Investigating the Factors Influencing Purchasing Behavior: A Statistical Analysis of the Account Dataset

Kazi Shahria

The project contains analysis and statistical testing on the ACCOUNT dataset in the regclass library. All the code and visuals are done on R Studio, with the parts broken apart from association analysis to a simple linear regression.

1. INTRODUCTION

Before diving into any dataset, it is vital to understand the variables.

Library: Regclass
Dataset: ACCOUNT
Description: Customers have marketed a new type of account at a bank.
Format: A data frame with 24,242 observations on the following eight variables (will only use 500 observations):

- Purchase - a factor with levels No Yes.
- Tenure - a numeric vector, the number of years the customer has been with the bank.
- CheckingBalance - a numeric vector, the amount currently held in checking.
- SavingBalance - a numeric vector, the amount currently held in savings (0 or larger).
- Income - a numeric vector, yearly income in thousands of dollars.
- Homeowner - a factor with levels No Yes.
- Age - a numeric vector.
- Area.Classification - a factor with levels R S U for rural, suburban, or urban.

2. ANALYSIS I

Tables

For this analysis of two categorical variables, we would like to understand if there is an association between area classifier and purchasers. For example, this analysis can show if someone classified in a suburban area is more likely to purchase than someone in a rural area.

The variables which will be used for this analysis are:

- `data$Purchasers`
- `data$Area.Classification`

Question: Is there an association between purchase and area classifier?

Frequency Table

We generate a frequency table of purchasers and the area.

```
          No Yes
Rural     74  43
Suburban 109  66
Urban    148  60
```

Based on the frequency table, most of the data are non-purchases.

Relative Frequency Table

A relative frequency table was created to summarize the data of the two categorical variables and their distribution (like a frequency table).

```
            No   Yes
Rural    0.148 0.086
Suburban 0.218 0.132
Urban    0.296 0.120
```

Contingency Table

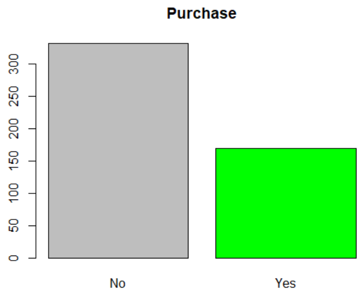
Creating a contingency table allows us to understand data distribution across each group and categorical values. It is also a great way to get the total sample size of the data.

```
          No Yes Sum
Rural     74  43 117
Suburban 109  66 175
Urban    148  60 208
Sum      331 169 500
```

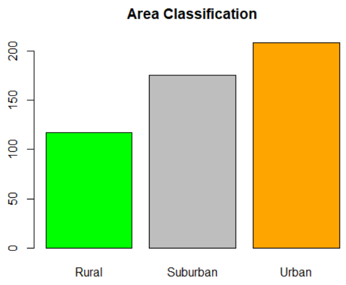
Plot

This section will contain visuals done using R. We will understand the categorical variables and how they are distributed between purchasers and area classification.

Bar Chart



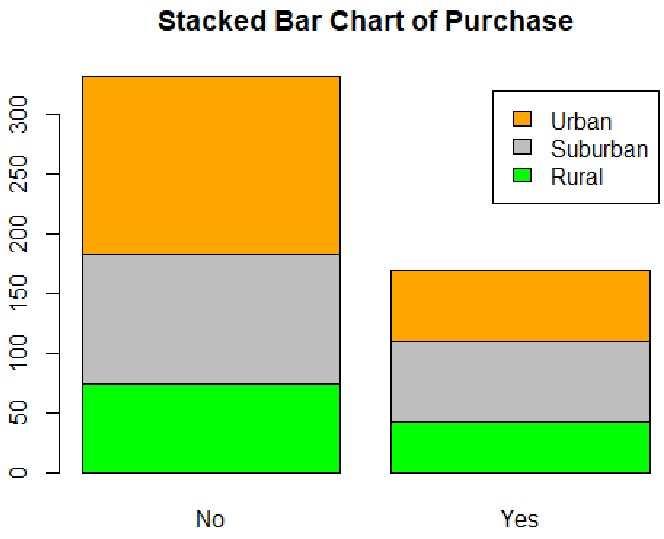
A significant portion of customers are non-purchasers of the account.



Customers classify more to be in an urban area.

