Multiple Regression Analysis of Bank Marketing Data

1. Research Question

This study examines how customer demographic and financial characteristics influence the likelihood of a client subscribing to a term deposit. Specifically, the analysis investigates whether age, account balance, duration of last contact, number of campaign contacts, and previous contacts significantly predict subscription outcomes.

2. Dataset Description

The dataset used in this analysis is the **Bank Marketing dataset** obtained from the **UC Irvine Machine Learning Repository**. The data comes from direct marketing campaigns of a Portuguese banking institution. It contains information on client demographics, campaign interactions, and whether the client subscribed to a term deposit. After cleaning, the dataset includes relevant numeric variables suitable for regression analysis.

3. Data Inspection and Cleaning

The dataset was loaded into R and inspected using structure and summary functions. Missing values were checked and handled appropriately. Non-numeric and irrelevant variables were excluded. The dependent variable (subscription outcome) was converted into numeric form to allow regression modeling.

4. Summary Statistics

Descriptive statistics including mean, minimum, maximum, and variability were calculated for all variables used in the regression. Account balance and duration of last contact showed substantial variation, while age and campaign-related variables were more evenly distributed. These statistics provide context for interpreting the regression results.

5. Model Specification

The multiple regression model is defined as:

Subscription = $\beta_0 + \beta_1(\text{Age}) + \beta_2(\text{Balance}) + \beta_3(\text{Duration}) + \beta_4(\text{Campaign}) + \beta_5(\text{Previous}) + \varepsilon$

- **Dependent Variable (Y):** Subscription outcome
- **Independent Variables (X):** Age, balance, duration, campaign contacts, previous contacts

6. Regression Results and Coefficient Interpretation

The regression results indicate that **duration of the last contact** has a strong positive relationship with subscription outcomes. Holding other variables constant, longer contact duration increases the likelihood of subscription. Account balance shows a modest positive effect, suggesting clients with higher balances are slightly more likely to subscribe. Age and campaign-related variables exhibit weaker relationships.

7. Statistical Significance

Based on p-values, **duration** and **balance** are statistically significant predictors at the 5% significance level. Other predictors are not statistically significant, indicating they do not reliably explain variation in subscription outcomes in this model.

8. Confidence Intervals

Ninety-five percent confidence intervals were calculated for all coefficients. The confidence intervals for duration and balance do not include zero, reinforcing their statistical significance. Confidence intervals for other variables include zero, confirming their lack of significance.

9. Coefficient of Determination (R^2)

The model's R^2 indicates that a meaningful portion of the variation in subscription outcomes is explained by the included predictors. This suggests moderate explanatory power, though additional variables may improve model performance.

10. Residual Diagnostics

Residual plots were examined to assess regression assumptions. The residuals appear approximately normally distributed, show constant variance, and do not display systematic patterns. These findings suggest that the assumptions of linear regression are reasonably satisfied.

11. Prediction

A prediction was generated for a hypothetical client with average age and balance but a longer-than-average contact duration. The model predicts a higher likelihood of subscription for this client, highlighting the importance of interaction duration in marketing effectiveness.

12. Conclusion and Limitations

This analysis demonstrates that **marketing contact duration** is the most influential predictor of term deposit subscription. Account balance also plays a role, while other demographic and campaign variables are less impactful. Limitations of the study include potential nonlinearity, omitted behavioral factors, and reliance on a linear probability approach. Future research could incorporate logistic regression and additional client behavior variables.