# Case Study #1: Nils Baker

*Alice Zhao, Valentin Vrzheshch*

*September 30, 2016*

## Executive Summary

The goal of the study is to answer the question *"Is the presence of a physical bank branch creating demand for checking accounts?"* To find the answer to this question, we analyzed the data set for the case study. The null hypotheses is that physical bank presence does not create create demand for checking accounts. A linear regression model is fitted to the data and the null hypotheses is rejected for the linear model. However, log-log model reveals that physical presense is insignificant.

## Introduction

The data provided for the case study contains 120 observations (see table 1). Each observation is described by three meaningful values **Total Households in Area**, **Households with Account** and **Inside/Outside Footprint** (the **ID** variable is omitted on purpose).

Table 1: Summary of Observations

| Variable | Name | Mean | St. Dev. |
|---|---|---|---|
| **Total Households in Area** | $TH$ | $1.6260267 \times 10^5$ | $2.769554 \times 10^5$ |
| **Households with Account** | $HA$ | 1992.26 | 3301.49 |
| **Inside/Outside Footprint** | $Out$ | 67 Outside (**1**),<br>53 Inside (**0**) | |

Since all of the observations fall into two categories (**Inside** or **Outside Footprint**) the points and the lines for two categories will be represented differently on the plots (see table 2).

Table 2: Dummy Variable Conversion & Illustration

| **Inside/Outside Footprint** | Value | Color | Point |
|---|---|---|---|
| Inside | 0 | red | circle |
| Outside | 1 | blue | triangular |

# Methodology

For this research we chose the $\alpha = 5\%$ level of significance.

The data set is broken into two subgroups by the **Inside/Outside Footprint** dummy variable. The first question that has to be answered is whether the subgroups are equal in some sense. To answer this question we will test whether mean and variance of two subsets are equal.

## 1. Initial Comparison of Two Subgroups

Visual analysis of box plots (see fig. 1) shows that the distributions for subgroups are different. We will analyze the ratio of **Households with Account** to **Total Households in Area** since we assume that physical presense determines a share of households that might have account. Fig. 1 shows boxplots for logarithms of **Households with Account** since the Log-Log model will be chosen as the best for this data set later.

