A Regression Analysis of Credit Card Expenditure

Thi Thu Ha Vo

December 11, 2024

1 Introduction

1.1 Problem description

Understanding credit card expenditure is crucial for financial institutions, as it helps evaluate risk effectively and potentially improve customer retention. By analyzing spending patterns, financial institutions can better understand their customers' behavior, which can lead to more informed decisions about credit offers and loan approvals. This project aims to apply regression modeling techniques to identify and model the factors that influence monthly credit card expenditure.

1.2 About the data set

This project uses the dataset called `CreditCard` from the `AER` package in R, which contains cross-sectional data on the credit history of a sample of applicants for a type of credit card. The data was collected by Greene (2003) and includes 1,319 observations, with 11 predictors and 1 response variable, as described below.

1. `card`: Factor. Was the application for a credit card accepted?
2. `reports`: Number of major derogatory reports.
3. `age`: Age in years plus twelfths of a year.
4. `income`: Yearly income (in USD 10,000).
5. `share`: Ratio of monthly credit card expenditure to yearly income.
6. `owner`: Factor. Does the individual own their home?
7. `selfemp`: Factor. Is the individual self-employed?
8. `dependents`: Number of dependents.
9. `months`: Months living at current address.
10. `majorcards`: Number of major credit cards held.
11. `active`: Number of active credit accounts.
12. `expenditure`: **Response variable**, average monthly credit card expenditure.

2 Analysis process

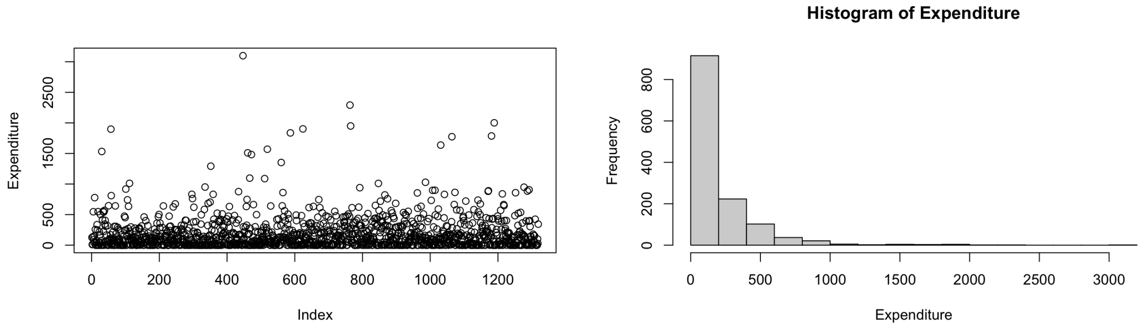
2.1 Exploratory data analysis

We start by loading the data, checking for missing values, and looking for duplicated rows. Fortunately, the data is clean with no missing values or duplicated rows. We also obtained a summary of the data, which is shown below.

Table 1: Summary statistics of the CreditCard dataset

	Min	1st Qu.	Median	Mean	3rd Qu.	Max
card	no: 296, yes: 1023					
reports	0.0000	0.0000	0.0000	0.4564	0.0000	14.0000
age	0.1667	25.4167	31.2500	33.2131	39.4167	83.5000
income	0.210	2.244	2.900	3.365	4.000	13.500
share	0.0001091	0.0023159	0.0388272	0.0687322	0.0936168	0.9063205
expenditure	0.000	4.583	101.298	185.057	249.036	3099.505
owner	no: 738, yes: 581					
selfemp	no: 1228, yes: 91					
dependents	0.0000	0.0000	1.0000	0.9939	2.0000	6.0000
months	0.00	12.00	30.00	55.27	72.00	540.00
majorcards	0.0000	1.0000	1.0000	0.8173	1.0000	1.0000
active	0.000	2.000	6.000	6.997	11.000	46.000

We are interested in studying the distribution of the response variable `expenditure` by checking both its scatterplot and histogram. Based on the visualizations, it appears that the distribution of the response is right-skewed, and the range of values is relatively large. The majority of the data points are concentrated within the range of 0 to 500, while the overall range extends from 0 to 3000. Given the wide range of values, it may be appropriate to consider applying a transformation to the response variable in the next step of our analysis to improve model performance.