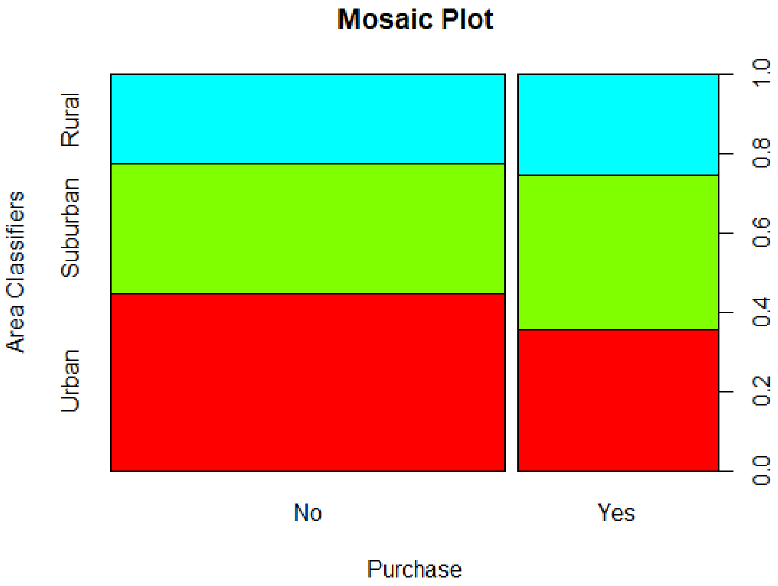
Stacked Bar Chart

The following section has additional visuals of purchases and area classifications.



The stacked bar chart allows us to understand the size and distribution of the data of purchasers and non-purchasers in the different area classifiers.

Mosaic Plot



The mosaic plot shows that many customers did not purchase an account, and there is a weak association. Next, we will test our null hypothesis (H_0) statistically.

Statistical Testing – Chi-Square Test

H_0: The area classifier is independent of purchasing behavior.
H_1: The area classifier is dependent upon purchasing behavior.
0.05 level of significance, degree of freedom of 2, and a critical value of 5.99.
The expected value table is generated using `chisq.test()` and `result$expected:`

```
              No    Yes
Rural     77.454 39.546
Suburban 115.850 59.150
Urban    137.696 70.304


X-squared = 3.9353, df = 2, p-value = 0.1398
```

Test Statistic: 3.9353

Convert Test Statistic to P-Value: `1-pchisq(3.9353, 2)` **≈ 0.139785.**

Running a permutation test of 1000, we get the following result:

```
                No       Yes
Rural    0.6324786 0.3675214
Suburban 0.6228571 0.3771429
Urban    0.7115385 0.2884615
Marginal 0.6620000 0.3380000

Permutation procedure:
Discrepancy          Estimated p-value
3.935267             0.129

With 1000 permutations, we are 95% confident that: the p-value is between 0.109 and
0.151.
```

Conclusion: Based on the result, we fail to reject the null hypothesis because, with 1000 permutations, we are 95% confident that the p-value is between 0.109 and 0.151, which is > 0.05. The marginal distribution is also very similar, and the mosaic table did show a weak association.

We accept the null hypothesis, which is that the area classifier is independent of purchasing behavior.

3. ANALYSIS II

For this analysis, we will be using one categorical and quantitative variable. We want to see if there's an association between purchasers and income.

The two variables in these experiments are:

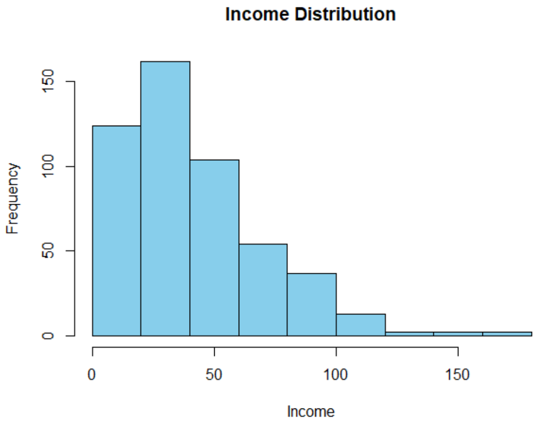
- `data$Purchasers`
- `data$Income`

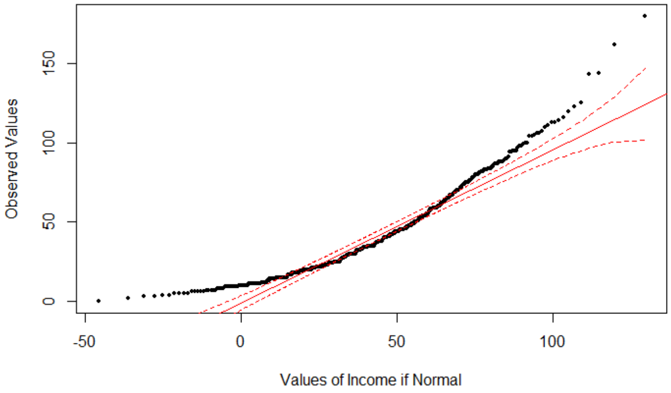
Question: Is there an association between the purchasers and income?

This analysis can help the bank understand if a person's income is associated with the account's purchase. Also, from a bank's perspective, this information can be valuable for their marketing campaigns.

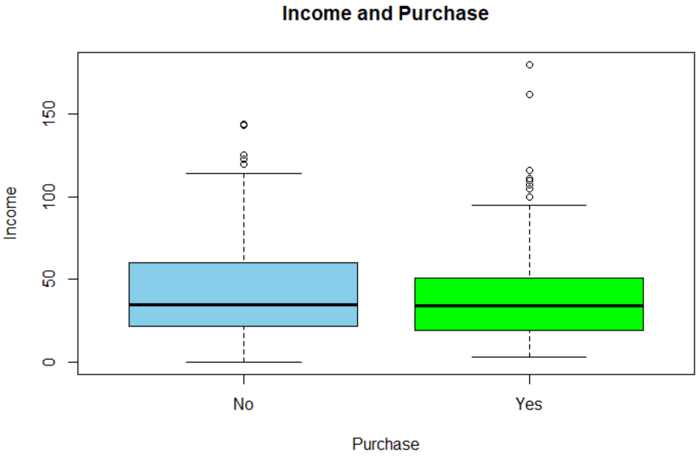
Histogram & QQ Plot for Income Distribution

The histogram shows that the distribution of income is right-skewed. The QQ plot also displays the distribution of income not to be normal. For the following reasons, the **median test** will be used.





Box Plot



The box plot does not show an association, as the means and medians are closely similar. Therefore, Mood's Median test will be conducted in the following steps for further analysis.

Median Test – Mood's Median Test

H$_0$: The median income is equal between the purchasers.
H$_1$: The median income is not equal between the purchasers.

We used a for-loop for Mood's Median Test to determine if an income was above or below the median. This resulted in the following table:

```
    Above Below Sum

No    164   167 331
Yes    80    89 169
Sum   244   256 500
```

Calculating the expected values was straightforward. The result returned using the chi-square test gave us the following output.

```
X-squared = 0.13912, df = 1, p-value = 0.7092
```

Convert Test Statistic to P-Value: `1-pchisq(0.13912, 2) ≈ 0.7092.`

Conclusion: Based on the results, the null hypothesis is accepted because the test statistic gave us a score of 0.13912, far from the reject area. Based on Mood's Median Test and the 5% confidence level, no statistical evidence exists of an association between purchase behavior and income.

4. ANALYSIS III

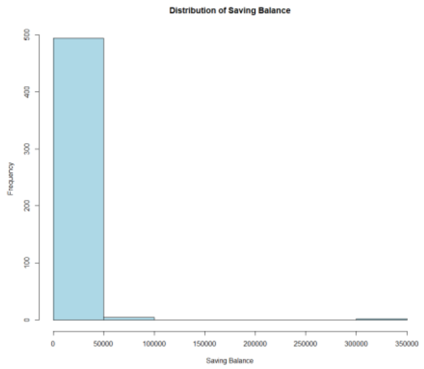
This final analysis is the association between two quantitative variables. For this test, we would like to see if there is an association between income and savings accounts.