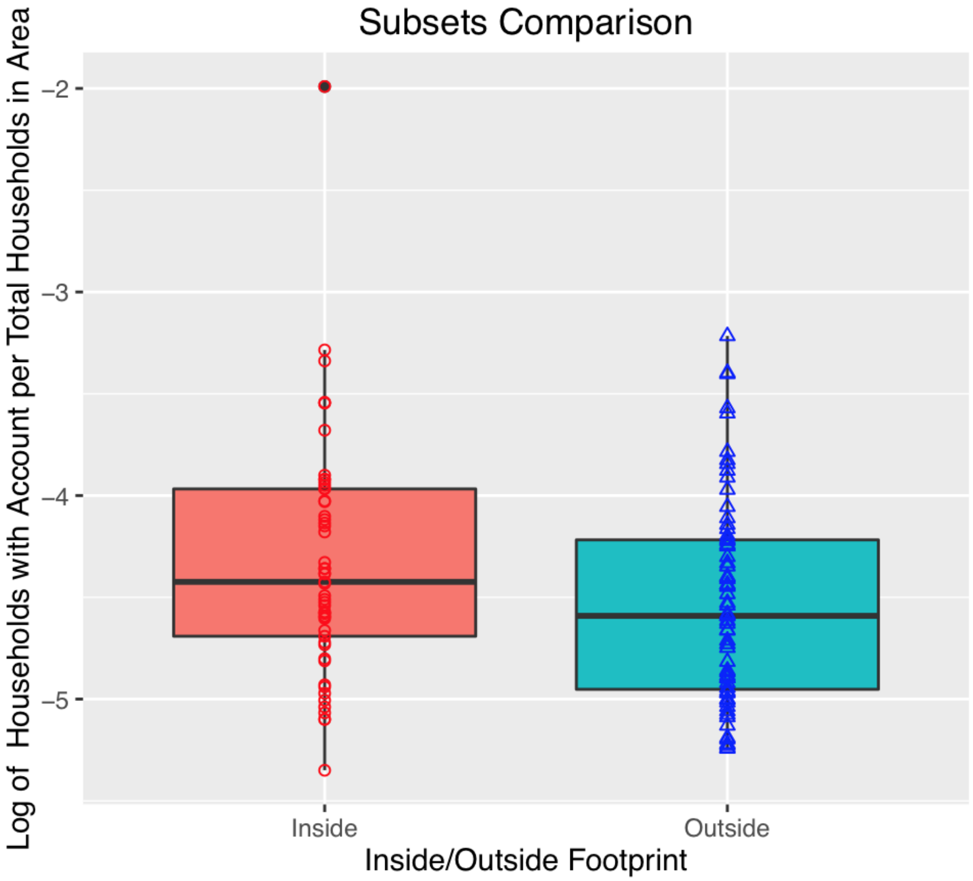

Figure 1: Subsets Comparison of Inside/Outside

### 1.1 Testing equivalence of two sample variances

We test the null hypothesis

$$\mathrm{H}_0 : \sigma^2_{Inside} = \sigma^2_{Outside}$$

against alternate hypothesis

$$\mathrm{H}_a : \sigma^2_{Inside} \neq \sigma^2_{Outside},$$

where $\sigma^2$ stands for variance of ratio of **Households with Account** to **Total Households in Area**.

Empirical statistics are

$$\hat{\sigma}_{Inside} = 3.3826252 \times 10^{-4}, \quad \hat{\sigma}_{Outside} = 5.4656384 \times 10^{-5},$$

and F-statistic is $F^* = 6.1888931$. 95% confidence interval for the ratio of variance is [3.7126612, 10.4847085] and p-value is $1.612932 \times 10^{-11}$. Hence we reject the null hypothesis on any reasonable significance level and conclude that the subsets have different variances. This result shows that when we test for equality of sample means, we have to assume non-equal variances.

### 1.2 Testing equivilance of two sample mean

The formal test is similar to that of the previous subsection. We test the null hypothesis

$$\mathrm{H}_0 : \mu_{Inside} = \mu_{Outside}$$

against alternate hypothesis

$$\mathrm{H}_0 : \mu_{Inside} \neq \mu_{Outside},$$

where $\mu$ stands for mean of ratio of **Households with Account** to **Total Households in Area**.

Empirical statistics are

$$\hat{\mu}_{Inside} = 0.0165366, \quad \hat{\mu}_{Outside} = 0.0124114,$$

and t-statistic is $t^* = 1.5375757$. 95% confidence interval for the difference of means is [-0.0012325, 0.0094829] and p-value is 0.128984. Hence we **cannot** reject the null hypothesis on **5%** significance level. Therefore we cannot conclude that the subsets have different means.

## 2. Two Simple Linear Regression Models

We know that the subgroups are not similar at least with respect to empirical variance. Hence we might want to run two separate regressions for each group in the form of

$$HA_i = \beta_0 + \beta_1 TH_i + \epsilon_i$$

The results of the two linear models are described in the table 5 in **Appendix**. The two SLR are as below:

$$Inside : HA_i = -194.456 + 0.02TH_i$$
$$Outside : HA_i = 251.537 + 0.01TH_i$$

The coefficients of determination $R^2$ are 0.926 and 0.867 for **Inside** and **Outside** subsets respectively. Both slopes are significant on **1%** level of significance. Hence we can conclude that **Households with Account** explains **Total Households in Area**, however the slope for **Inside** is twice greater than that for **Outside**. The top-left plot in figure 2 shows that there is a significant difference in the slopes. To determine if the difference in slope is statistically significant, we should perform a multiple regression analysis with an interaction term.

## 3. Multiple Linear Regression Models

### 3.1 Selecting Variables

Assuming that **Households with Account** can be fitted as a fraction of **Total Households in Area** we can run a linear model

$$HA_i = \beta_0 + \beta_1 TH_i + \beta_2 Out_i + \beta_3 TH_i Out_i + \epsilon_i$$

The interaction term $TH_i Out_i$ is significant on **1%** level for the whole set model. This supports the assumption that physical presence matters. Since the dummy variable alone is not significant, let's try removing it and compare the two models.

