Effect of Physical Retail Bank Presence On Number of Checking Accounts

Andre Duarte and Francisco Calderon
MSAN 601

October 5, 2016

1 Introduction

This report is aimed at determining whether the presence of a physical branch creates demand for checking accounts at Large Regional. Intuitively, having a physical bank location nearby could be a significant factor in customers' decision to open a checking account.

To test this hypothesis, a linear regression model was constructed and evaluated. The purpose is to predict the number of households with checking accounts as a function of the total number of households in the area and the bank's footprint (whether it has a local branch). The statistical significance of the latter is of particular interest.

The results show that having a physical branch in the vicinity does not seem to have a direct relation with the local demand for checking accounts at Large Regional.

2 Method

2.1 Data

The key features used to model the number of households with checking accounts are the total number of households in the a metropolitan area and the physical branch presence. The first was numeric, whereas the latter was binary: "inside" if there is a branch in the area, and "outside" if not. The response variable consisted of the number of households with checking accounts, which was also numeric. Each case was a metropolitan area where Large Retail has customers with checking accounts. Data points for 120 metropolitan areas were available.

2.2 Linear Model

The use of a linear model in this case allows inference in the results. Thus, it is possible to conclude predictive ability of the total number of households and the physical presence of a branch on the number of checking accounts with values that are interpretable. Simplicity of the model offers actionable insights and business decisions for Large Regional.

2.3 Before Building the Model

Before building a model, an initial scatter plot was used to visually inspect the relationship between the two numeric variables. This is shown in Figure 1. Most of the points are concentrated near the origin of the graph and the space between points increases as the number of households increases.

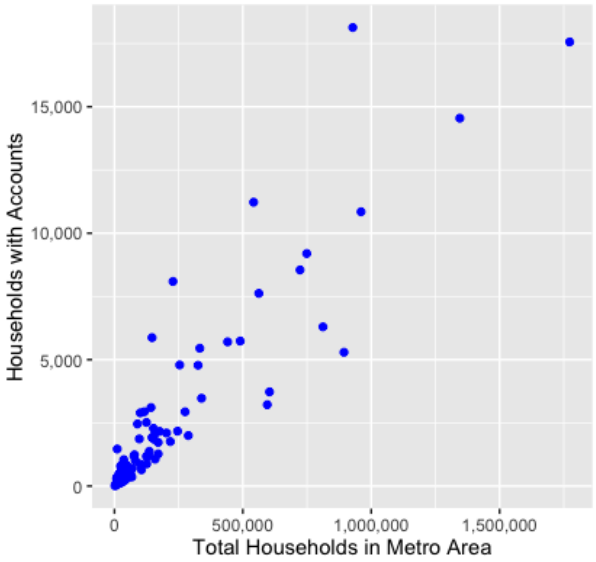

Figure 1: Scatter plot of total households vs. households with accounts

A log-log transformation was applied in order to correct this issue. The resulting scatter plot shows a linear relationship with more equidistant points, as shown in Figure 2. The transformed data is kept in order to build the linear model.

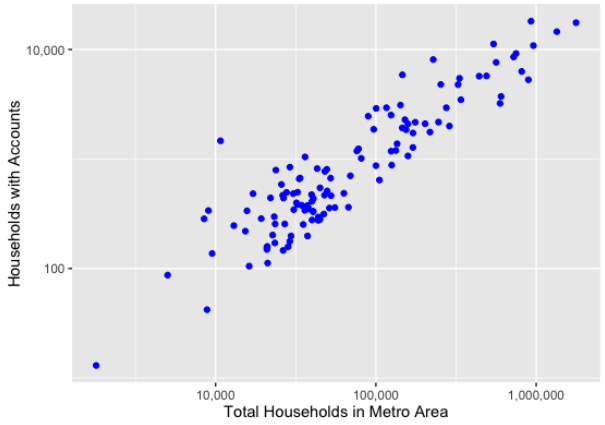

Figure 2: Scatter plot of total households vs. households with accounts after log-log transformation

In addition, the physical presence of a branch is well divided along both possibilities (53 "Inside" and 67 "Outside"). No transformation is done on this variable.

2.4 Evaluating the Model

In order to evaluate the linear regression model, several tests are run to determine normality of the residuals, homoskedasticity, independence, and presence of outliers.

In addition, multi-colinearity among the predictor variables was checked through the correlation matrix and scatter plot matrix. Lastly, Ramsey's RESET test was used to indicate whether any further intelligence could be gained from higher-order or interaction terms.

All tests were conducted with a significance level of $\alpha = 0.05$.