The two variables in these experiments are:

- `data$Income`
- `data$SavingBalance`

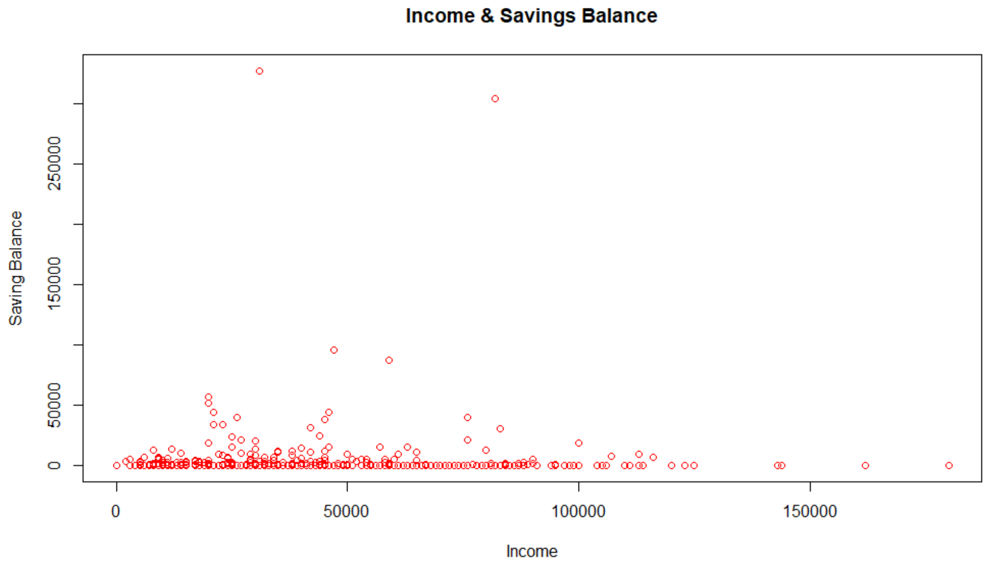
Question: Is there an association between income and saving balance?

H$_0$: There is no association between income and saving balance.
H$_1$: There is an association between income and saving balance.

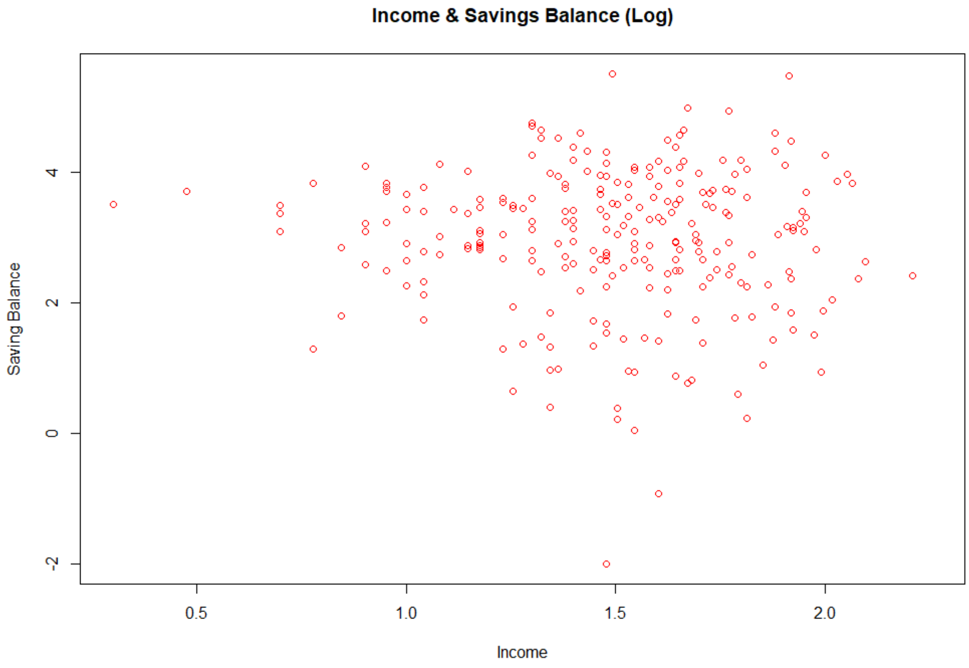


Plot: Creating a scatter plot on income and saving shows a few outliers above the 250k savings threshold. However, after scaling the data using the log function, the distribution of the data seems to have a weak distribution.