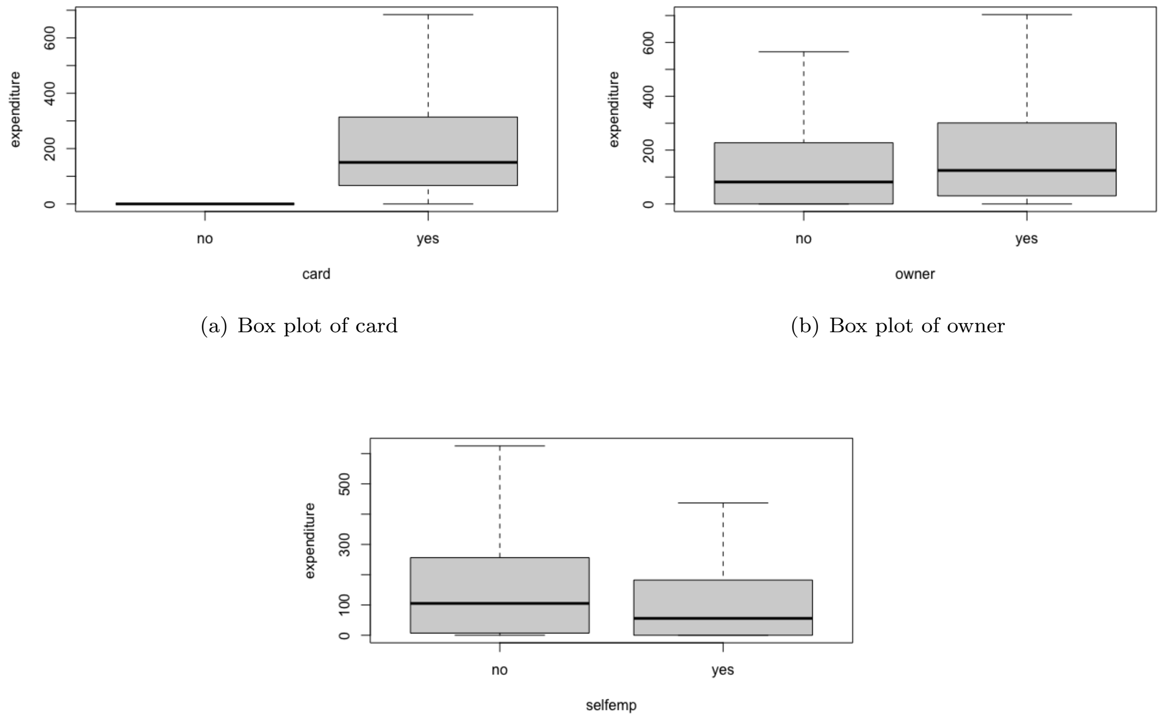


(a) Scatter plot of expenditure

(b) Histogram of expenditure

Figure 1: Scatter plot and the histogram of the response

Since the predictor variables consist of both numerical values and categorical factors, we can start by using a Box Plot to examine the relationship between the response variable and the factor predictors. For the numerical predictor variables, scatter plots are more appropriate.



(a) Box plot of card

(b) Box plot of owner

(c) Box plot of selfemp

Figure 2: The box plots for 3 factor variables

As shown in the box plot for `card`, the value of `expenditure` is always 0 when the factor `card` is set to `no`. Therefore, we are only interested on the rows where `card` equals `yes`, since for the `no` group, the `expenditure` is consistently 0.

The scatter plot between the numerical variables and the response shows that `share` appears to have the strongest linear relationship with `expenditure` with the increasing variance, followed by `income` and `reports`. As for the relationships between the predictor variables, it is difficult to identify clear patterns because most of the data points are concentrated near the origin. The plot also shows one outlier in the response variable and some noticeable outliers in `age`. We decide to remove these outliers since `expenditure` values over 2500 were excluded as they are much larger than others and could skew the results. Moreover, rows with `age` < 5 are also removed because such values are unrealistic and likely errors.

