points, further proving non-independence of the variables in this model (Appendix Figure 11). Additionally, the results of the BP test corroborated with the residual plot, further confirming non-constant variance (since p-value$_{BP} = 5.242 \times 10^{-3} < 0.05$, we rejected the null hypothesis that the variance was constant).

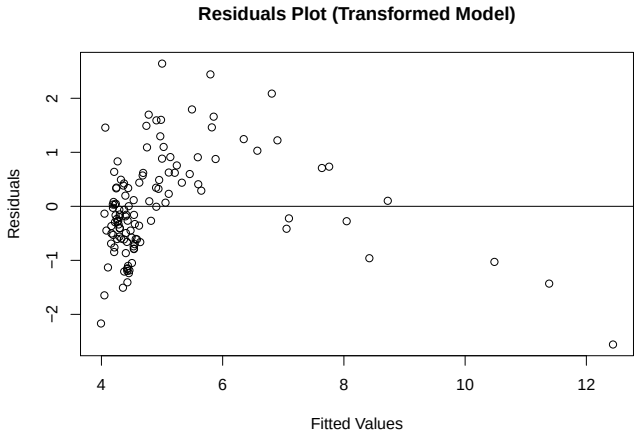


Figure 3: Fitted Values vs. Residuals

3. Robust Linear Regression Model

Having noted the presence of outliers earlier in our data exploration, we ran a robust linear regression model to see if such a model might work. The BP test results remained unchanged from previous iterations, while the Shapiro-Wilk test returned a slightly smaller p-value. This indicated a higher significance in non-normality of the residuals but, ultimately, this model did not get us any closer to answering Mr. Baker's initial question.

4. Excluding Outliers

Finally, as an exploratory last resort, we attempted to remove outlier observations from the data set. By limiting the scope of the data to a subset of MSAs that have similar attributes, we might supply an answer that was applicable to that subset. We ran a Bonferroni outlier test to detect extreme observations, cutting off at the 5% significance level. The test identified four outliers (observations 5,12, 22, and 32). We excluded these points from the data and regenerated a fourth model, running the same battery of diagnostics as the previous versions (Appendix Table 6). The only change found in this iteration was the improved significance of the dummy variable (p-value of 0.0353). However, the same symptoms as the previous modeling attempts were found (patterned residual plots, heteroskedasticity, non-normality of the residuals) (Appendix Figures 16-20). Another Box-Cox transformation was conducted on this pared-down data set; unfortunately, the same results of heteroskedasticity and non-normality were found (Appendix Figures 21-25).

Despite minor improvements from our initial model in the modified linear models, as well as attempts to exclude outliers, we concluded that none of these models were appropriate solutions for determining if local bank branches create demand for checking accounts.

Subsetting Sample Groups

Despite these results, we decided to further explore potential solutions by subsetting the data: one group of MSAs with banks inside their area and another group of MSAs without (Appendix Tables 7-8). The hypothesis at this stage is that MSAs with banks inside their region display differing attributes (namely, increase demand for checking accounts) when compared to those without bank branches.

(a) Total Households in Area (X_1) (b) Households with Accounts (Y)