Scatter Plots

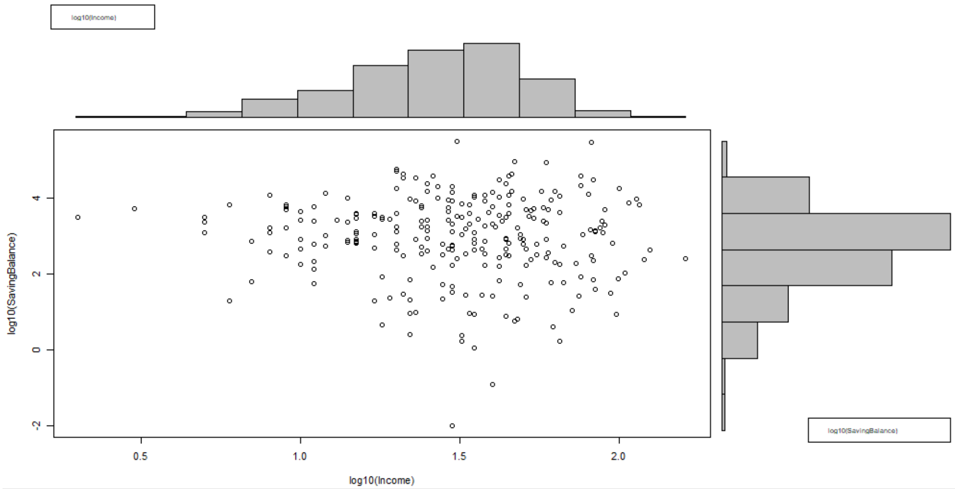


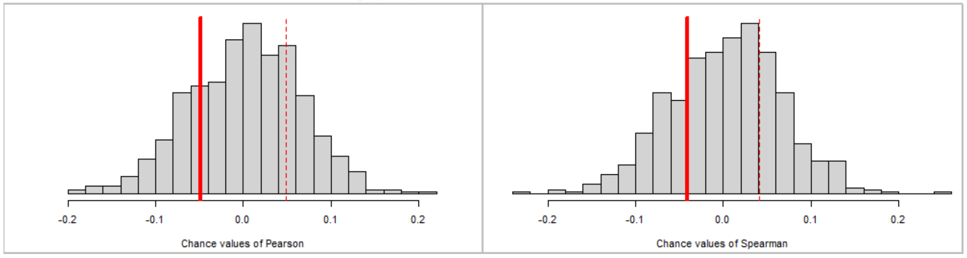
This scatter plot needs to give us a complete understanding of the distribution. It would be better to scale the data using the log function.



The scatter plot shows a nonlinear monotonic form of relationship, and the plot also displays outliers and heteroscedasticity.

Spearman’s Rank Correlation





Using the `associate()` function and increasing the `permutations` to 1000 results:

```
                                     Value      Estimated p-value
Pearson's r                    -0.04859340      0.466
Spearman's rank correlation    -0.04126306      0.521

With 1000 permutations, we are 95% confident that:
The p-value of Pearson's correlation (r) is between 0.435 and 0.497.
the p-value of Spearman's rank correlation is between 0.49 and 0.552.
```

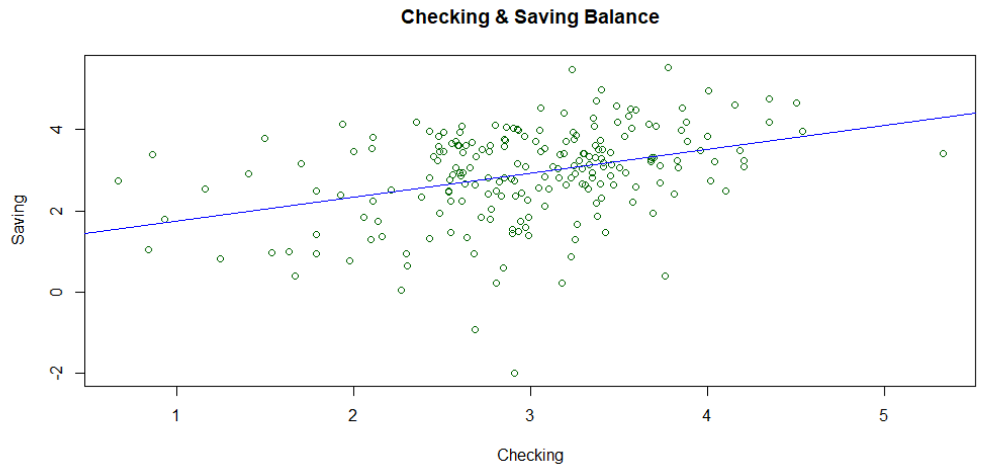
Conclusion: After running the Spearman's rank correlation test on the data, we found a correlation coefficient of -0.041, indicating a weak negative association between income and saving balance. Since the data appears to have a monotonic relationship, Spearman's rank correlation was an appropriate measure of association. However, with 1000 permutations, we are 95% confident that the p-value of the correlation coefficient is between 0.49 and 0.552, suggesting that the weak correlation we observed may have occurred by chance. Therefore, we fail to reject the null hypothesis and conclude that there needs to be more evidence to support the claim that there is a significant association between income and saving balance.

