AFRICAN INSTITUTE FOR MATHEMATICAL SCIENCES

(AIMS RWANDA, KIGALI)

Regression Model for Predicting Observed Default Rate

Name: Jean Aime IRAGUHA
Course: Statistical Regression

Assignment Number: 2
Date: November 23, 2024

1 Introduction

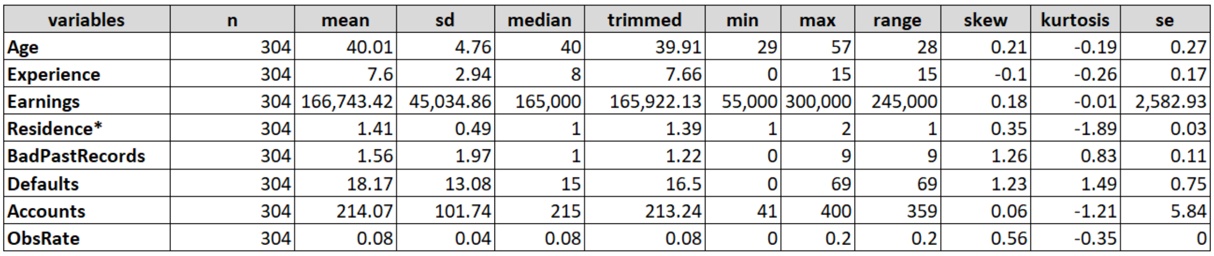
Understanding factors that influence default rates is critical for effective financial risk management. This study develops a predictive model for default rates (`ObsRate`) using demographic, financial, and behavioral factors. By applying a generalized linear model (GLM) with a quasibinomial family, the study identifies key predictors and evaluates their impact. The findings aim to provide actionable insights for decision-making in credit risk assessment.

2 Descriptive Analysis

In this part descriptive analysis was conducted to understand the characteristics of the dataset and identify patterns that could influence the default rate.

2.1 Summary Statistics

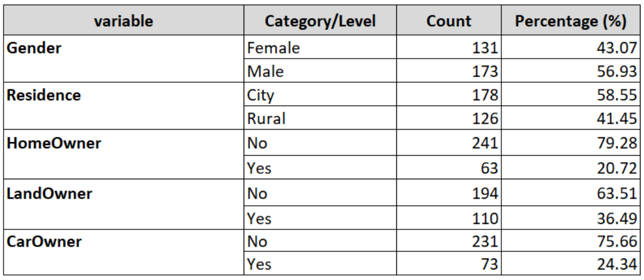
Table 2 presents summary statistics for key numerical variables.



variables	n	mean	sd	median	trimmed	min	max	range	skew	kurtosis	se
Age	304	40.01	4.76	40	39.91	29	57	28	0.21	-0.19	0.27
Experience	304	7.6	2.94	8	7.66	0	15	15	-0.1	-0.26	0.17
Earnings	304	166,743.42	45,034.86	165,000	165,922.13	55,000	300,000	245,000	0.18	-0.01	2,582.93
Residence*	304	1.41	0.49	1	1.39	1	2	1	0.35	-1.89	0.03
BadPastRecords	304	1.56	1.97	1	1.22	0	9	9	1.26	0.83	0.11
Defaults	304	18.17	13.08	15	16.5	0	69	69	1.23	1.49	0.75
Accounts	304	214.07	101.74	215	213.24	41	400	359	0.06	-1.21	5.84
ObsRate	304	0.08	0.04	0.08	0.08	0	0.2	0.2	0.56	-0.35	0

Figure 1: summary statistics of Numeric variable

The summary statistics of categorical variable is of the form :



variable	Category/Level	Count	Percentage (%)
Gender	Female	131	43.07
	Male	173	56.93
Residence	City	178	58.55
	Rural	126	41.45
HomeOwner	No	241	79.28
	Yes	63	20.72
LandOwner	No	194	63.51
	Yes	110	36.49
CarOwner	No	231	75.66
	Yes	73	24.34

Figure 2: summary statistics for categorical variable

From the table 1 and 2 The numerical variables show that Age has a mean of 40.01, Experience averages 7.60, and Earnings has a mean of 166,743.42, while BadPastRecords has a mean of 1.56. For the categorical variables, the dataset includes 131 females and 173 males, with 241 homeowners and 178 city residents.

