The Relation Between Credit Limits, Income and Age

Siddharth Nand (76648070),

Helin Wang (73965865),

Azure Gao (82789934)

Introduction

Credit card companies use different criteria to set credit limits for their customers. Two important factors that influence credit card limits are the age and income level of the customer. In this data analysis assignment, we will explore the relationship between the credit card limit, age and income level of customers.

This relationship can be of interest to individuals who are applying for credit cards or seeking to increase their credit limits. By understanding how their age and income level may impact their credit limit, individuals can make more informed decisions about how to manage their credit.

Secondly, this relationship can also have broader implications for financial policy and economic research. By examining patterns and trends in credit card limits across different age and income groups, researchers can gain insights into broader economic trends and issues related to income inequality, financial access, and consumer behaviour.

We will use the "Predicting Credit Card Customer Segmentation" dataset from Kaggle. The information was gathered in six months and consists of 10,000 records and 23 columns. The dataset contains records for each client, each of which contains details about the customer's age, gender, marital status, level of education, income, and credit limit. Additionally, the dataset contains details about how the customer uses their credit card, including their average monthly cost, the number of transactions they make each month, and the length of time they have been customers.

Analysis

Definitions

$Let\ Y\ \in [1438,\ 34516]\ be\ a\ person's\ credit\ limit$, this is our response variable

Our exploratory variables are:

$Let\ X_1 \in [26,\ 73]\ be\ a\ person's\ age$

$$Let\ X_2 = \begin{cases} 0, & "Unknown" \\ 1, & "Less\ than\ \$40K" \\ 2, & "\$40K\ -\ \$60K" \\ 3, & "\$60K\ -\ \$80K" \\ 4, & "\$80K\ -\ \$120K" \\ 5, & "\$120K\ +\ " \end{cases} \quad be\ a\ person's\ income\ category$$

We will be using two additive linear models:

$$(1)\quad Y_1 = \beta_{0,1} + \beta_{1,1}X_1 + \epsilon_1$$

$$(2)\quad Y_2 = \beta_{0,2} + \beta_{1,2}X_1 + \beta_{2,2}X_2 + \epsilon_2$$

$$where\ \beta_{2,2} = (b_{2,1},\ ...,\ b_{2,6})^T\ is\ a\ vector\ of\ slopes\ for\ each\ income\ category$$

Visualizations

Pre-Transformed Models

Below, figures 1 and 2 are the non-transformed residual vs fitted plots and Normal Q-Q plots for models 1 and 2 respectively. You can see that our standard residuals are not normally distributed, so we can not apply a linear model yet. Furthermore, the residual mean is more biased towards negative residuals, signifying a loosely fitted model. Therefore, we must apply a transformation to make our model more linear and have a tighter fit.