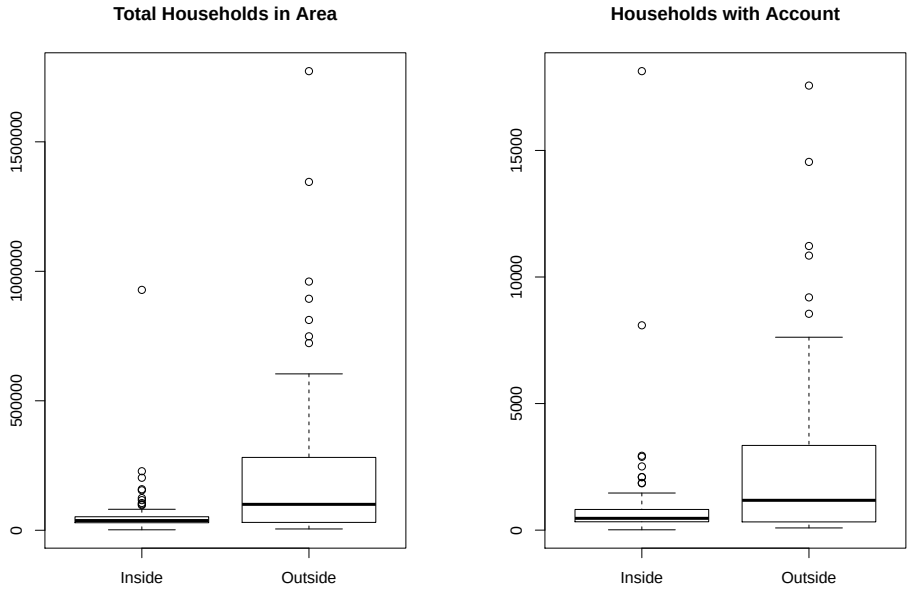


Figure 4

From the boxplots above, we still see a noticeable skew within both independent groups of data. The scatter plots also corroborate the linear relationship between both independent variables within each group (correlation coefficients $r = 0.96$ and $r = 0.93$, respectively) (Appendix Figures 26-27).

With these two distinct groups of data, we conducted an ANOVA test (unpaired t-test on difference of means with unequal variance) in order to detect any impact or difference from the presence of physical bank branches. In order to do this, we normalized the data points to more fairly compare between subsets, taking the ratio between the number of households with accounts and the total number of households in area.

Group	No. of Obs (N)	Mean ($\bar{x}$)	Variance (s^2)
1	53	0.01653	3.38263×10^{-4}
2	67	0.012411	5.46564×10^{-5}

Table 1: Summary of Groups 1 & 2

We tested the following hypothesis for groups 1 and 2:

$$H_0 : \bar{x}_1 = \bar{x}_2$$
$$H_1 : \bar{x}_1 \neq \bar{x}_2$$

For the unpaired t-test, in conjunction with the values in Table 1 above, we utilized the following formula to calculate the t and t_{crit}:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}} = 1.53758$$

$$df = v = \frac{\left(\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}\right)^2}{\frac{(s_1^2/N_1)^2}{N_1 - 1} + \frac{(s_2^2/N_2)^2}{N_2 - 1}} = 65.30194$$

$$t_{crit}(\alpha, v) = t_{crit}(0.05, 65.30194) = 1.996963$$

Since $t < t_{crit}$, we failed to reject the null hypothesis that the means of these two groups were equal. This ultimately means that, for this data set, the presence of an actual bank branch within an MSA bears no statistical difference from an MSA without a bank branch.

Conclusion and Recommendations

In summary, our findings from this data suggest that the physical presence of bank branches within these areas **does not** spur demand for checking accounts. In addition, the data and statistical inferences provided here do not imply any causal relationships that may or may not exist in reality. Considering the inherent limitations of this data set, we would suggest further data collection (primarily more variables, but not necessarily more observations). Examples of such variables might include (but are not limited to):

- How many competitor banks are in each of these regions, their customer visit frequency and volume (area share): a closer, physical branch may steal market share
- Average distance from home to any bank within any given area: as the nearest bank, customers may choose to open accounts rather than stay with a further competitor

Additionally, it may behoove the bank to conduct a qualitative survey to non-customers in each area. The results could be used to identify unmet needs that building a local bank might fulfill, such as the foreign currency exchange mentioned in Mr. Baker's situation. Other related, useful learnings could also be gleaned from this exercise, such as identifying account features that future clients may want, or creating product offerings not available through their current banks.

There may be other considerations, but these additional variables and data points would aid in developing more rigorous models to explore which, in turn, would help impact executives efforts to acquire more customers.

Appendix

	Mean	SD	Median	Min	Max
X_1	162602.67	276955.40	46082.50	1799.00	1772960.00
Y	1992.26	3301.49	565.00	13.00	18133.00
X_2	0.44	0.50	0.00	0.00	1.00

Table 2: Summary Statistics

	Estimate	Std. Error	t value	Pr($>\|t\|$)
(Intercept)	251.5372	164.3832	1.53	0.1287
X_1	0.0101	0.0004	24.98	0.0000
X_2	-445.9930	237.5807	-1.88	0.0630
X_1:X_2	0.0099	0.0012	7.93	0.0000