Thus, we decided to pick the first-order model without dummy variable because AIC, BIC and PRESS scores are lower for the first-order model without dummy variable alone:

$$HA_i = \beta_0 + \beta_1 TH_i + \beta_3 TH_i Out_i + \epsilon_i$$

The comparison of two models is provided in table 3.

| | model | | SSE | $F^*$ | $R^2$ | $R_a^2$ | AIC | BIC | PRESS |
|---|---|---|---|---|---|---|---|---|---|
| 1 | HA | TH + Out + TH:Out | 139722186.26 | 320.29 | 0.89 | 0.89 | 2026.67 | 2040.60 | 152238183.97 |
| 2 | HA | TH + TH:Out | 143966829.49 | 468.56 | 0.89 | 0.89 | 2028.26 | 2039.41 | 156927240.53 |

Table 3: Comparison of Two MLR Models

### 3.2 Refining the Model

From now on we will analyze and compare two models:

- The first-order model without dummy variable but with interaction term:

$$HA_i = \beta_0 + \beta_1 TH_i + \beta_2 TH_i Out_i + \epsilon_i$$

- The Log-Log model without dummy variable but with interaction term:

$$log(HA_i) = \beta_0 + \beta_1 log(TH_i) + \beta_2 TH_i Out_i + \epsilon_i$$

The F-statistic for the first-order model without dummy variable is lower than that of the first-order model with dummy variable and it is 468.5613099 and the p-value is $1.4116975 \times 10^{-56}$, which is much smaller than $\alpha = 5\%$. Therefore, we can reject the null-hypothesis and conclude that the first-order without the dummy variable model provides a better fit than the intercept-only model.

Let's also check for quadratic terms for the model. After running Ramsey RESET test for power of 2, we get a p-value of 0.3. Therefore, there is no significant evidence to reject the $H_0$ that there is no quadratic terms in the model.

Thus, we are safe to proceed with our first order linear model.

### 3.3 Test Model Assumptions

#### 3.3.1 Homoscedasticity

Let's run Breusch-Pagan Test to check homoscedasticity. Residual Plot can be found on fig. 3 in appendix. From the BP-test, we get a p-value of 0.003. For $\alpha = 0.05$, there is signficant evidence to reject null hypothesis that the residuals are homoskedastic. We have detected heteroscedaticity issue.

#### 3.3.2 Normality of Errors

Next, we will run Pearson's Chi-squared Test for goodnes of fit to test residuals for normality:

$$\mathrm{H}_0 : e_i \sim \mathrm{N}(0, \sigma^2) \; iid$$

against the alternate hypothesis that the residuals are not normally distributed

$$\mathrm{H}_a : e_i \nsim \mathrm{N}(0, \sigma^2) \; iid$$

QQ plot can be found in appendix (fig. 4). From the test we get p-value of $1.6685693 \times 10^{-19}$. Therefore, there is a signficant evidence to reject the null hypothesis that the errors are normal. We have detected non-normality issue.

### 3.4 Addressing Violations of Homoscedasticity and Normality

#### 3.4.1 Homoscedasticity

To address the heteroskedasticity issue, we tranform both the predictor variable **Total Households** and the response variable **Households with accounts** with log.

$$log(HA_i) = \beta_0 + \beta_1 log(TH_i) + \beta_2 TH_i Out_i + \epsilon_i$$

We have omitted the dummy variable in Log-Log model because it is insignificant on 10% significance level. This happened because $log(TH) * Out$ and $Out$ are highly correlated ($\hat{\rho}_{log(TH)*Out,Out} = 0.984$), so we deal with a multicollinearity issue. It is worth noting that we do not encounter the same multicollinearity issue in case of first-order model since $\hat{\rho}_{TH*Out,Out} = 0.426$.

| | model | | SSE | $F^*$ | $R^2$ | $R^2_a$ | AIC | BIC | PRESS |
|---|---|---|---|---|---|---|---|---|---|
| 1 | log(HA) | log(TH) + log(TH):Out | 34.47 | 309.05 | 0.84 | 0.84 | 198.84 | 209.99 | 36.34 |
| 2 | log(HA) | log(TH) + Out + log(TH):Out | 34.20 | 206.13 | 0.84 | 0.84 | 199.93 | 213.87 | 37.07 |

Table 4: Comparison of Two MLR Models after log transormation

For the new Log-Log model, the F-statistic is 309.0511934 and the p-value is $2.0301339 \times 10^{-47}$, which is much smaller than $\alpha = 5\%$. The residual plot can be found in **Appendix** (fig. 5).

