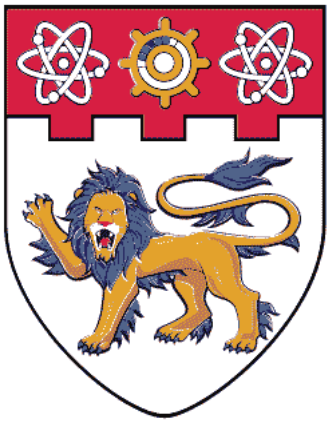


NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

BC3406 BUSINESS ANALYTIC CONSULTING

Nils Baker Case Analysis

Name: Kong Ming Yeow

Matriculation: U2010866H

Submitted on 30 January 2023

No. of pages: 3

With reference the Nils Baker case, regarding his question on:

Is the presence of a physical bank branch creating demand for checking account?

To provide the answer for this question, I have analysed the data set for 120 MSAs, with variables *Total Households in Area*, *Households with Account* and *Inside/Outside Footprint*, and determined that a physical bank presence **does not create** demand for the checking accounts.

I understand that it is counter-intuitive that the data does not support a relationship between them but let me run through a brief summary of my findings and analysis that drawn me to this conclusion.

1. Statistical Analysis

For this research, I have chosen the α = 5% level of significance.

Figure 1: Summary of Observations

	Inside.Outside.Footprint	mean	sd	min	lower	middle	upper	max
1	Outside	1.24%	0.74%	0.53%	0.71%	1.01%	1.47%	4.01%
2	Inside	1.65%	1.84%	0.47%	0.92%	1.20%	1.89%	13.66%

I used the ratio of **Households with Account** to **Total Households in Area** for comparison, and on a first glance, there is a higher percentage of checking account in area with a physical bank (1.65%) than area without a physical bank (1.24%). However, just a comparison of ratio is not enough to determine whether there is a relationship between the variables, thus, to test the pattern observed, the mean of both subsets (Inside/Outside) of this ratio will be used for a two-sample T-test.

1.1 Two-Sample T-test (One-Tail)

Our hypothesis is that there is more checking account in area with physical bank branch, hence, we test the null hypothesis:

$$H_O : \mu_{Inside} = \mu_{Outside}$$

Against alternate hypothesis:

$$H_a : \mu_{Inside} > \mu_{Outside}$$

With the empirical statistics result in Appendix A, the p value of 0.06449 is greater than the 0.05 (5%) significant level, hence we **cannot** reject the null hypothesis of no difference and say with a high degree of freedom that there is no statistical difference between the two means.

The result of this t-test implies that there is no significance in the difference of mean on **Households with Account** to **Total Households in Area**, meaning that even if there is higher percentage of checking account in area with a physical bank, the difference is not significant, and could be due to external factors or due to chance.

1.2 Multiple Linear Regression Models

To test whether the variable of physical bank branch is significant, multiple linear regression is used.

$$Households.with.Account = \beta_0 + \beta_1\ Total.Households.in.Area + \beta_2\ Inside$$

With the statistics result in Appendix B, the interaction term of Inside is not significant on 5% level with a p-value of 0.118. Hence, this supports the assumption that physical presence does not matters.

1.3 Testing Regression Model Assumptions

1.3.1 **Homoscedasticity**

Based on the Scale-Location plot in Appendix C, the residuals are showing an upward trend as the fitted values increase, this implies that the errors are unequal, or heteroskedasticity. Thus, the assumption on homoscedasticity of residuals variance is violated.

1.3.2 **Normality of Errors**

Based on the Normal Q-Q plot in Appendix C, the residuals do not follow the reference line, indicating that a non-normality issue.

1.4 Nonlinear Transforming to Address Violations of Assumptions

To address the non-normality and heteroskedasticity issues, both the dependent variable (*Households.with.Account*) and independent variable (*Total.Households.in.Area*) will be transform with log.

$$Log(Households.with.Account) = \beta_0 + \beta_1 Log(Total.Households.in.Area) + \beta_2 Inside$$

Based on the diagnostic plots of the new Log-Log model in Appendix D, the Log-Log model's residuals are more homoscedastic and normally distributed. Hence, the Log-Log model will be deemed as a better fit for this regression analysis.

With the statistical output in Figure 2 below, the interaction term of Inside is not significant on the 5% level with a p-value of 0.084, denying the significance of this independent variable. Thus, it is concluded that the presence of a physical bank branch **does not significantly affect** the demand for checking accounts.