Table 3: Linear Model

	Estimate	Std. Error	t value	Pr($>\|t\|$)
(Intercept)	4.3320	0.1399	30.96	0.0000
X_1	0.0000	0.0000	13.35	0.0000
X_2	-0.3535	0.2022	-1.75	0.0831
X_1:X_2	0.0000	0.0000	3.21	0.0017

Table 4: Linear Model (Box-Cox Transformation)

	Value	Std. Error	t value
(Intercept)	27.8787	65.2873	0.4270
X_1	0.0107	0.0002	67.0365
X_2	-257.4899	94.3587	-2.7288
X_1:X_2	0.0089	0.0005	18.0491

Table 5: Linear Model (Robust)

	Estimate	Std. Error	t value	Pr($>\|t\|$)
(Intercept)	123.2374	111.4725	1.11	0.2713
X_1	0.0102	0.0003	36.38	0.0000
X_2	-341.9488	160.4663	-2.13	0.0353
$X_1 : X_2$	0.0090	0.0009	10.55	0.0000

Table 6: Linear Model (Excl. Outliers)

	Mean	SD	Median	Min	Max
X_1	69434.15	129139.26	37553.00	1799.00	928274.00
Y	1189.60	2672.44	466.00	13.00	18133.00

Table 7: Summary Statistics (Group 1: Areas with a physical bank branch)

	Mean	SD	Median	Min	Max
X_1	236303.15	335672.59	100089.00	5019.00	1772960.00
Y	2627.19	3619.28	1178.00	87.00	17563.00

Table 8: Summary Statistics (Group 2: Areas without a physical bank branch)