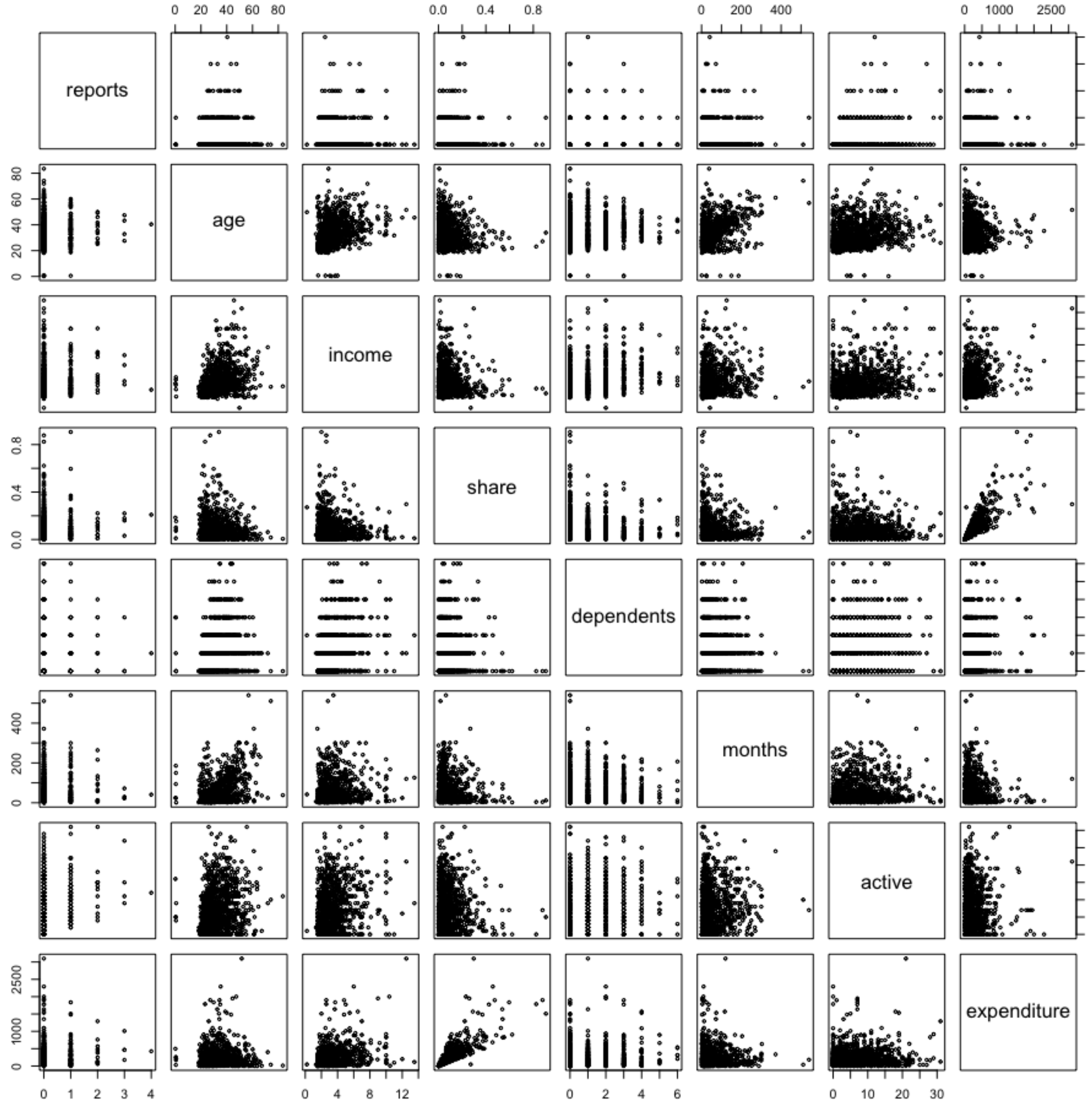
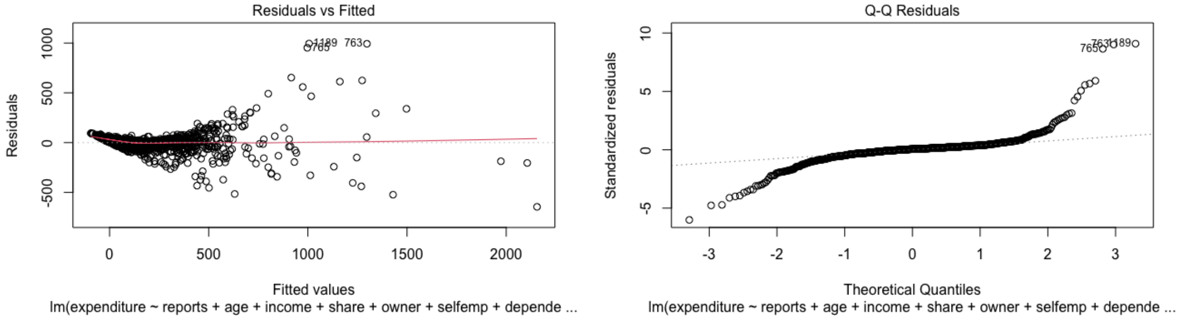