3 Results

The model used for this study is in the form

$$E\{\log Y\} = \beta_0 + \beta_1 \log X_1 + \beta_2 X_2$$

where

- Y: households with checking accounts
- X_1: total households in area
- X_2: presence of a physical branch

This model suggested that there was a linear relationship between the explanatory variables and Y. The statistical significance of the regressors was assessed, and in particular, β_2 is the coefficient that is of interest in this case study.

As an equation, this can be written as

$$\log Y = b_0 + b_1 \log X_1 + b_2 X2$$

where b_i are the estimators for β_i and their values are summarized in Table 1. Both b_0 and b_1 are statistically significantly different from 0, whereas b_2 is not.

Table 1: Summary of estimators

Estimator	Value	p-value
b_0	-4.171	6.28e-16
b_1	0.985	$<$ 2e-16
b_2	-0.185	0.084

For this regression model, the coefficient of determination is $R^2 = 0.8403$, and the adjusted coefficient is $R_a^2 = 0.8376$. 84% of the variability in the number of households with accounts is explained by the two predictors.

The correlation between the two explanatory variables is -0.250. This result shows that there is no multicolinearity between them. In addition, a Ramsey RESET test on this model gives a p-value of $0.3714 > \alpha = 0.05$. The null hypothesis that there are no significant quadratic or interaction terms to be included is therefore accepted.

In order to better interpret these results, and how the estimators affect the predicted value, it is useful to transform the previous equation into the form

$$Y = e^{b_0 + b_2 X_2} \cdot X_1^{b_1}$$

An increase in the total number of households in the area (X_1) by 10% leads to an increase in the number of households with an account (Y) by 9.85%.

However, it can be seen from Table 1 that the estimator for β_2 was not significant at the 5% significance level. The null hypothesis that $b_2 = 0$ was not rejected. The presence of a bank branch in the area does not have a statistically significant impact on the number of households with checking accounts. Indeed, the partial coefficient of determination for this variable was $R^2_{Y\ X_2|X_1} = 0.02596$. This means that adding the information of presence of a physical branch to the simple linear regression model (that includes only X_1) only marginally improved the model but was not significant.

3.1 Model validation tests

Before accepting this final model, several tests were run in order to verify that all linear regression assumptions are maintained. In particular, normality of the residuals, independence, homoskedasticity, and outliers are tested.

3.1.1 Normality of the residuals

Normality of the residuals was assessed visually using a QQ-plot, shown in Figure 3. Most standardized residuals fall on the diagonal line with some skewness present on the tails. In particular, one data point fell way off the diagonal line, which could be due to it being an outlier. Overall, the residuals approximated normal distribution.

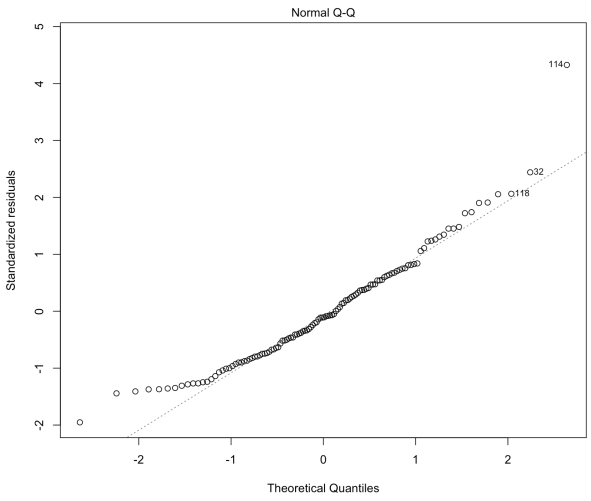

Figure 3: Q-QPlot of standarized residuals

3.1.2 Independence of the residuals

The data was received sorted by the total number of households in the area, and no information concerning the time and/or location of data gathering is available. Consequently, a sequence plot cannot be traced, as this information is missing. For the purposes of this report, independence of the observations is assumed to be true.

3.1.3 Homoskedasticity

Homoskedasticity was verified through a Breusch-Pagan test, where the null hypothesis was that there is no relationship between the predictor variables and the residuals. The p-value for this test was $0.036 > \alpha = 0.05$. The null hypothesis was therefore not rejected. The Residuals in the model were assumed to be homoskedastic.

3.1.4 Outliers

In order to detect potential outliers, the standardized residuals were plotted alongside the studentized residuals, both included the appendix. Observation number 114 in the data was possibly an outlier. This observation had a higher number of households with accounts for the number of total households than would be expected. This may have had many possible explanations, but without additional information about the data set, no conclusive clarification can be provided. Not enough evidence was available to remove this observation from the data set.