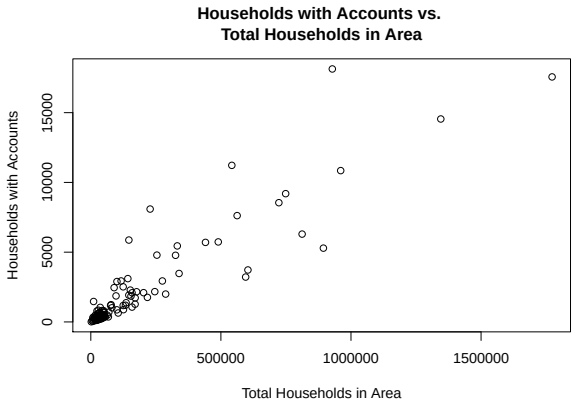


Figure 5: Aggregate Y vs. X_1

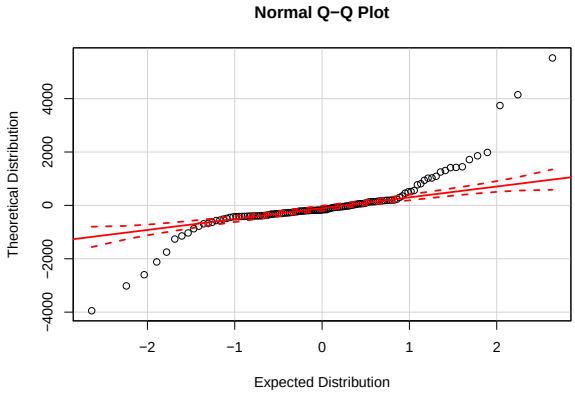


Figure 6: Aggregate Q-Q Plot

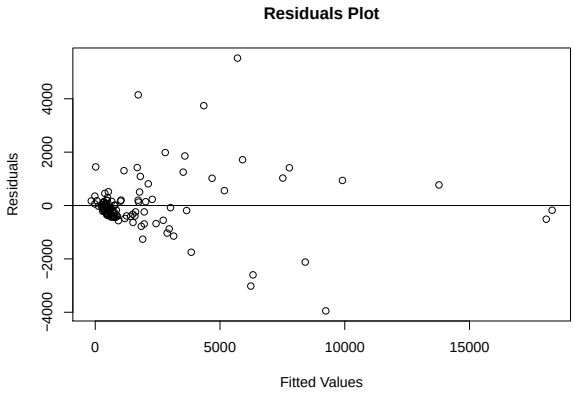


Figure 7: Fitted Values vs. Residuals

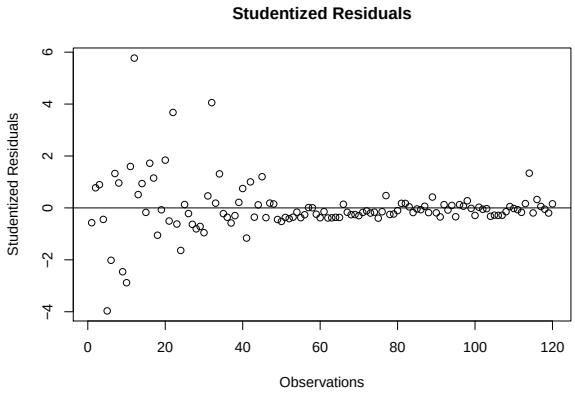


Figure 8: Studentized Residuals

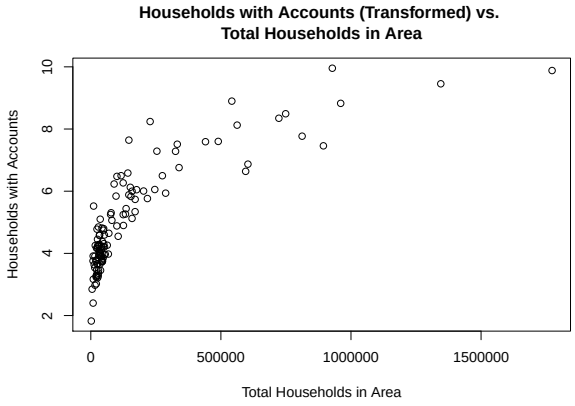


Figure 9: Transformed Y vs. X_1

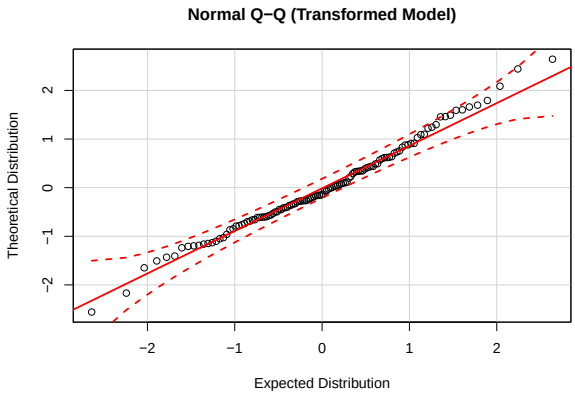


Figure 10: Transformed Q-Q Plot

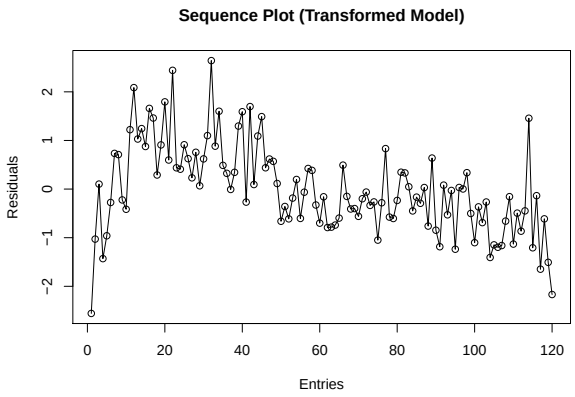