Then let's run Breusch-Pagan Test again. This time we get a p-value of 0.036. We still cannot conclude the null hypothesis that residuals are homoscedastic, however this p-value is closer to $\alpha = 0.05$ than the p-value for first-order model for the same BP-test which is 0.003.

#### 3.4.2 Normality of Errors

We will again run Pearson's Chi-squared Test for goodness of fit to test residuals for normality for the transformed model. QQ plot can be found in **Appendix** (fig. 6).

After Log-Log transformation, we get p-value of 0.013. We still cannot conclude the null hypothesis that residuals are normally distributed, however this p-value is much closer to $\alpha = 0.05$ than the p-value for first-order model for the same Pearson test for normality which is $1.6685693 \times 10^{-19}$.

# Results

The main result of the research is that the dummy variable Inside/Outside Footprint that indicates *the presence of a physical bank branch* is insignificant. To get to this conclusion we start with simple first-order models for two subgroups:

$$Inside : HA_i = -194.456 + 0.02TH_i$$

$$Outside : HA_i = 251.537 + 0.01TH_i$$

and get relatively high $R^2$'s - 0.926 and 0.867 for **Inside** and **Outside** subsets respectively. Since the slope for **Inside** is twice greater than that for **Outside** we perform a multiple regression analysis with an interaction term to determine if the difference in slope is statistically significant.

To check the significance of physical presense we estimate a multiple linear regression model in the form of

$$HA_i = \beta_0 + \beta_1 TH_i + \beta_2 Out_i + \beta_3 TH_i Out_i + \epsilon_i$$

The interaction term $TH_i Out_i$ is significant on **1%** significance level for the whole data set model. Hence the first-order model shows that physical presence matters. Table 5 in **Appendix** shows statistics for this model.

Table 5:

| | *Dependent variable:* | | |
|---|---|---|---|
| | HA | | log(HA) |
| | (1) | (2) | (3) |
| TH | 0.020***<br>(0.001) | 0.019***<br>(0.001) | |
| Out | 445.993*<br>(237.581) | | |
| TH:Out | −0.010***<br>(0.001) | −0.009***<br>(0.001) | |
| log(TH) | | | 1.000***<br>(0.044) |
| log(TH):Out | | | −0.018*<br>(0.010) |
| Constant | −194.456<br>(171.531) | 38.026<br>(119.956) | −4.322***<br>(0.466) |
| Observations | 120 | 120 | 120 |
| $R^2$ | 0.892 | 0.889 | 0.841 |
| Adjusted $R^2$ | 0.889 | 0.887 | 0.838 |
| Residual Std. Error | 1,097.498 (df = 116) | 1,109.273 (df = 117) | 0.543 (df = 117) |
| F Statistic | 320.288*** (df = 3; 116) | 468.561*** (df = 2; 117) | 309.051*** (df = 2; 117) |
| *Note:* | | | *p<0.1; **p<0.05; ***p<0.01 |

However, further analysis reveals that first-order linear model fails to deliver homoscedastic and normally distributed residuals. Hence we have to come up with another model that provides residuals which satisfy the assumptions of multiple regression model framework.