2.2 Correlation Analysis

A correlation analysis was performed to assess the relationships between numerical variables. high correlations were observed between defaults and past records , Earnings and Experiences and default with Accounts. The summary correlation graph is presented below

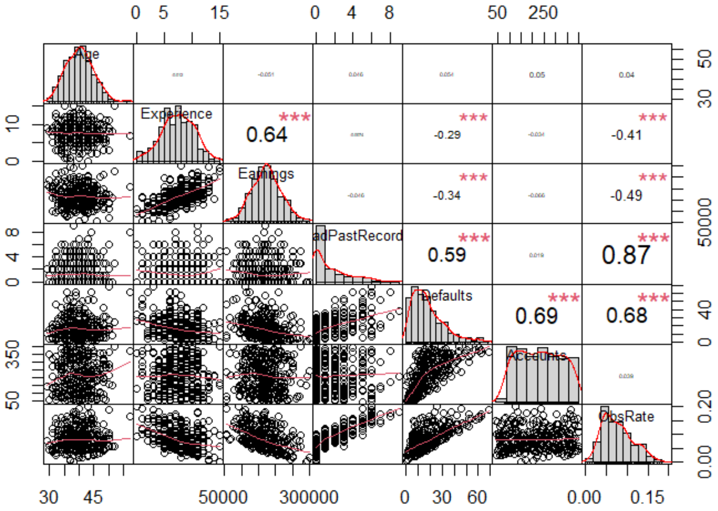


Figure 3: Correrration Plot

From this graph 3 we can also access the distribution of each variable . we can se that the age, Experiences and Earnings are normary distributed since they are symetric obout their mean.

3 Regression Model Development

Initial Model

We have decided to use Generalized linear model because the response variable is proportion and it is bounded between 0 and 1 and we set . The initial model included all available predictors:

$$\texttt{ObsRate} \sim \texttt{Age} + \texttt{Gender} + \texttt{Experience} + \texttt{Earnings} + \texttt{Residence} + \texttt{HomeOwner} + \texttt{LandOwner} +$$
$$\texttt{BadPastRecords} + \texttt{CarOwner} + \texttt{Accounts}.$$

Results shows that :

- Residual Deviance: 0.680 (Null Deviance: 6.433).
- Non-significant predictors ($p > 0.05$): `Age`, `Gender`, `Residence`, `CarOwner`, `Accounts`.
- Significant predictors ($p \leq 0.05$): `Experience`, `Earnings`, `HomeOwner`, `BadPastRecords`.

While the model fit the data , its retaining non-significant predictors decrease the model accuracy. That why we decide to remove the insignificant parameters.

Modified Model

The modified model retained only significant predictors:

$$\texttt{ObsRate} \sim \texttt{Experience} + \texttt{Earnings} + \texttt{HomeOwner} + \texttt{BadPastRecords}.$$

The generalized linear model (GLM) used for predicting the observed default rate (`ObsRate`) follows a logit link function. The logit transformation is defined as:

$$\text{logit}(\text{ObsRate}) = \log\left(\frac{\text{ObsRate}}{1-\text{ObsRate}}\right) = -1.86 - 0.0339 \cdot X_1 - 0.00000389 \cdot X_2 - 0.079 \cdot X_3 + 0.193 \cdot X_4.$$

Where: X_1 is `Experience`, X_2 is `Earnings`, X_3 is `HomeOwnerYes`, X_4 is `BadPastRecords`.

To convert the logit back to a probability, we use the inverse logit transformation:

$$\text{Obs}\hat{\text{R}}\text{ate} = \frac{e^{\text{logit}(\text{ObsRate})}}{1 + e^{\text{logit}(\text{ObsRate})}}.$$

Substituting the logit equation:

$$\text{Obs}\hat{\text{R}}\text{ate} = \frac{\exp(-1.860 - 0.03398 \cdot X_1 - 0.000003898 \cdot X_2 - 0.07936 \cdot X_3 + 0.1932 \cdot X_4)}{1 + \exp(-1.860 - 0.03398 \cdot X_1 - 0.000003898 \cdot X_2 - 0.07936 \cdot X_3 + 0.1932 \cdot X_4)}.$$

Interpretation of Coefficients

- **Intercept:**($\beta_0 = -1.860$) When all predictor variables are zero, the baseline log-odds of default is -1.860.More years of experience significantly reduce the likelihood of default.
- **Experience ($\beta_1 = -0.03398$):** For each additional year of experience, the log-odds of default decrease by 0.03398, indicating a reduced probability of default with more experience.
- **Earnings ($\beta_2 = -0.000003898$):** For each additional CFA of earnings, the log-odds of default decrease by 0.000003898, indicating that higher earnings reduce the likelihood of default.
- HomeOwnerYes ($\beta_3 = -0.07936$): If an individual is a homeowner, the log-odds of default decrease by 0.07936, suggesting homeownership is associated with a lower probability of default.
- **BadPastRecords ($\beta_4 = 0.1932$):** For each additional bad past record, the log-odds of default increase by 0.1932, highlighting the strong impact of past financial issues on the likelihood of default.