Figure 11: Transformed Sequence Plot

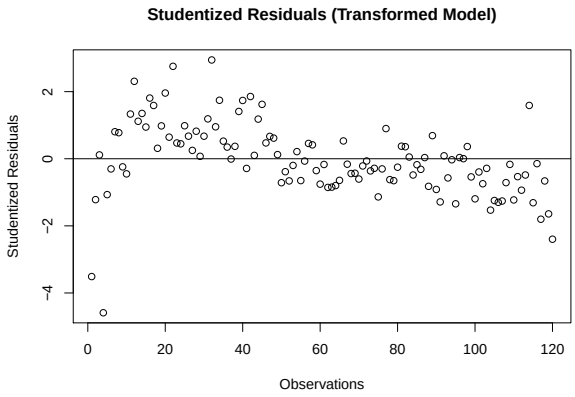


Figure 12: Transformed Studentized Residuals

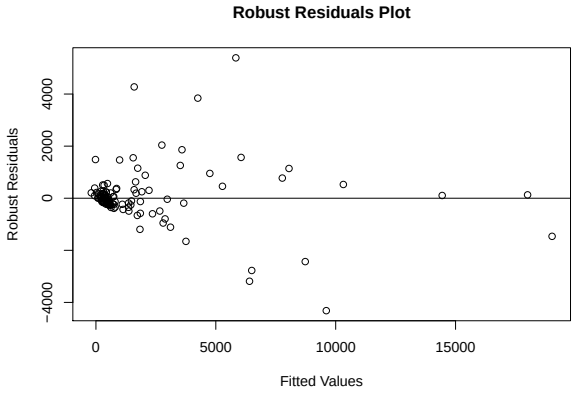


Figure 13: Robust Fitted vs. Residuals

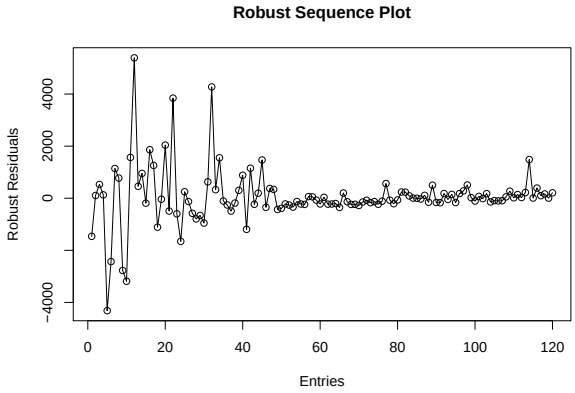


Figure 14: Robust Sequence Plot

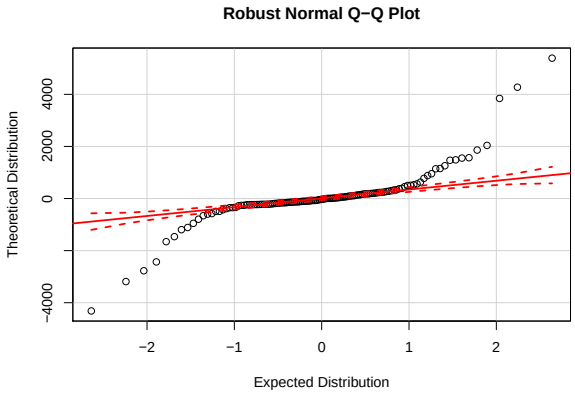


Figure 15: Robust Q-Q Plot

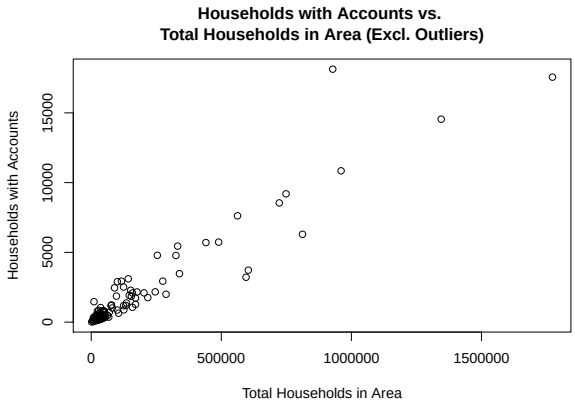


Figure 16: Y vs. X_1 Excl. Outliers

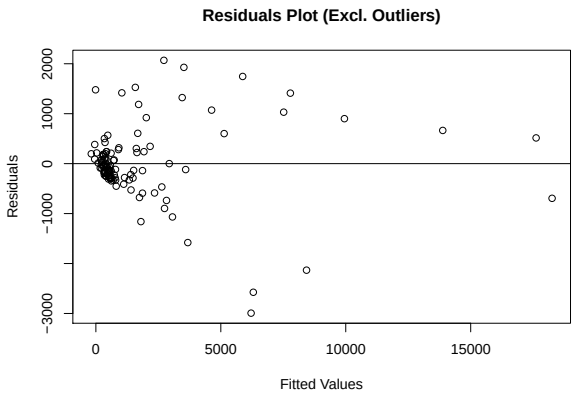