4 Conclusion

In this report, a linear regression model was developed to assess the hypothesis that having a physical bank branch in a metropolitan area increases the number of households with checking accounts. The results obtained from the model suggested that this theory does not hold: the presence of a branch does not have a significant relation with the number of accounts.

In the end, this physical is not useful for predicting the number of households with checking accounts in the area. A simpler linear regression model could be used for this purpose. Such a model would look like

$$E\{\log Y\} = \beta_0 + \beta_1 \log X_1$$

This simpler model was not studied here since the purpose of this analysis was to assess the impact of having a physical branch in the area.

Although the conclusion was negative, it does not mean that having a physical branch has no effect on the number of accounts for a metropolitan area. In fact, other significant factors may have been omitted from this study, such as the proximity of metropolitan areas, advertising, marketing techniques, and product offerings. These factors could help determine what actions Large Regional can take to directly increase the demand for checking accounts in specific markets.

Appendix

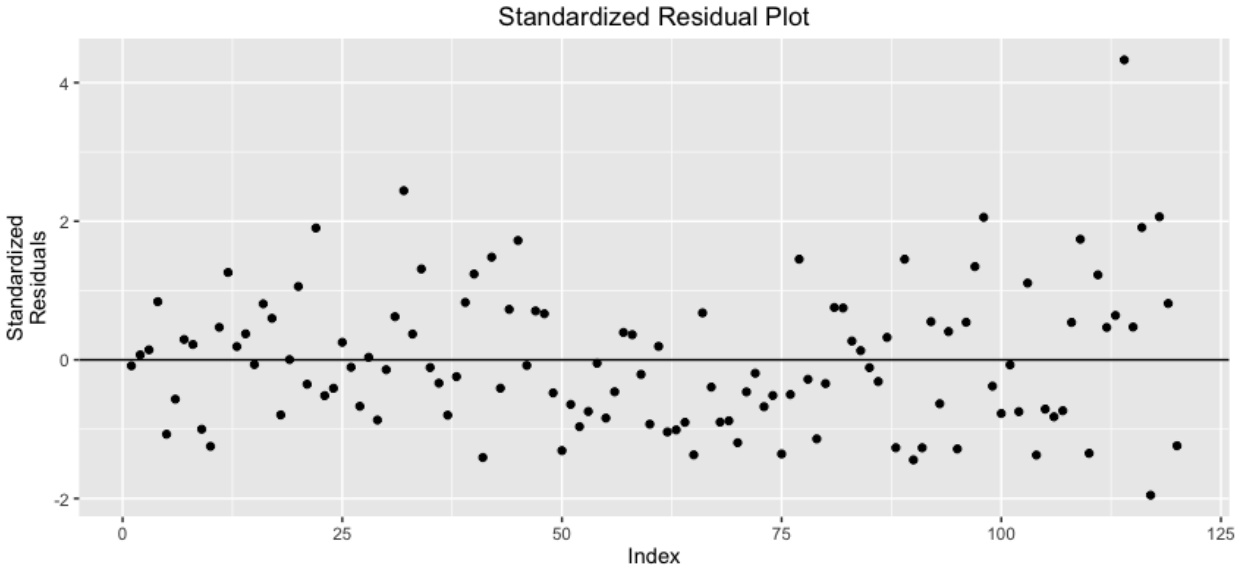

Figure 4: Standarized residuals

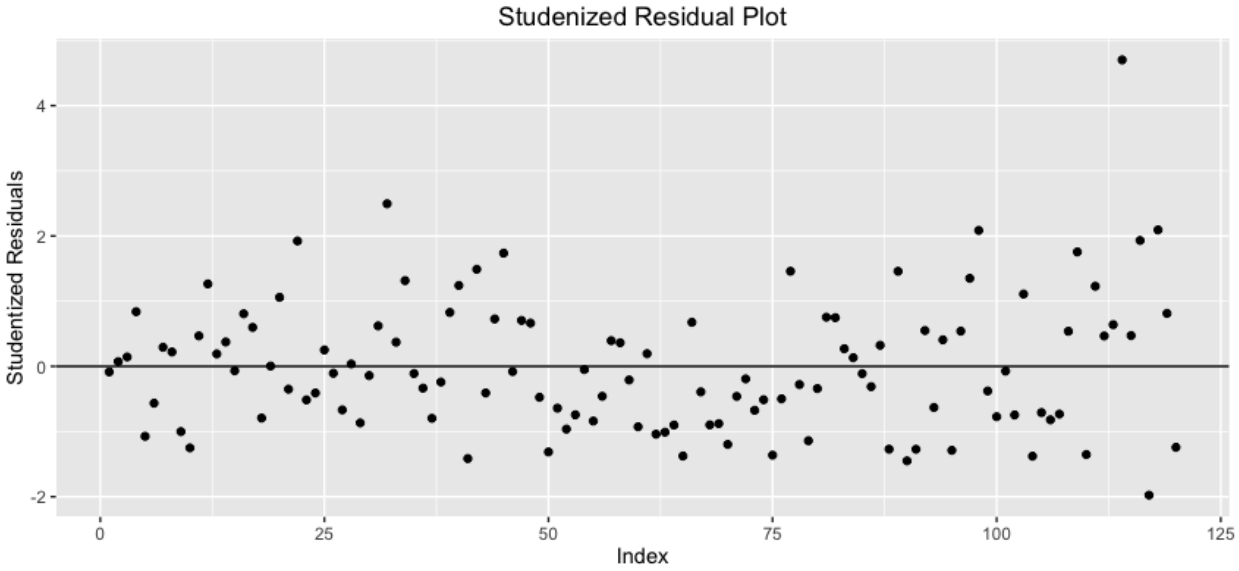

Figure 5: Studentized residuals

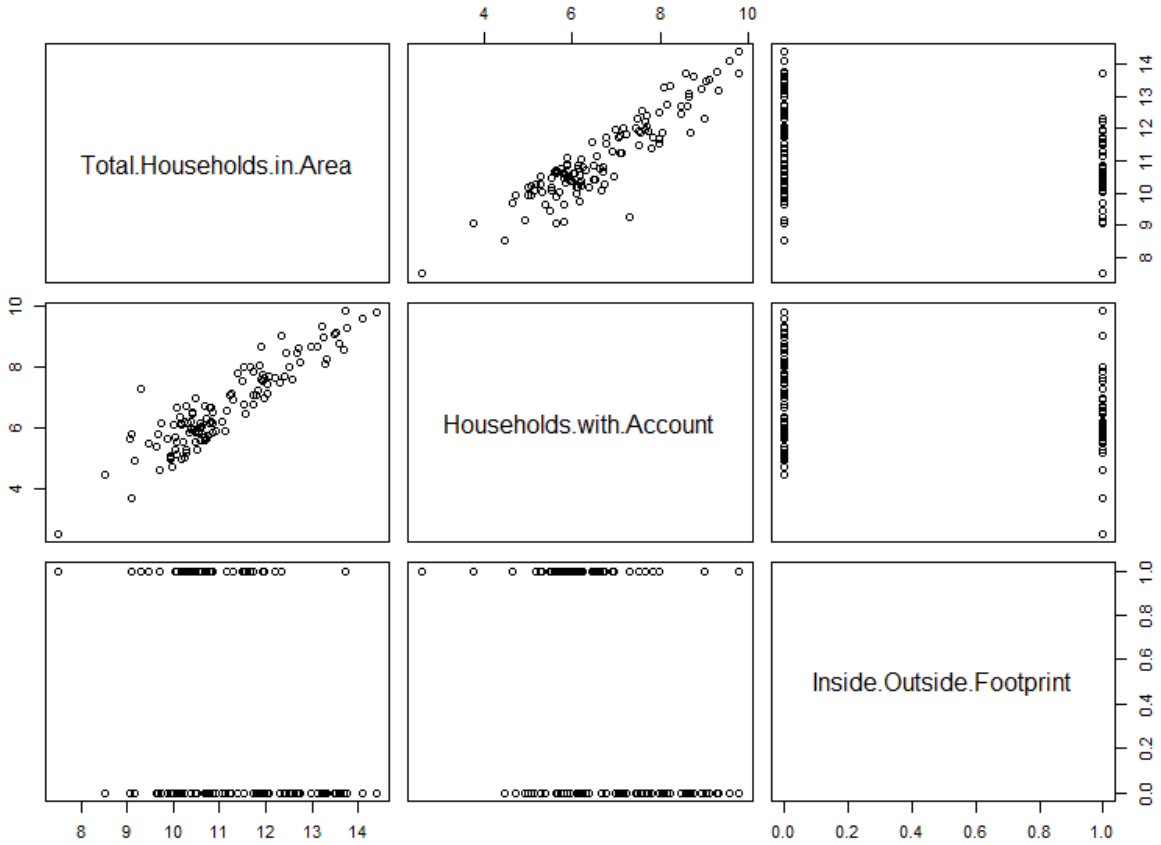

Figure 6: Scatterplot matrix with transformed X and Y

Table 2: Correlation matrix with transformed variables

	Total.Households.in.Area	Households.with.Account	Inside.Outside.Footprint
Total.Households.in.Area	1.0000000	0.9144215	-0.3399767
Households.with.Account	0.9144215	1.0000000	-0.2503362
Inside.Outside.Footprint	-0.3399767	-0.2503362	1.0000000