3.1 Model Selection Procedure

We observed a high correlation between `Earnings` and `Experience`, and decided to remove one of them in order to reduce the number of parameters in the model. Based on the residual deviance and Likelihood Ratio Test (LRT), the initial model with four predictors demonstrated a significantly better fit than the modified model with three predictors. The deviance of the initial model was 0.6881372, while the modified model had a higher deviance of 1.077507. The LRT yielded a highly significant p-value ($< 2.2 \times 10^{-16}$), confirming that the initial model provided a better fit.

4 Model Diagnostics

4.1 Assumption Checks

In a generalized linear model (GLM) with a quasibinomial family and a logit link, the key assumptions that need to be checked are Linearity of the Logit, Independence of Observations,No Multicollinearity ,Homoscedasticity and No Over dispersion. Perfoming Durbin-Watson test for autocorrelation we ontain the 2.0477 which is closer to 2 and impilies that there is no autocorrelation and Multicolineatity assumption holds . The other assumption is can be indicated by the plots in figure 4.

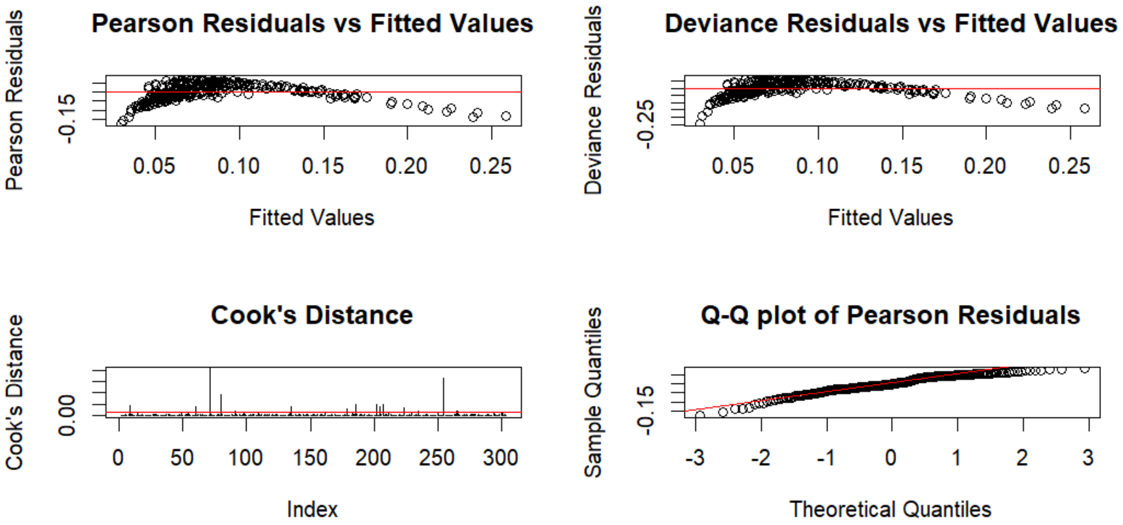


Figure 4: Assumption of model check graph

In the figure 4, the Plot the Pearson **Deviance residuals vs fitted values** indicate that homoskedaciticity assumption holds. Also the grath of **residuals vs fitted value** prove the assumption of lineality.futher more **Q-Q plot of peason resdual** prove the normality assumption.

4.2 Potential Model Improvements

While the model meets all necessary assumptions, there are opportunities for refinement to enhance its predictive power and robustness. First, addressing potential multicollinearity, such as between Experience and Earnings, by combining them into a composite variable or using regularization like ’Lasso or Ridge regression’ can improve model stability. Additionally, including interaction terms like **HomeOwner * BadPastRecords** could capture combined effects that might be overlooked. Incorporating splines or polynomial terms for predictors like Experience can address any subtle non-linear relationships.

5 Conclusion

This study analyzed factors influencing default rates and developed a predictive model using ‘Experience‘, ‘Earnings‘, ‘HomeOwner‘, and ‘BadPastRecords‘. The model met all statistical assumptions, and key results showed that higher experience and earnings lower default risk, while bad past records increase it. The model is reliable and provides insights for financial risk management.