Figure 17: Fitted vs. Residuals Excl. Outliers

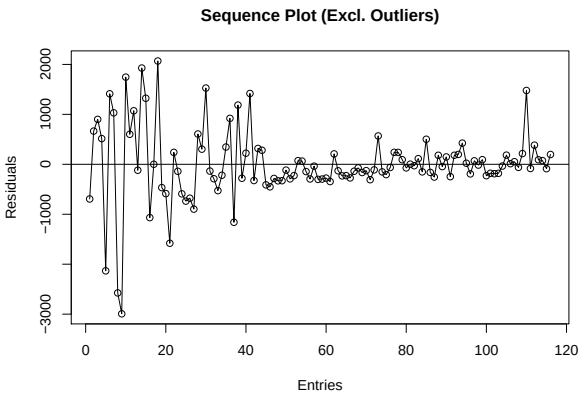


Figure 18: Sequence Plot Excl. Outliers

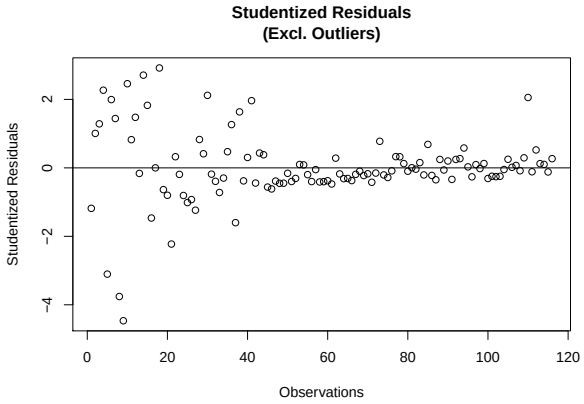


Figure 19: Studentized Residuals Excl. Outliers

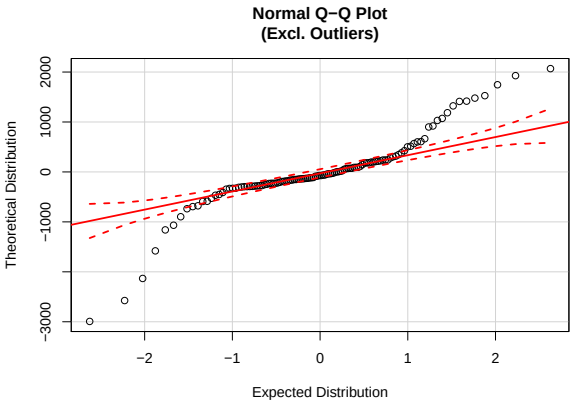


Figure 20: Normal Q-Q Excl. Outliers

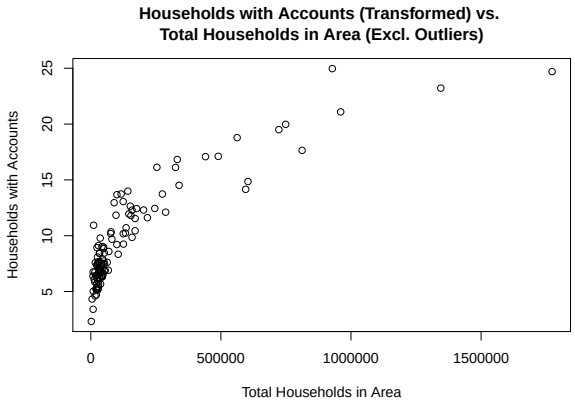


Figure 21: Y vs. X_1 Excl. Outliers Transformed

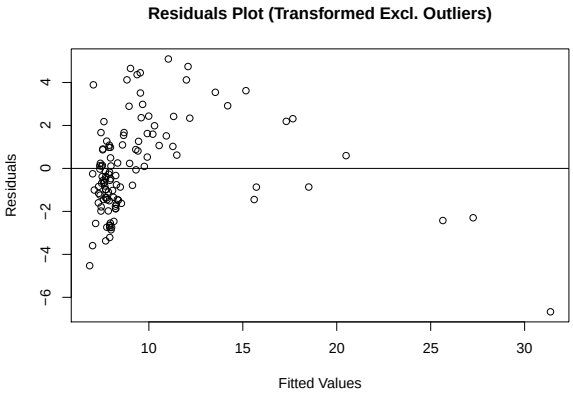


Figure 22: Fitted vs. Residuals Excl. Outliers Transformed