Next we tranform the variables with logarithm and analyze the Log-Log model. The Log-Log model provides a smaller but very close $R^2$. Unlike first-order model, the Log-Log model proves the physical presense to be insignificant. Moreover, the Log-Log model's residuals are homoscedastic and normally distributed on *1%* level of significance. Hence, we contemplate the Log-Log model as a better fit. Since the Log-Log model denies the significance of dummy variable, we conclude that *the presence of a physical bank branch* **does not** significantly affect the *demand for checking accounts.*

## Conclusion

Based on the provided data set we conclude that *the presence of a physical bank branch* **does not** significantly affect the *demand for checking accounts.* We come to this conclusion by comparing two model approaches - first-order model and Log-Log model. Although in the first-order model approach the coefficient for dummy variable of **Inside/Outside Footprint** happens to be significant on *a priori* chosen 5% level of significance, the Log-Log model proves to be a better fit for the data set and the Log-Log model shows that *the presence of a physical bank branch* is insignificant. The limited amount of data does not allow us to fully address the issue of the true (*theoretical*) model that generates this data. However, the Log-Log model results in much less heteroscedastic and much more normally distributed residuals. Hence, the Log-Log model should be chosen to make inference. Therefore, we conclude that physical presence does not statistically significantly affect the demand for checking accounts.

# Appendix

## Methodology

### 2. Two Simple Linear Model

We compare four types of models:

- Whole data set vs. two subsets;
- First-order regressors vs. logarithms of **TH** (*Total Households in Area*) and **HA** (*Households with Account*).

The results of these models are shown on the plots in figure 2. * Top row - first-order models;

- Bottom row - Log-Log models;
- Left column - separate SLR models for both subgroups;
- Right column - SLR models for the whole data set.

