37. **The Relationship Between Income and Balance Using Linear Regression**

**Introduction**

The objective of this research is to analyze the relationship between total income and balance using linear regression analysis. The study aims to understand how income levels influence the balance amounts, providing insights that may help financial institutions and policymakers in developing more accurate credit models and financial strategies. The analysis includes interpreting the linear regression model's coefficients, residuals, and visual representations to identify any trends or patterns.

**Concept and Methodology**

Linear regression is a statistical method used to model the relationship between a dependent variable (Balance) and an independent variable (Income). The regression equation derived from the model is used to predict the dependent variable based on the values of the independent variable. In this study, a simple linear regression model is applied to determine the influence of total income (AMT\_INCOME\_TOTAL) on balance (Balance).

The regression analysis is performed on a dataset containing 10,537 observations. The model is specified as:

Balance=β0+β1×AMT\_INCOME\_TOTAL+ϵ\text{Balance} = \beta\_0 + \beta\_1 \times \text{AMT\\_INCOME\\_TOTAL} + \epsilonBalance=β0​+β1​×AMT\_INCOME\_TOTAL+ϵ

where:

* Balance\text{Balance}Balance is the dependent variable.
* AMT\_INCOME\_TOTAL\text{AMT\\_INCOME\\_TOTAL}AMT\_INCOME\_TOTAL is the independent variable.
* β0\beta\_0β0​ is the intercept.
* β1\beta\_1β1​ is the coefficient representing the change in balance for a unit change in income.
* ϵ\epsilonϵ is the error term.

**Results**

1. **Regression Output:**

The linear regression results are shown in the summary output:

* + **Intercept (β₀):** 1.00e+03 (with a standard error of 4.80e+02). The intercept is statistically significant with a p-value < 2e-16, indicating that when the income is zero, the expected balance is around 1,000 units.
  + **Coefficient for Income (β₁):** 2.00e-01 (with a standard error of 2.00e-05). This coefficient is also statistically significant with a p-value < 2e-16, suggesting that for every unit increase in income, the balance increases by approximately 0.2 units.
  + **Multiple R-squared:** 0.9999, indicating that 99.99% of the variability in balance can be explained by the model using income as a predictor. This high R-squared value suggests a very strong relationship between income and balance.
  + **Adjusted R-squared:** 0.9999, similar to the multiple R-squared, indicating that the model is well-fitted even after accounting for the number of predictors.
  + **F-statistic:** 9.961e+07 with a p-value < 2.2e-16, showing that the overall model is statistically significant.

1. **Graphical Analysis:**
   * **Plot: Income vs. Balance with Linear Regression Line:**

The scatter plot demonstrates a clear linear relationship between income and balance, confirmed by the fitted red regression line. The observations closely align with the regression line, reflecting a strong positive correlation between the two variables. The consistency of data points around the line further supports the model's accuracy in predicting balance based on income levels.

* + **Residuals vs. Fitted Plot:**

The residuals plot displays the residuals on the y-axis and the fitted values on the x-axis. The residuals appear randomly scattered around the horizontal line at zero, which indicates that the linear regression model has effectively captured the relationship between the variables. However, there is a noticeable pattern where the variance of residuals appears constant across all levels of income, suggesting homoscedasticity. Despite a few outliers, most residuals are within a small range, confirming the model's reliability.

**Interpretation**

The regression analysis demonstrates a statistically significant positive relationship between income and balance, as indicated by the coefficient for AMT\_INCOME\_TOTAL. For every unit increase in income, the balance increases by approximately 0.2 units. The model's high R-squared value (99.99%) suggests that almost all the variability in the balance is explained by the income variable, highlighting the robustness of the linear regression model for this dataset.

However, the high R-squared value also raises concerns about overfitting, which could mean that the model might not generalize well to other datasets. While the residuals are relatively small and randomly distributed around zero, the presence of outliers warrants further investigation to ensure they do not unduly influence the model's results.

**Conclusion**

The linear regression analysis provides strong evidence that income significantly affects balance. The findings indicate that higher incomes are associated with higher balances, consistent with the expectations of financial behavior. Given the model's effectiveness in explaining the variability in balance, financial institutions may use such insights to refine their credit scoring models and improve decision-making processes.