AFRICAN INSTITUTE FOR MATHEMATICAL SCIENCES

(AIMS RWANDA, KIGALI)

Name: Jean de Dieu NGIRINSHUTI Assignment Number: 2

Course: Statistical Regression Date: November 23, 2024

Introduction

This report provides an analysis of lending data to predict default rates. The tasks include descriptive data analysis, building a regression model, interpreting the results, checking assumptions, and discussing possible improvements.

Descriptive Analysis

Summary statistics

The descriptive statistics for key numerical variables in the dataset are summarized in Table 1.

Table 1: Summary of Numerical Variables

Variable	Min	1st Quartile	Median	Mean	Max
Age	29.00	36.75	40.00	40.01	57.00
Experience	0.00	6.00	8.00	7.60	15.00
Earnings ($)	55,000	135,000	165,000	166,743	300,000
BadPastRecords	0.00	0.00	1.00	1.56	9.00
Defaults	0.00	8.75	15.00	18.17	69.00
Accounts	41.00	124.50	215.00	214.10	400.00
ObsRate	0.00000	0.05153	0.07871	0.08399	0.20008

The dataset provides an overview of the demographic distribution of loan applicants. There are 131 females and 173 males. In terms of residence, 178 individuals live in urban areas, while 126 reside in rural areas. Regarding ownership, 241 individuals do not own a home, and 194 do not own land, with 63 homeowners and 110 landowners. Finally, 231 individuals do not own a car, while 73 do. This breakdown offers valuable context for analyzing financial behaviors and loan risk.

Visualizations

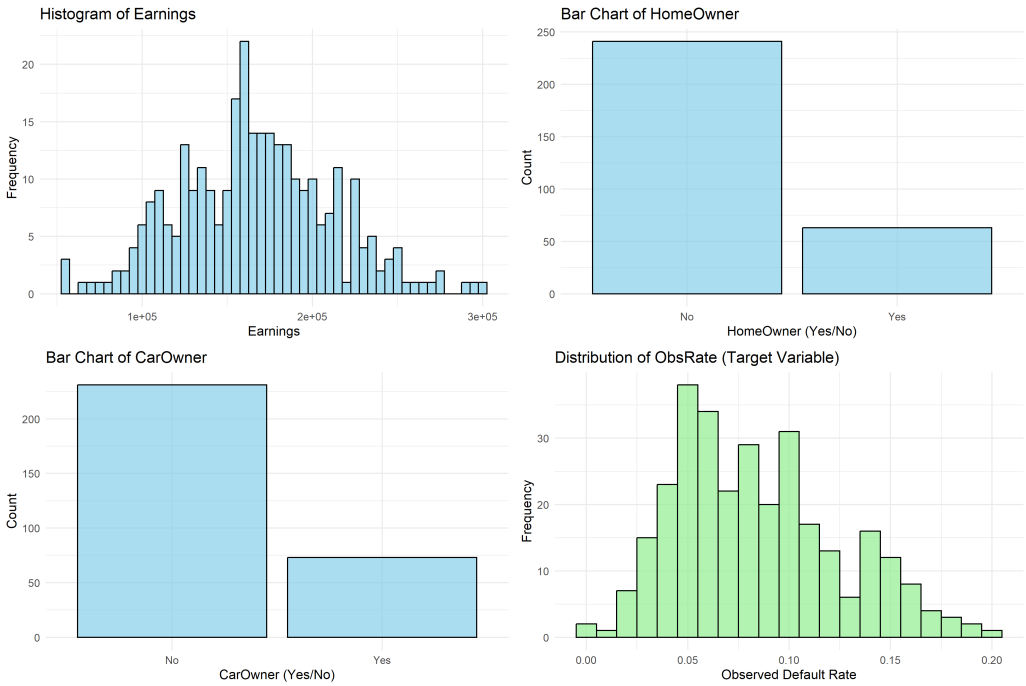

Figure 1: Key Visualizations

The visualizations in figure 1 offer a clear overview of the dataset. Most individuals earn between \$135,000 and \$195,000. Around 79% of the individuals are non-homeowners, and about 76% do not own a car. The default rate, or "ObsRate," has a mean of 0.08399, with most values falling between 0.05 and 0.1.

Model Fitting

Full Model Results

The initial model included all predictors and explained 98.13% of the variance in the observed default rate ($R^2 = 0.9813$, Adjusted $R^2 = 0.9807$). However, several predictors, including *Age*, *Gender*, *LandOwner*, and *CarOwner*, were not statistically significant ($p > 0.05$).

To fit the full model, the following equation was used for predicting the observed default rate (ObsRate):

$$\begin{aligned}\text{ObsRate} = {} & 0.1283 - 0.000101 \cdot \text{Age} - 0.000193 \cdot \text{GenderMale} - 0.002609 \cdot \text{Experience} \\ & - 2.644 \times 10^{-7} \cdot \text{Earnings} - 0.006069 \cdot \text{HomeOwnerYes} + 6.589 \times 10^{-5} \cdot \text{LandOwnerYes} \\ & + 0.01633 \cdot \text{BadPastRecords} - 7.165 \times 10^{-4} \cdot \text{CarOwnerYes} + 2.865 \times 10^{-4} \cdot \text{Defaults} \\ & - 2.612 \times 10^{-5} \cdot \text{Accounts}\end{aligned}$$

Refined Model Results

To improve model interpretability, non-significant predictors were removed. The refined model retained *Experience*, *Earnings*, *HomeOwner*, and *BadPastRecords* as predictors. This model explained 98.02% of the variance ($R^2 = 0.9802$, Adjusted $R^2 = 0.9799$), with all predictors being statistically significant ($p < 0.001$).