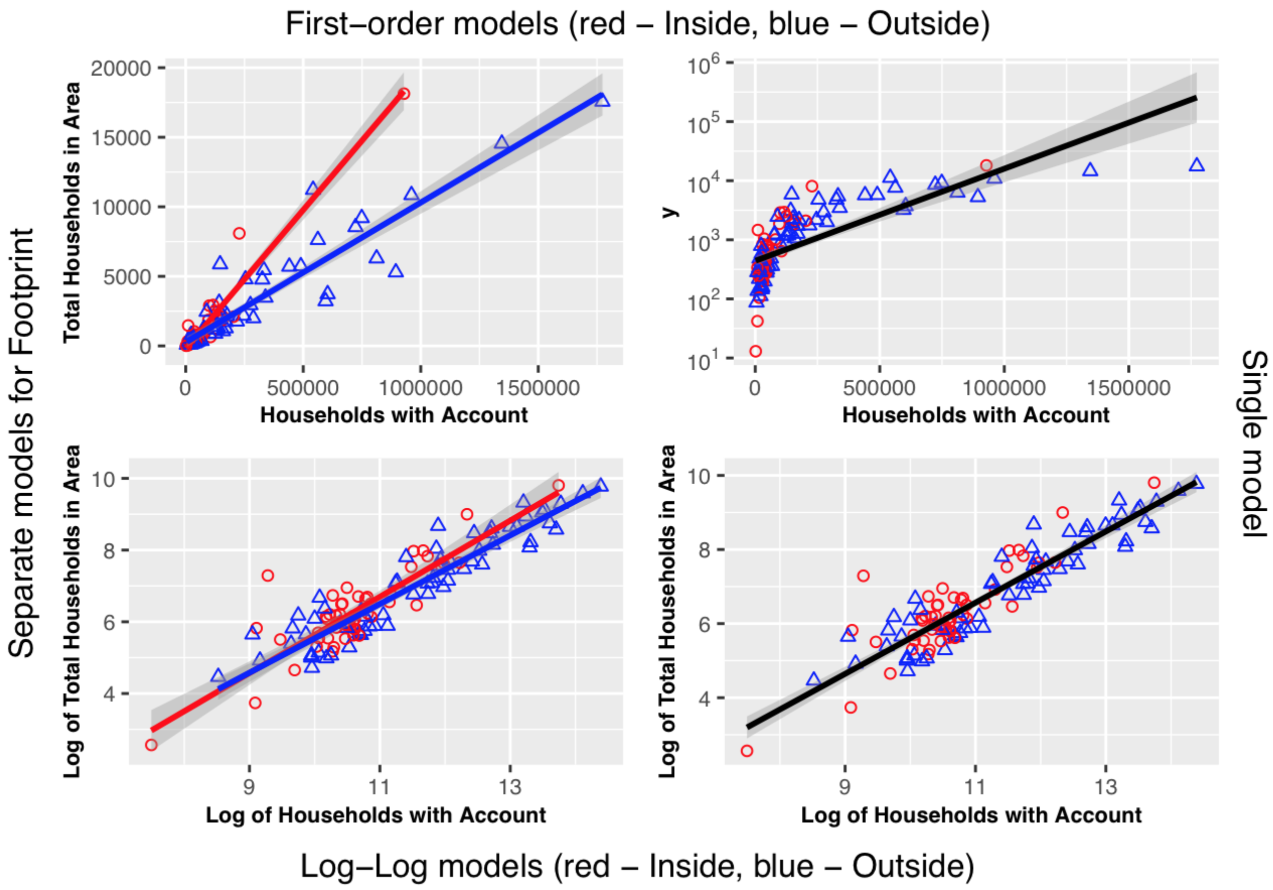

Figure 2: Comparison of SLR Models

The Summary of the two separated linear models can be found below:

Table 5:

| | *Dependent variable:* | |
|---|---|---|
| | HA | |
| | Inside Model | Outside Model |
| | (1) | (2) |
| TH | 0.020$^{***}$ | 0.010$^{***}$ |
| | (0.001) | (0.0005) |
| Constant | $-$194.456$^{*}$ | 251.537 |
| | (113.310) | (197.413) |
| Observations | 53 | 67 |
| R$^2$ | 0.928 | 0.869 |
| Adjusted R$^2$ | 0.926 | 0.867 |
| Residual Std. Error | 724.987 (df = 51) | 1,318.019 (df = 65) |
| F Statistic | 655.575$^{***}$ (df = 1; 51) | 432.674$^{***}$ (df = 1; 65) |
| *Note:* | | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