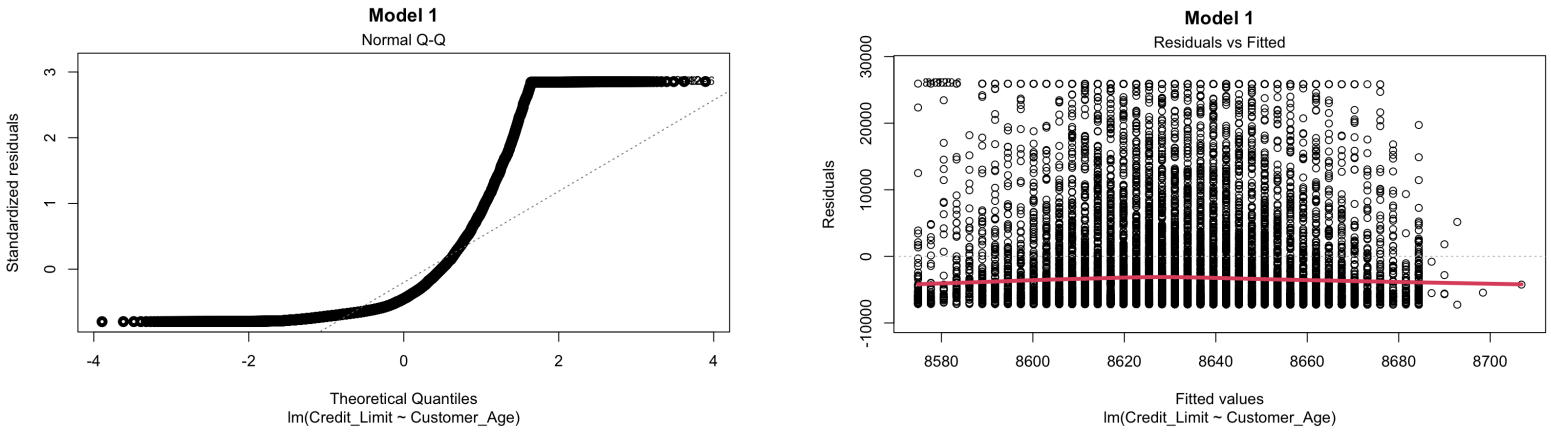

Figure: 1: Untransformed Normal Q-Q and Residual vs Fitted plots for model 1

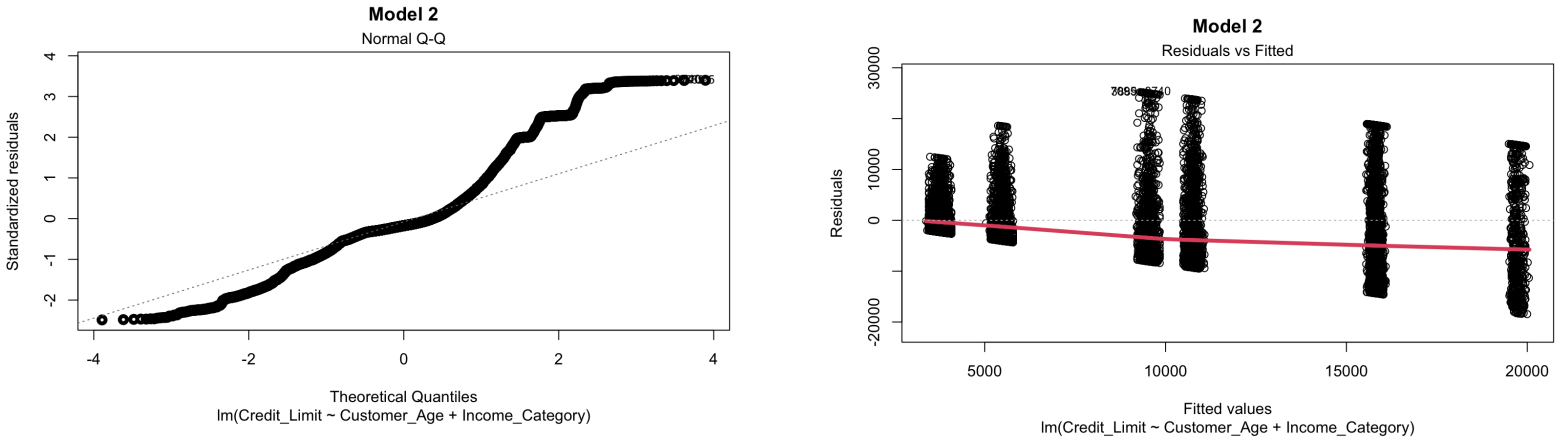

Figure 2: Untransformed Normal Q-Q and Residual vs Fitted plots for model 2

Transformed Models

After playing around with various functions, we will be using applying the tan function to the response variable for model (1) and the log function to the response variable for model (2). Figures 3 and 4 demonstrate an improvement in our model fit. For model 1, the residual vs fitted value has a mean of 0, compared to the untransformed plot in figure 1 and the Normal Q-Q plot is a better fit, however, it has heavy tails at the ends similar to the untransformed Normal Q-Q plot. For model 2, the standardized residuals are more normally distributed and the residual vs fitted value has a mean of 0.

Note: when referring to models (1) and (2) from now onwards, this paper will be referring to the transformed versions.