To fit the refined model, the following equation was used for predicting the observed default rate (ObsRate):

$$\begin{aligned}\text{ObsRate} = 0.1259 - 0.002832 \cdot \text{Experience} - 2.770 \times 10^{-7} \cdot \text{Earnings} \\ - 0.006612 \cdot \text{HomeOwnerYes} + 0.01742 \cdot \text{BadPastRecords}\end{aligned}$$

Key findings from the refined model include:

- **Experience:** Each additional year of experience reduces the observed default rate by approximately 0.0028.
- **Earnings:** Higher earnings are associated with a slight reduction in the default rate.
- **Home Ownership:** Homeowners tend to have a default rate approximately 0.0066 lower than non-homeowners.
- **Bad Past Records:** Each additional bad record increases the default rate by 0.0174.

The variables `Accounts` and `Defaults` were excluded from the refined model despite being statistically significant. This decision was made because their contribution to improving the model's explanatory power was minimal. By removing these variables, the model became simpler and easier to interpret without significantly impacting its overall performance.

Model Comparison

The results, shown in Table 2, indicate that at $\alpha = 0.05$, the full model is statistically significant compared to the refined model . However, this does not necessarily mean that the full model is better in terms of practical performance—it simply means that the additional predictors in the full model contribute significantly to explaining variance in the response variable.

Table 2: ANOVA Comparison Between Full and Refined Models

Model	Residual DF	RSS	Df	Sum of Sq	p-value
Full Model	293	0.0091643	-	-	-
Refined Model	299	0.0097131	-6	-0.00054879	0.0087 **

The refined model demonstrates strong predictive power and provides actionable insights into the factors influencing default risk.

Model Assumptions Check

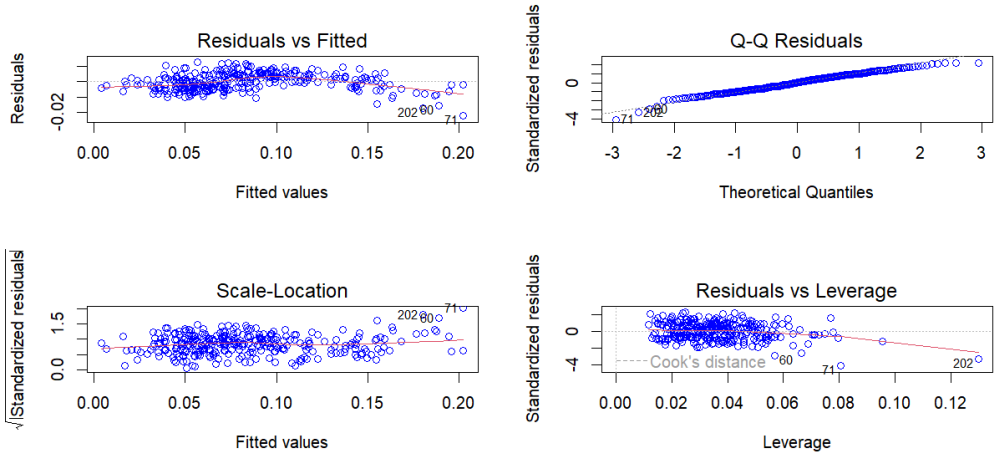

Figure 2: Plots for Model Assumption Checking

The diagnostic plots provide key insights into how well the assumptions of the linear model hold. The **Residuals vs Fitted** plot shows a slight curve, which suggests that the relationship between the predictors and the response might not be perfectly linear.

The **Q-Q Plot** indicates that most residuals follow a normal distribution, but some deviations at the extremes suggest a few outliers or a slight departure from normality, especially at the tails.

In the **Scale-Location Plot**, the spread of residuals increases with the fitted values, suggesting that the variance is not constant (heteroscedasticity). This pattern indicates that the assumption of homoscedasticity might be violated .

Finally, the **Residuals vs Leverage Plot** highlights a few points with high leverage. These points have the potential to strongly influence the model.

Model Improvement Discussion

- Applying transformations such as logarithmic or square root adjustments to the response variable or predictors could help linearize relationships and stabilize variance, mitigating non-linearity and heteroscedasticity.
- Points with high leverage have the potential to disproportionately influence the model. A deeper examination of these points is necessary to assess their validity and impact on model performance.
- Incorporating regularization methods like Ridge or Lasso regression could address multicollinearity and help improve the model's predictive performance, especially in the presence of correlated predictors.
- If transformations and regularization do not adequately resolve the non-linearity, considering non-linear models like decision trees or random forests could capture the complex relationships more effectively.

End.