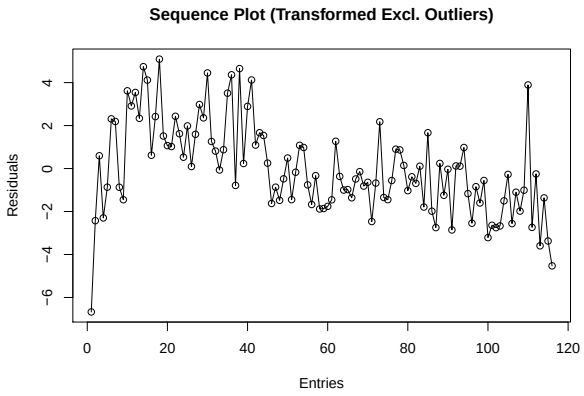


Figure 23: Sequence Plot Excl. Outliers Transformed

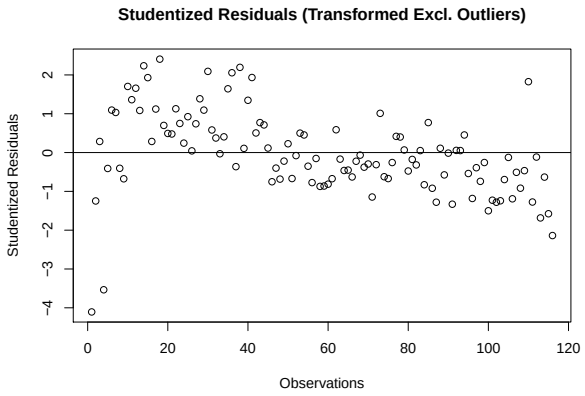


Figure 24: Studentized Residuals Excl. Outliers Transformed

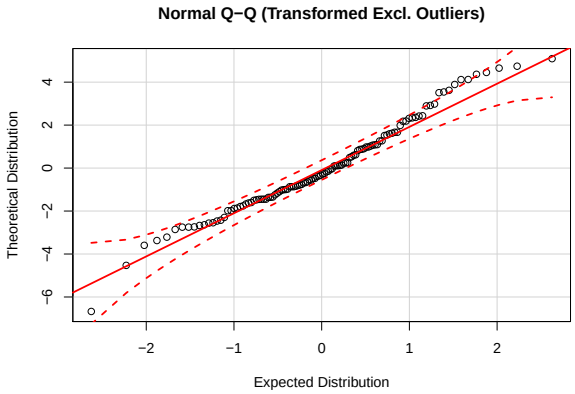


Figure 25: Normal Q-Q Excl. Outliers Transformed

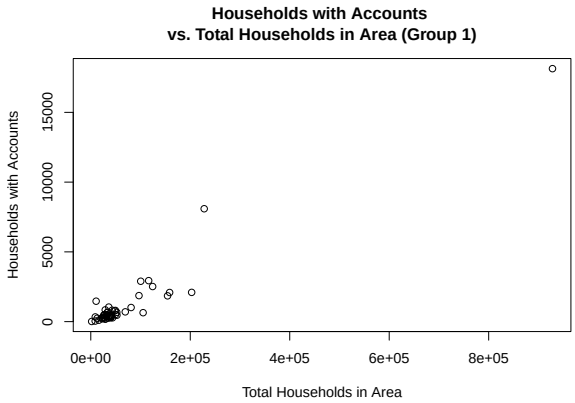


Figure 26: Y vs. X_1 (Group 1: Bank Branch Inside)

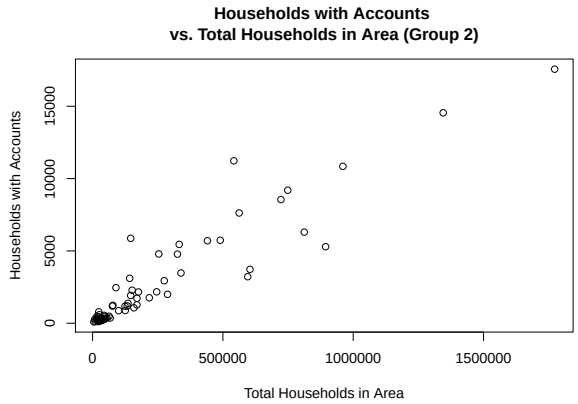


Figure 27: Y vs. X_1 (Group 2: Bank Branch Outside)