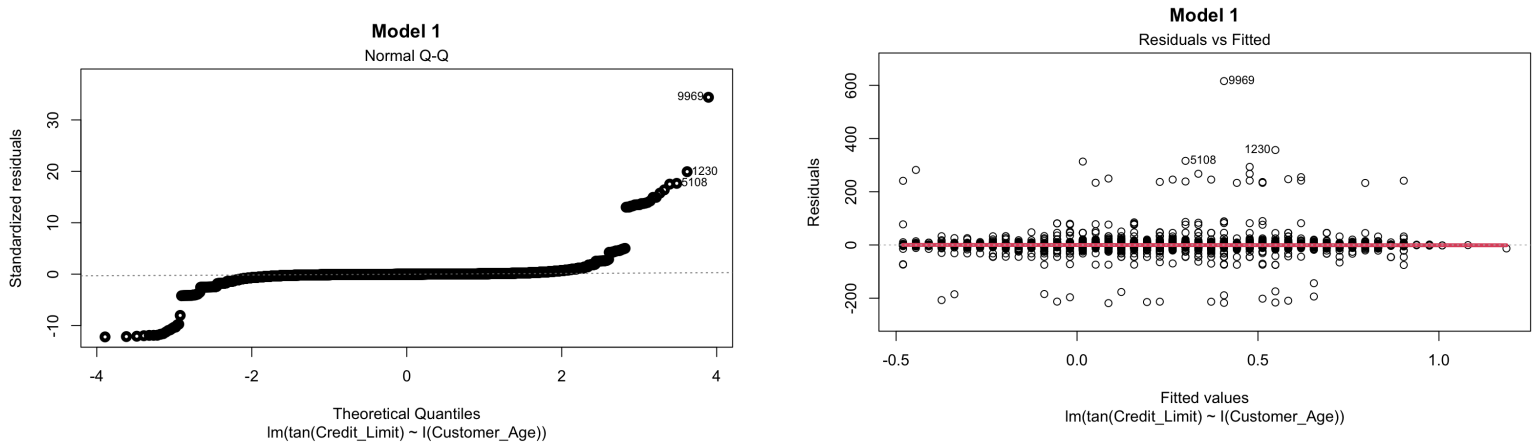

Figure 3: Transformed Normal Q-Q and Residual vs Fitted plots for model 1

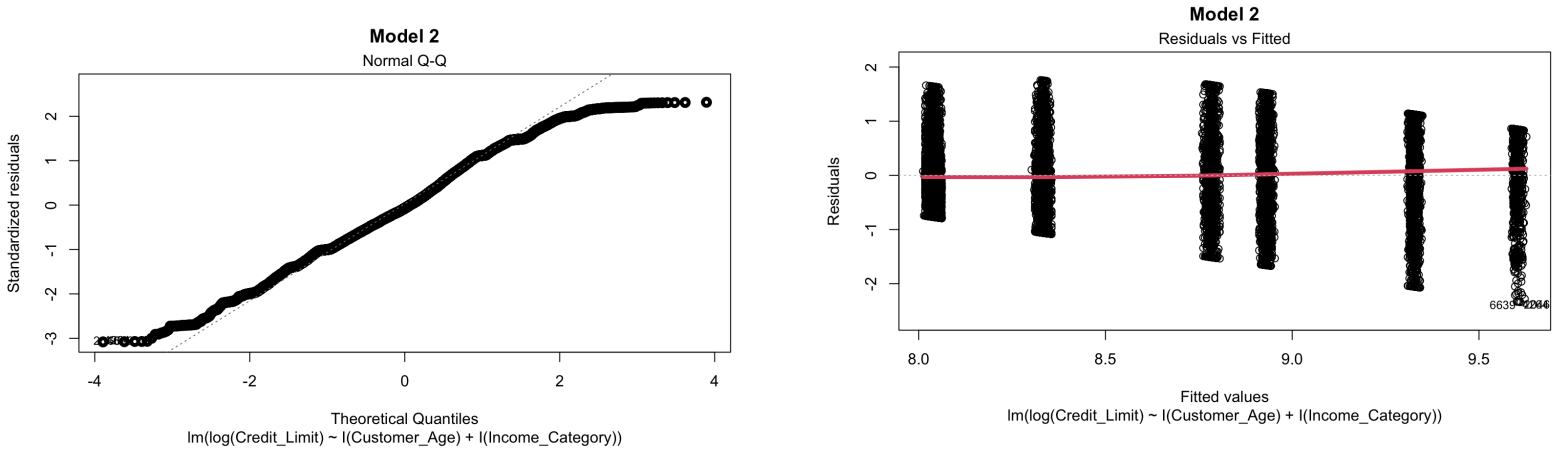

Figure 4: Transformed Normal Q-Q and Residual vs Fitted plots for model 2

Visualizing Models

Below are visualizations of models (1) and (2). Figure 5 is for model (1). It shows that there is no linear relationship between age and credit limit defined by the visually flat slope. From the plot above, we can clearly find that the income category indeed impacts the person's credit limit along with age. Figure 6 is for model (2), it shows that there are non-horizontal slopes. So before running any hypothesis tests, it is expected that a customer's age does not have a linear relationship with their credit limit, however, if we

add a customer's income to the problem, then we have a linear relationship between a person's age, income and credit limit.