Figure 3: Scatterplot between the response and other numerical variables

2.2 Diagnostics and transformation

First, we fit a simple linear regression model with all predictors as the starting point. Then, we plot the residuals versus the fitted values and create a Q-Q plot.



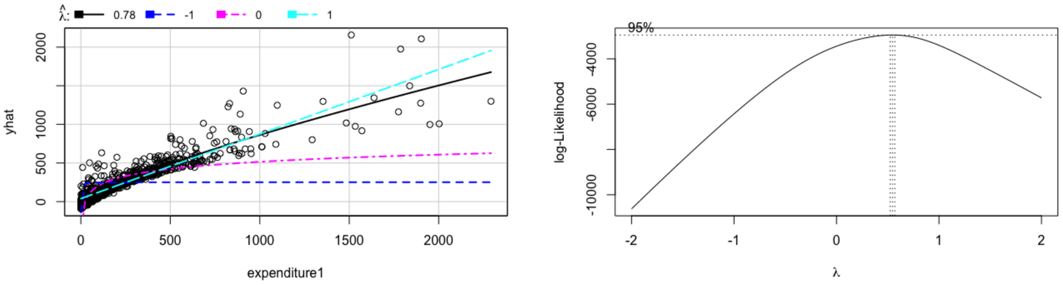
(a) Residual plot

(b) Q-Q plot

Figure 4: The residual plot and Q-Q plot for the baseline model

This plot shows that the baseline model does not meet the assumptions of linearity and constant variance. The curved pattern in the plot and the small residuals when the response is small suggest a non-linear relationship. Additionally, the spread of the residuals increases as the fitted values grow larger. Therefore, methods like transformation might help address these issues. Regarding the Q-Q plot, The residuals in the left side are smaller than expected, while the residuals in the right side are larger than expected, this suggests that the residuals are not normally distributed.

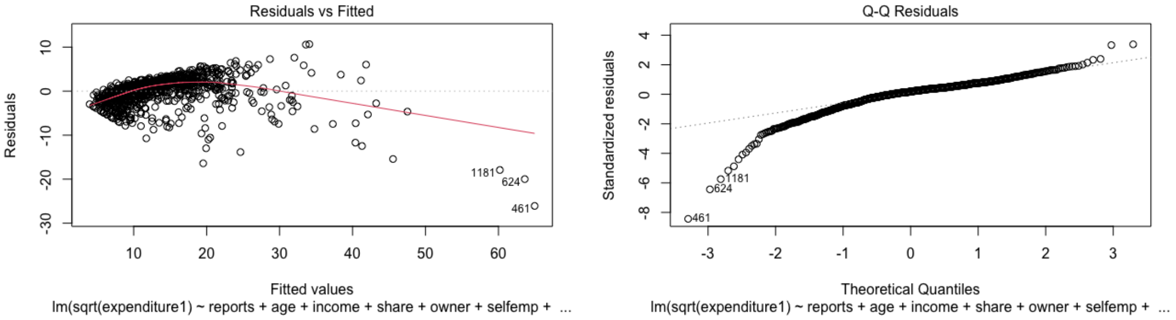
One approach that might help in this case is finding the appropriate transformation for both the response and the predictors. From the scatter plot (Figure 3), we observe that transformations can only be applied to continuous numerical variables with a wide range. In this case, the relevant variables are `expenditure`, `months`, `income`, `share`, and `age`. First, we focus on finding the appropriate transformation for the response variable using both the `inverseResponsePlot` and the Box-Cox method. The `inverseResponsePlot` suggests a value of $\lambda = 0.78$, while the Box-Cox method suggests $\lambda \approx 0.5$. Based on this, we decide to use $\lambda = 0.5$, which corresponds to the square root transformation.