## 3. Multiple Linear Regression Models

### 3.3 Test Model Assumptions

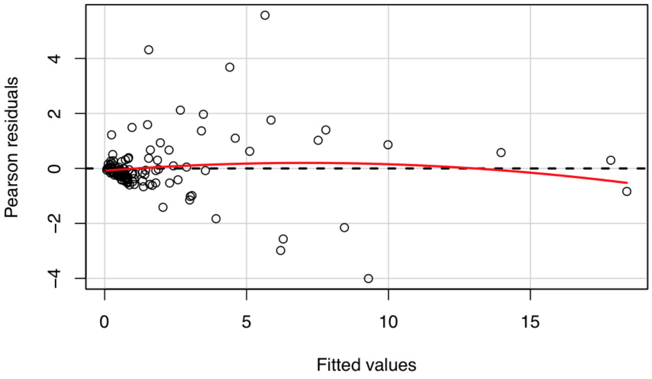

Figure 3: Residual Plot for the first-order model with interaction term

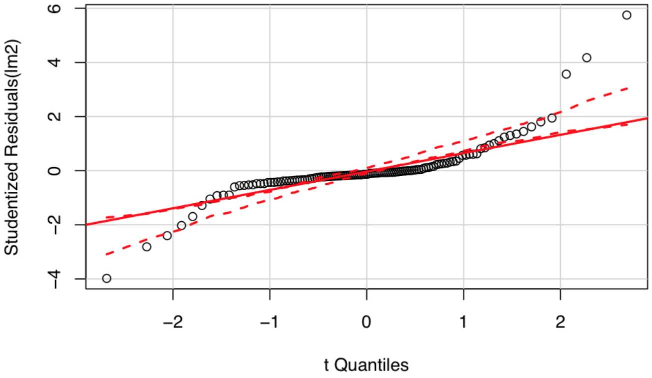

Figure 4: QQPlot for the first-order model with interaction term

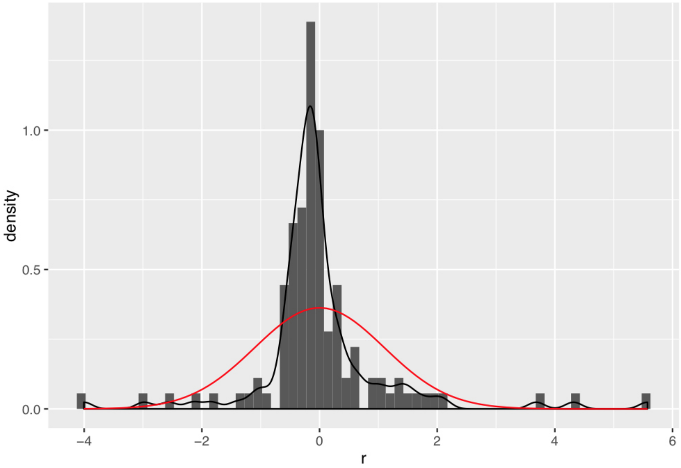

Figure 5: Histogram of the first-order model with interaction term

### 3.4 Addressing Violations of Homoscedasticity and Normality

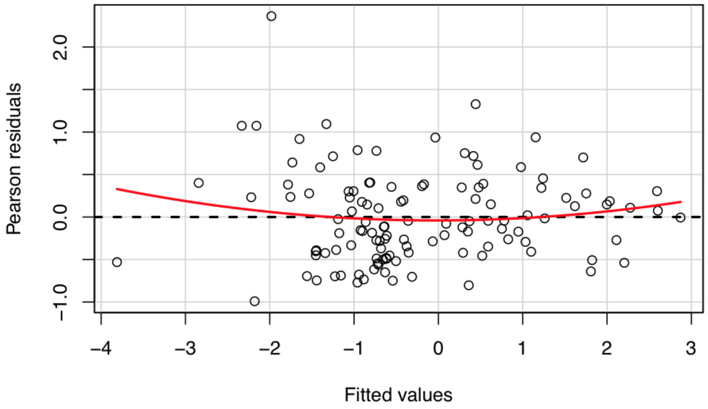

Figure 6: Residual Plor for the Log-Log model with interaction term

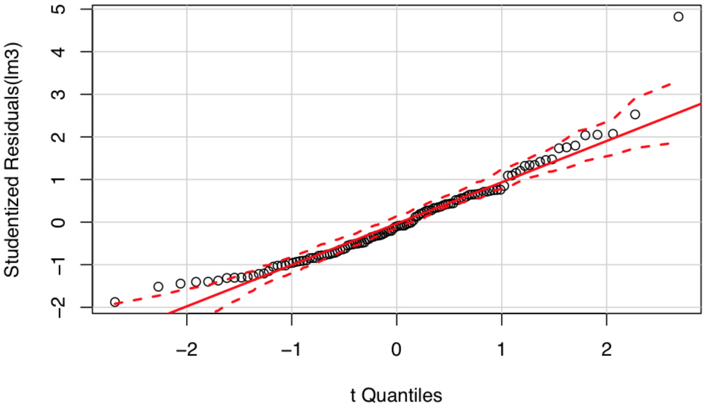

Figure 7: QQPlot for the Log-Log model with interaction term

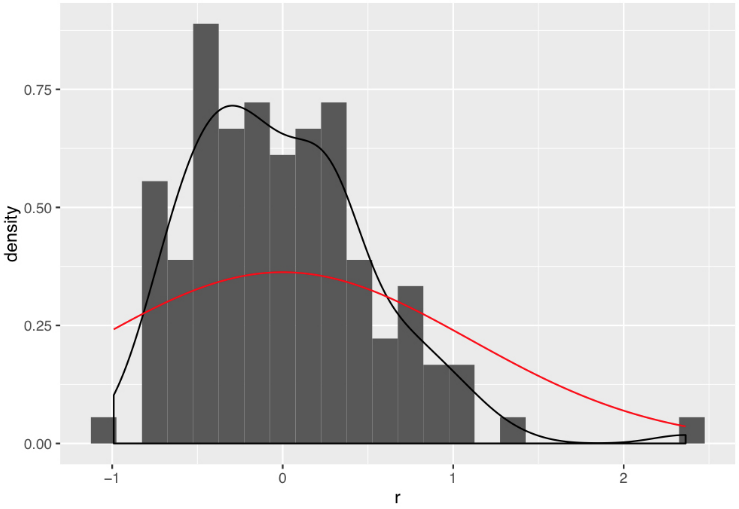

Figure 8: Histogram of the Log-Log model with interaction term