Figure 2: Log-Log Model Statistical Output

```
Residuals:
    Min      1Q  Median      3Q     Max
-1.0440 -0.4053 -0.0587  0.3255  2.3180

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                  -4.35645    0.47912  -9.093 2.97e-15 ***
log(Total.Households.in.Area)  0.98521    0.04127  23.870  < 2e-16 ***
Inside.Outside.FootprintInside  0.18519    0.10627   1.743    0.084 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

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
```

2. Limitations and Recommendation

The presence of a physical bank branch was assumed to increase the demand for checking accounts, however after the statistical analysis, it is noted that the presence of a physical bank branch **does not significantly affect** the demand for checking accounts.

2.1 Correlation, Causality, and Limitation of Analysis

However, the presence of a physical bank may increase the accessibility of banking services to households in the area, leading to an increase in percentage of households with bank accounts. Thus, the correlation between the presence of a physical branch and demand for checking account is likely to be complex, and it does not prove that the presence of physical bank causes the demand for checking accounts.

Even if there is assumed to be a correlation, it does not prove the causality between the two variables. There are limitations in this analysis that restricts us to draw conclusion about the causality.

1. **Temporal Precedence**. We do not have data as to whether customers create their accounts before the presence of a physical bank branch or after, thus we are unable to establish that the cause occurs before the effect.
2. **Elimination of Alternative Explanations**. The alternative explanations for the cause-effect relationship are not considered or not eliminated as this analysis only looks into the presence of physical bank, not any other variables.

Furthermore, there could be various factors such as income levels, education levels, competitive interest rates or credit ratings that play a role in determining the percentage of households with bank accounts. But in this analysis, a limitation will be the availability of data where only the variable used for comparison is the presence of a physical bank.

2.2 Possible Alternative to Increase Demand for Checking Accounts

With many other factors that create demands for checking accounts, having a physical bank branch may not be economically feasible and viable to implement as the increase may not be significant. Thus, a possible method is to lower administrative/processing fees. Based on the case provided, the lady opted to receive euros at her local bank branch instead of through a secure courier mail, possibly due to the substantially higher cost. Thus, by lowering the cost of these fees, the bank may be able to attract more customers for their checking accounts without the need for a physical bank branch, but more research and analysis are needed to be done to prove this hypothesis.

3. Conclusion

All in all, the statistical analysis has proved that the presence of physical bank branch **does not significantly affect the demand** for checking account. There are many external factors that create the demand for checking accounts, but we are constrained by the limitations on the dataset to prove a cause and effect relationship. Mr Baker can look into other possible alternatives to draw in the demand of checking accounts.

4. Appendices

Appendix A: Statistics Result from Two-Sample T-test

```
        Welch Two Sample t-test

data:  dt$Percentage.Of.Total[dt$Inside.Outside.Footprint == "Inside"] and
dt$Percentage.Of.Total[dt$Inside.Outside.Footprint == "Outside"]
t = 1.5376, df = 65.302, p-value = 0.06449
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -0.0003513254           Inf
sample estimates:
 mean of x  mean of y
0.01653659 0.01241139
```

Before the T-test, it has to be established whether the variances of the two samples are equal or not through F-test. Therefore, we test the null hypothesis:

$$H_O : \sigma^2_{Inside} = \sigma^2_{Outside}$$

against alternative hypothesis

$$Ha : \sigma^2_{Inside} \neq \sigma^2_{Outside}$$

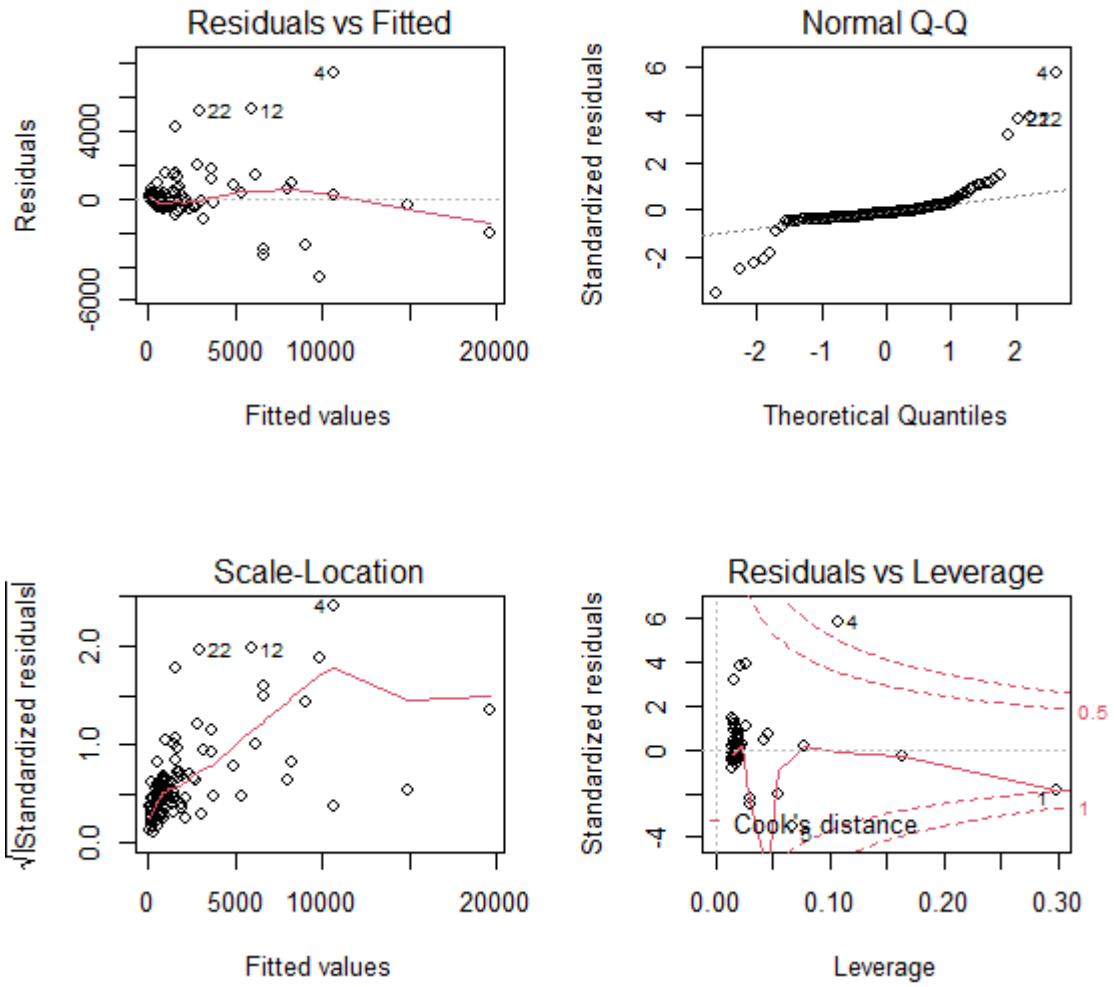
With the empirical statistics of the F-test below, the p-value is 1.613×10^{-11} , hence we reject the null hypothesis on any reasonable significance level and conclude that the subsets have different variances. Thus, Welch's T-test is used as there is unequal variance.

```
        F test to compare two variances

data:  dt$Percentage.Of.Total[dt$Inside.Outside.Footprint == "Inside"] and
dt$Percentage.Of.Total[dt$Inside.Outside.Footprint == "Outside"]
F = 6.1889, num df = 52, denom df = 66, p-value = 1.613e-11
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
  3.712661 10.484708
sample estimates:
ratio of variances
          6.188893
```