(a) Inverse response plot

(b) Box-Cox method

Figure 5: The inverse response plot and the box-cox plot for the response



(a) Residual plot

(b) Q-Q plot

Figure 6: The residual plot and Q-Q plot after applying transformation for the response

After applying the square root transformation, the residuals versus fitted values plot looks more random and less patterned. This suggests that the transformation has reduced heteroscedasticity and improved the linearity of the model. The Q-Q plot shows a better alignment of the residuals in the tails. This indicates that the residuals are closer to a normal distribution after the transformation. Moreover, we obtained some points that seems to be the influential. Based on this Diagnostic plot, we decide to remove the point 461, 624 and 1181 as the influential.

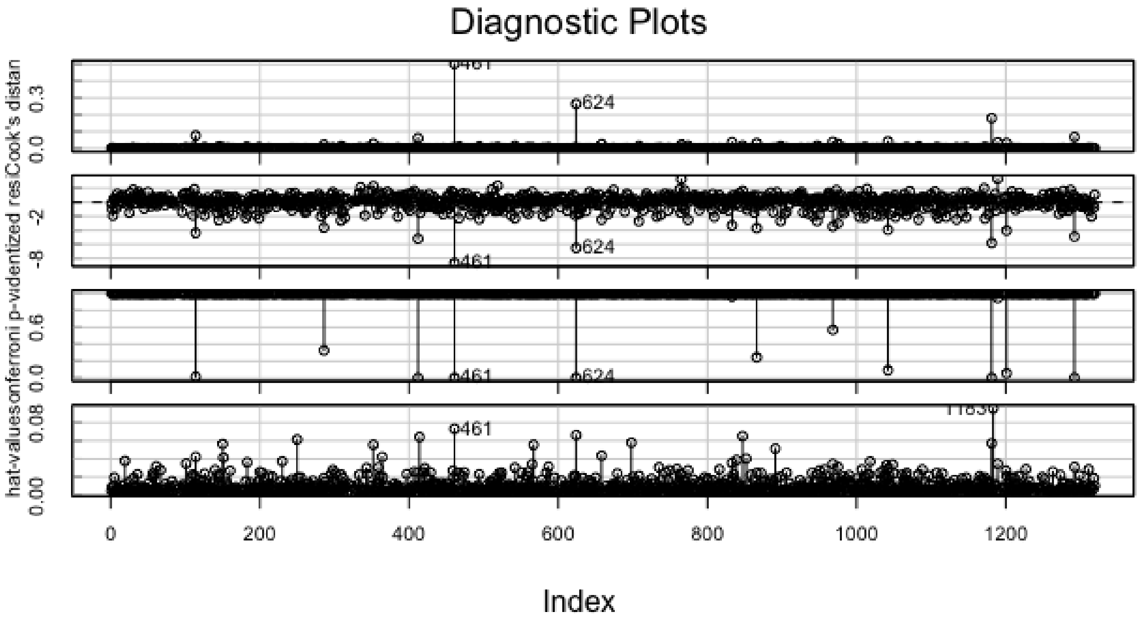


Figure 7: Diagnostic plots after applying transformation for the response

We are also interested in applying transformations of the predictors. Among these variables, `expenditure` and `months` have zero values, so we add a small constant to make the transformations work. We use the `powerTransform` function to help choose the transformations. We decide to use the inverse square root for `age`, and the logarithm for `income` and `share`.

Table 2: bcPower Transformations

Variable	**Est Power**	**Rounded Pwr**	**Wald Lwr Bnd**	**Wald Upr Bnd**
age	-0.5471	-0.50	-0.7494	-0.3448
income	-0.1146	-0.11	-0.1967	-0.0325
share	0.2610	0.26	0.2228	0.2992
months1	0.1209	0.12	0.0796	0.1622

Comparing the models using the summary table, we see a small improvement in terms of residual standard error (RSE), R^2, and R^2_{adj}. The RSE decreases from 3.267 to 3.168, and the adjusted R^2 increases from 0.8133 to 0.8249. Additionally, the coefficients for these predictor variables become more significant.