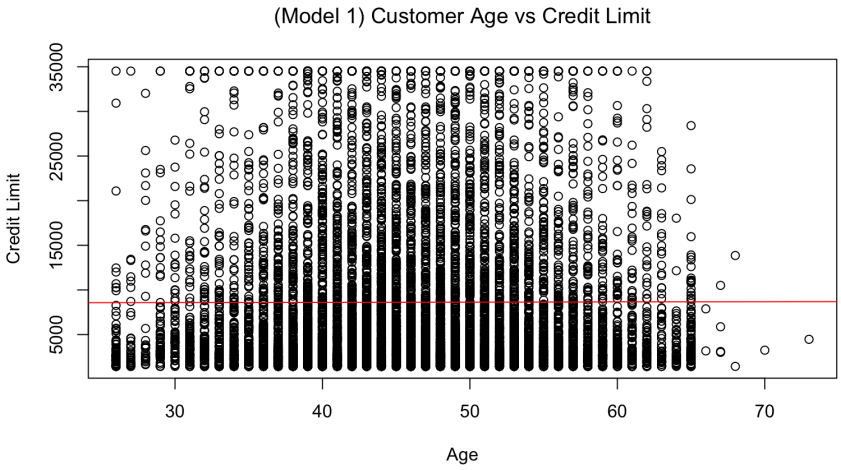

Figure 5: Linear regression plot of model 1

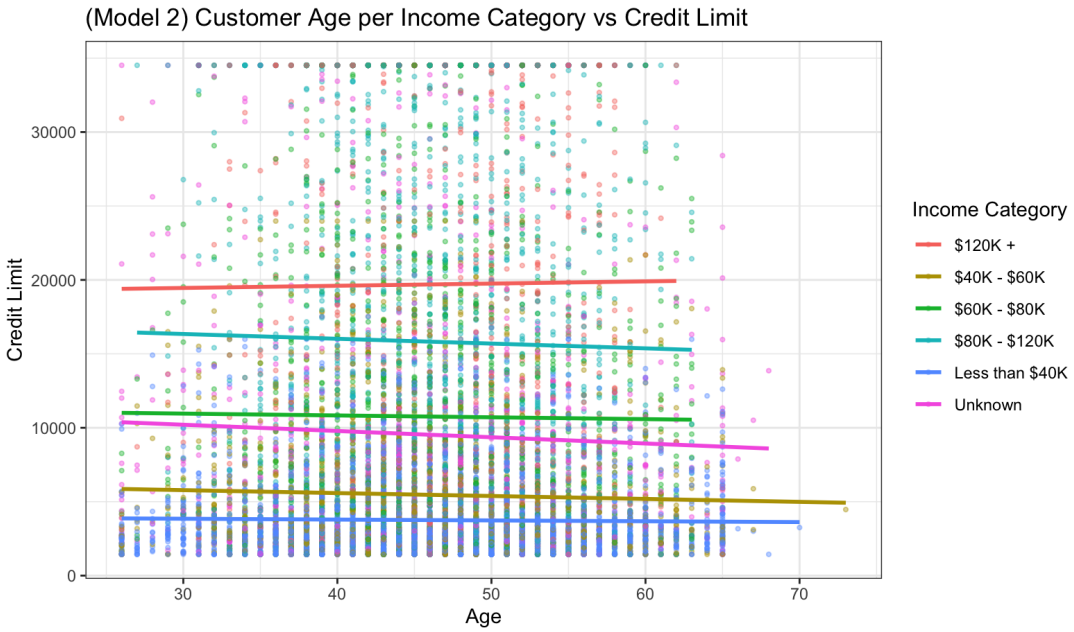

Figure 6: Linear regression plot of model 2

Hypothesis Testing

We will conduct two hypothesis tests:

I. *Whether a person's age does not vary with their credit limit against the alternative that there does exist a linear dependence between age and credit limit.*

$$H_0 : \beta_{1,1} = 0 \qquad H_A : \beta_{1,1} \neq 0$$

II. *Whether a person's age and income category do not vary with their credit limit against the alternative that there does exist a linear dependence between age, income and credit limit.*

$$H_0 : \beta_{1,2} = b_{2,1} = \ \ldots \ = b_{2,6} = 0 \qquad H_A : Not\ H_0$$

I. For the first hypothesis test

$H_0 : \beta_{1,1} = 0 \qquad H_A : \beta_{1,1} \neq 0$

The output from the `lm` function for model (1) is

	Estimate	Std. Error	t-value	P-value
$\beta_{0,1}$	8501.901	529.707	16.050	<2e-16
$\beta_{1,1}$	2.807	11.267	0.249	0.803

Multiple R-squared: 6.132e-06,
Adjusted R-squared: -9.263e-05

The p-value for $\beta_{1,1}$ is 0.803, assuming a 5% significance level, we do not reject H_0. Therefore, a person's age does not vary with their credit limit.

II. For the second hypothesis test

$H_0 : \beta_{1,2} = b_{2,1} = \ \ldots \ = b_{2,6} = 0 \qquad H_A : Not\ H_0$

We will first do a model comparison using ANOVA between models (1) and (2) to see if there is a significant difference. The results of the model comparison is:

	Res.Df	RSS	DF	Sum of Squares	F-statistic	P-value
Model (1)	10125	8.3646e+11				
Model (2)	10120	5.5796e+11	5	2.785e+11	1010.2	< 2.2e-16

Assuming a 5% statistical significance, the above results show that model (2) improved the model significantly.

The output from the `lm` function for model (2) is

	Estimate	Std. Error	t-value	P-value
$\beta_{0,2}$	10252.148	481.050	21.312	<2e-16
$\beta_{1,2}$	-15.896	9.215	-1.725	0.0846
$b_{2,1}$	-5761.802	255.078	-22.588	<2e-16
$b_{2,2}$	-4057.297	283.524	-14.310	<2e-16
$b_{2,3}$	1237.332	298.187	4.150	<2e-16
$b_{2,4}$	6295.752	292.407	21.531	3.36e-05
$b_{2,5}$	-10221.875	354.359	28.846	<2e-16

Multiple R-squared: 0.333,
Adjusted R-squared: 0.3326

The P-value for $\beta_{1,2}$ = 0.0846 and assuming a 5% significance level, we would reject the H_0. We also know that age is not significant for this model. Therefore, there does exist a linear dependence between income categories and credit limits.

Results and Discussion

In the first model, we explored the association between the age of customers and their credit limits. The hypothesis test shows a p-value of 0.803 for the coefficient of the age input variable. Therefore, we failed to reject the null hypothesis that a person's age does not affect his/her credit limit. Additionally, the R-squared and Adjusted R-squared values of this Simple Linear Regression are very small, suggesting that age alone does not do a good job of explaining the observations in our dataset. Therefore, we concluded that age alone does not play a significant role in determining credit limits.

For the second model, we created an additive multilinear regression model with response variable credit limit, and input variable age and income categories. The coefficients for income categories are significant with very low p-values, indicating that different income levels are associated with different credit limits. The coefficient for age is greater than 0.05, which suggests that age is not a significant factor and has little effect on a person's credit limit. We also created plots to help us assess the assumptions of linear models. To examine heteroscedasticity, we created residual plots. The residuals appear to be randomly distributed, while the plot shows the points to be uniformly scattered. Moreover, we created a QQ plot to evaluate the normality assumption. Most of the points near the middle are on the 45-degree dotted line. However, some points closer to the ends were not on the theoretical quantile line, indicating that our model is not perfectly normal. We can use techniques like data transformation to improve the normality of our data and the fit of our model.

The model has an Adjusted R-square of 0.3326, meaning the model accounts for over 30 percent of the variation in the dependent variable credit limit is explained by the independent variables age and income. While this is an improvement from the first model that only considered age, there is still room for improvement. To create a better model that explains more variation in the dataset, we need to include more variables in our model. We could use Forward Selection Method to find the best LR model to predict the customer credit limit.

The models we created give us a clear view of how credit limits are affected by customer age and income. In conclusion, our analysis suggests that income is a significant factor in determining credit limits while age is not. We can use this information to inform

credit decisions and develop better credit risk models. Further research could explore additional factors that affect credit limits, such as credit history and debt-to-income ratio, to create even more accurate predictive models.