Appendix B: Linear Regression Model Statistical Output

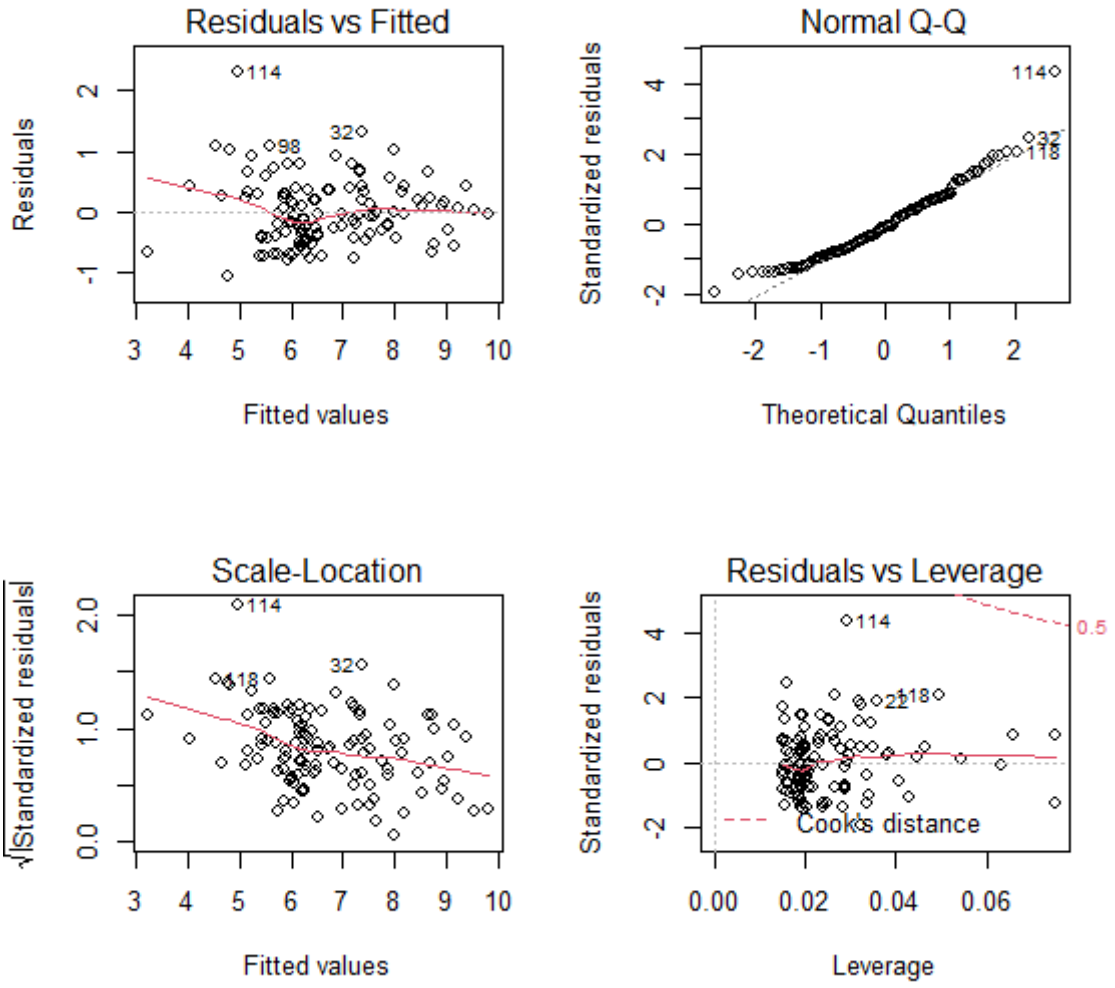
```
Call:
lm(formula = Households.with.Account ~ . - Percentage.Of.Total,
    data = dt)

Residuals:
    Min      1Q  Median      3Q     Max
-4626.9  -452.8  -213.8   184.0  7423.0

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                     7.718e+00  1.997e+02   0.039    0.969
Total.Households.in.Area        1.109e-02  4.710e-04  23.535   <2e-16 ***
Inside.Outside.FootprintInside 4.122e+02  2.616e+02   1.576    0.118
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1357 on 117 degrees of freedom
Multiple R-squared:  0.8338,    Adjusted R-squared:  0.831
F-statistic: 293.6 on 2 and 117 DF,  p-value: < 2.2e-16
```

Appendix C: Diagnostic Plots of Linear Regression Model



Plot 1: Residuals Vs Fitted	In our graph, it is observed that there is no pattern in the residual plot, suggesting that we can assume a linear relationship between the predictors and outcome variables.
Plot 2: Normal Q-Q	In our plot, the residuals clearly depart from the reference line, indicating that they do not follow a normal distribution.
Plot 3: Scale-Location	In our graph, the residuals are showing an upward trend as the fitted values increase, this implies that the errors are unequal, or heteroskedasticity.
Plot 4: Residuals Vs Leverage	In our plot, there is a possible 2 influential outliers, as they fall outside the Cook's distance. As the outliers are legitimate observation of the population, there are not removed.

Appendix D: Diagnostic Plots of Log-Log Regression Model



Plot 1: Residuals Vs Fitted	In our graph, it is observed that there is no pattern in the residual plot, suggesting that we can assume a linear relationship between the predictors and outcome variables.
Plot 2: Normal Q-Q	In our plot, the residuals are closer to the reference line, indicating that they follow a normal distribution.
Plot 3: Scale-Location	In our graph, even though there is a small downward trend, the residuals are generally more spread out and better compared to the one in Appendix C.
Plot 4: Residuals Vs Leverage	In our plot, the residuals fall inside the Cook's distance, thus there is no influential outliers.