Table 3: Regression summary for sqrt(expenditure)
and other transformations on predictor variables

Variable	**Estimate**	**Std. Error**	**t-value**	**p-value**
Intercept	0.369469	0.459312	0.804	0.4214
reports	0.646051	0.248287	2.602	0.0094 **
1/sqrt(age)	0.000000	-	-	-
log(income)	6.776569	0.265409	25.533	$<$ 2e-16***
share	66.522992	1.027264	64.757	$<$ 2e-16***
owneryes	0.010400	0.227377	0.046	0.9635
selfempyes	-0.716618	0.419691	-1.707	0.0880
dependents	0.127791	0.088906	1.437	0.1509
log(months1)	-0.183044	0.085622	-2.138	0.0328 *
majorcards	-0.193244	0.276508	-0.699	0.4848
active	0.001326	0.017685	0.075	0.9403

Residual Standard Error: 3.168 on 1004 degrees of freedom
Multiple R-Squared: 0.8264
Adjusted R-Squared: 0.8249
F-Statistic: 531.2 on 9 and 1004 DF, p-value: < 2e-16

2.3 Variables selection

In this part, we apply both forward selection and backward elimination methods using the AIC criterion to identify the most suitable subset of predictor variables. Both methods produce the same subset of significant predictors. There is a slight improvement in model performance, with the residual standard error (RSE) decreasing from 3.165 to 3.161 and the adjusted R^2 increasing from 0.8249 to 0.8253. These results indicate that forward selection and backward elimination effectively remove irrelevant predictors while keeping the most significant ones.

Table 4: Summary of the model with the predictors from forward selection and backward elimination

Predictor	**Estimate**	**Std. Error**	**t-value**	**p-value**
Intercept	0.22664	0.40276	0.563	0.5738
reports	0.64383	0.24227	2.658	0.0080 **
log(income)	6.75877	0.25616	26.385	< 2e-16 ***
share	66.50575	1.02318	64.999	< 2e-16 ***
selfempyes	-0.71881	0.41873	-1.717	0.0864 .
dependents	0.13245	0.08637	1.534	0.1255
log(months1)	-0.17873	0.08275	-2.160	0.0310 *

2.4 Interaction

If we look closely at the description of the dataset, `share` is defined as the ratio of monthly credit card expenditure to yearly income. By multiplying `share` by `income`, we can create a new predictor variable that might be correlated with the response variable `expenditure`, which represents the average monthly credit card expenditure. Using the subset of variables from the previous part, we perform a partial F-test to determine whether this interaction term should be included in the model.

Table 5: ANOVA Table comparing models with and without interaction term

Model	**Residual DF**	**Residual SS**	**Sum of Squares (SS)**	**F-Statistic**	**p-value**
Without Interaction	1009	10081.4			
With Interaction	1008	6197.2	3844.2	631.78	$< 2.2 \times 10^{-16}$

The ANOVA table shows that the p-value is extremely small, indicating that we should reject the null hypothesis. In other words, the interaction term significantly improves the model and provides additional explanation for the response variable `expenditure`.

3 Final model summary and interpretation

First, we summarize the final model after applying variable selection and including the interaction term. This summary will help us determine the key predictors that significantly influence the response variable.

Table 6: Regression summary of the final model for sqrt(expenditure)

Variable	**Estimate**	**Std. Error**	**t-value**	**p-value**
Intercept	4.4384	0.35807	12.395	$< 2\text{e-}16^{***}$
reports	0.35612	0.19053	1.869	0.0619
log(income)	2.43675	0.26464	9.208	$< 2\text{e-}16^{***}$
share	22.49589	1.92698	11.674	$< 2\text{e-}16^{***}$
selfempyes	-0.64391	0.32873	-1.959	0.0504
dependents	0.02319	0.06797	0.341	0.7330
log(months1)	-0.03365	0.06525	-0.516	0.6062
share:income	14.99263	0.59672	25.125	$< 2\text{e-}16^{***}$

Residual Standard Error: 2.481 on 1006 degrees of freedom
Multiple R-Squared: 0.8933
Adjusted R-Squared: 0.8926
F-Statistic: 1203 on 7 and 1006 DF, p-value: < 2e-16