Simple Linear Regression

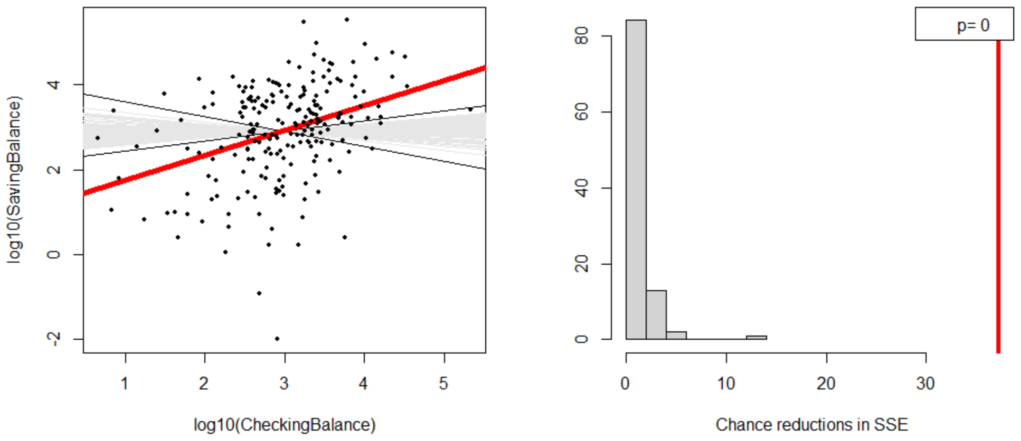
The next experiment will be a simple linear regression modeling the distribution of a checking and saving balance. This experiment will tell us if there is a linear relationship and if the connection is strong or weak.

H$_0$: Checking and saving balances do not correlate.
H$_1$: Checking and saving balances correlate.

Note: The data is filtered for balances more significant than 0 and values that are not nan. The data is also scaled.



The result shown is the linear model with the line of best fit.



The visual shown is 1000 permutations on the dataset. The red line on the left is the observed regression in the data. The slope is steep compared to what happens "by chance," and the SSE reduction is much more significant than "by chance." The regression is statistically significant.

```
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)               1.1583     0.3142   3.686  0.00029
log10(CheckingBalance)    0.5869     0.1027   5.718 3.66e-08

Residual standard error: 1.068 on 211 degrees of freedom
Multiple R-squared:  0.1341,    Adjusted R-squared:   0.13
F-statistic: 32.69 on 1 and 211 DF,  p-value: 3.66e-08
```

Simple Linear Regression: $\mu_{y|x} = 1.1583 + 0.5869x$

It would be helpful to convert the numbers to non-logarithmic and in thousands of dollars to understand them better. The simple linear regression equation would look like this: $\boldsymbol{\mu_{y|x} = 14{,}397.92 + 3{,}862.78x}$. This means for every dollar in the checking balance, an increase of about $3,862.78.
Things to note by this model:

1. The RMSE is 1.068 at 211 degrees of freedom, or if you convert it to the dollar amount, $11,694 away from the model's predicted line.
2. The R^2 value of 0.1341 indicates that the checking balance can explain approximately 13% of the variance in the savings balance.
3. The f-statistic tells us that variance does exist in the model.
4. The p-value of 3.66e-08 is much smaller than the significance level of 0.05, indicating that we can reject the null hypothesis and conclude that there is a significant linear relationship between the checking balance and savings balance.

ANOVA Test

```
                         Df  Sum Sq Mean Sq F value    Pr(>F)
log10(CheckingBalance)    1  37.279  37.279  32.691 3.66e-08
Residuals               211 240.618   1.140
```

The SSE of the regression is 240.618, and the SSR, which is the reduction in SSE, is 37.279. The p-value is again < 0.05.

Confidence Interval

```
                           2.5 %    97.5 %
(Intercept)            0.5387950 1.7777314
log10(CheckingBalance) 0.3845673 0.7892793
```

Conclusion: From the test, we are 95% confident that the "true slope" is between 0.3845873 and 0.7892693. Based on all the evidence, the regression line is statistically significant.