As we can see, the residual standard error (RSE) has significantly decreased from 3.161 to 2.481, indicating an improved fit of the model. Both R^2 and R^2_{adj} have also increased notably, from 0.8264 to 0.8933 and 0.8253 to 0.8926, respectively. These improvements demonstrate that the final model better explains the variability in the response variable compared to the previous models (Table 4). And the final model can be written as:

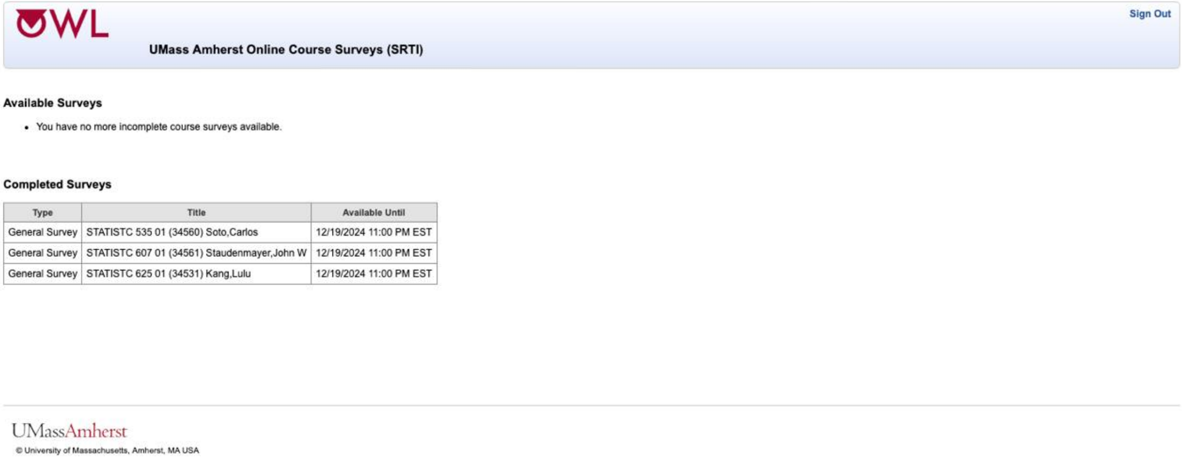
$$\begin{aligned}\sqrt{\texttt{expenditure}} = {} & 4.4384 + 0.35612\texttt{reports} + 2.43675\log(\texttt{income}) \\ & + 22.49589\texttt{share} - 0.64391\texttt{selfempyes} + 0.02319\texttt{dependents} \\ & - 0.03365\log(\texttt{months1}) + 14.99263\texttt{share}\cdot\texttt{income}.\end{aligned}$$

where `selfempyes` is a factor with two levels, 0 and 1, and `months1` is the sum of `months` and a small constant.

The summary table shows that log (`income`), `share`, and the interaction between `income` and `share` are the most significant predictors, with strong positive effects, as their p-values are very small. In this case, as income and the ratio of monthly credit card expenditure to yearly income increase, the square root of expenditure also increases. Their product is even more interpretable, as it captures the average expenditure. The factor `selfempyes` also shows a correlation with the response variable. If we have a look at the Box Plot (Figure 2) in the EDA section, people who are self-employed seem to spend less on average.

4 Conclusion

In this project, we study the effect of several predictors to determine the monthly credit card expenditure. We begin by exploring the data, diagnosing the linearity assumption and normal distribution, and identifying outliers and influential points. Additionally, transformations are applied to both the response variable and the predictors. Variable selection is performed to identify the best subset of predictors, and interaction terms are tested to improve the model's significance. The results show that the average monthly expenditure has a strong relationship with income, the ratio between expenditure and income, and self-employment status. These findings are reasonable and align with general intuition.



UMass Amherst Online Course Surveys (SRTI)

Available Surveys

- You have no more incomplete course surveys available.

Completed Surveys

Type	Title	Available Until
General Survey	STATISTC 535 01 (34560) Soto,Carlos	12/19/2024 11:00 PM EST
General Survey	STATISTC 607 01 (34561) Staudenmayer,John W	12/19/2024 11:00 PM EST
General Survey	STATISTC 625 01 (34531) Kang,Lulu	12/19/2024 11:00 PM EST

UMassAmherst

© University of Massachusetts, Amherst, MA USA