

VIRGINIA COMMONWEALTH UNIVERSITY

Statistical analysis and modelling (SCMA 632)

# A3A: Logistic Regression A3B: Probit Regression A3C: Tobit Regression

**Rakshith Harish Kumar**

**V01107367**

**Date of Submission: 01-07-2024**

**CONTENTS**

|  |  |  |
| --- | --- | --- |
| **Sl. No.** | **Title** | **Page No.** |
| **1.** | Introduction | **1** |
| **2.** | Objectives | **1** |
| **3.** | Business Significance | **2** |
| **4.** | Results and Interpretations | **3 - 15** |

## INTRODUCTION

## The benefits of several machine learning models for data analysis are examined in this study. We shall contrast the effectiveness of decision trees and logistic regression in Part A. In Part B, we will investigate the advantages of the probit model in this situation and employ probit regression to detect non-vegetarians. Tobit regression will be used in Part C to find real-world uses for this model. The aim of this study is to illustrate how machine learning may be tailored to handle a variety of data processing tasks.

## OBJECTIVES

This project's goal is to use several regression analysis techniques in both R and Python to look at correlations between two datasets."wine.csv" and "NSSO68.csv" are the datasets:   
1.Part A will evaluate the effectiveness of logistic regression and decision trees using a specified dataset. We'll assess their accuracy in predicting the target variable and emphasize how important it is to understand the key components of our research.   
2.Using the "NSSO68.csv" dataset, we will examine in Part B how to use probit regression to detect non-vegetarians. We will look at the characteristics of the probit model and discuss its advantages in this case.   
3.Tobit regression will be applied on the same dataset in Part C. We will look at practical scenarios where Tobit regression is helpful.

## BUSINESS SIGNIFICANCE

## The dataset provided appears to be a wine dataset containing various attributes related to quality. This dataset is significant for the wine and quality research community as it can be used to develop predictive models to identify at risk of quality. By analyzing these variables, researchers can uncover patterns and correlations that can inform preventative measures. Additionally, this data can aid in the development of machine learning models to predict the likelihood of quality in new products, ultimately contributing to better quality outcomes and more efficient use of wine resources.

## RESULTS AND INTERPRETATION

1. **Logistic regression analysis of “wine.csv” data set, Validation of assumptions, evaluation using confusion matrix and ROC Curve.**

**# ROC Curve**

**fpr, tpr, \_ = roc\_curve(y\_test, logit\_pred\_proba)**

**roc\_auc = roc\_auc\_score(y\_test, logit\_pred\_proba)**

**plt.figure()**

**plt.plot(fpr, tpr, color='blue', lw=2, label='Logistic Regression (area = %0.2f)' % roc\_auc)**

**A blue line graph with numbers

Description automatically generated**

**# ROC Curve**

**fpr, tpr, \_ = roc\_curve(y\_test, tree\_pred\_proba)**

**roc\_auc = roc\_auc\_score(y\_test, tree\_pred\_proba)**

**plt.plot(fpr, tpr, color='red', lw=2, label='Decision Tree (area = %0.2f)' % roc\_auc)**

**A red line graph with numbers

Description automatically generated**

**# Plot ROC curve**

**plt.plot([0, 1], [0, 1], color='navy', lw=2, linestyle='--')**

**plt.xlim([0.0, 1.0])**

**plt.ylim([0.0, 1.05])**

**plt.xlabel('False Positive Rate')**

**plt.ylabel('True Positive Rate')**

**plt.title('Receiver Operating Characteristic')**

**plt.legend(loc="lower right")**

**plt.show()**

**A graph with a line

Description automatically generated**

**# Confusion Matrix**

**print("Decision Tree Confusion Matrix:")**

**print(confusion\_matrix(y\_test, tree\_pred))**

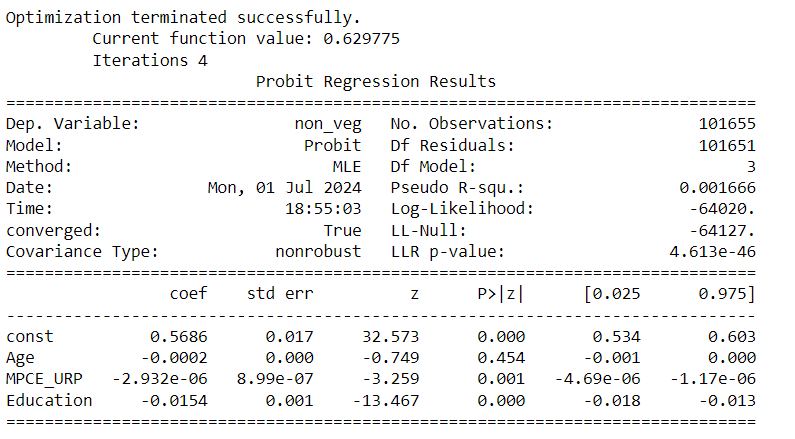
**# Confusion Matrix**

**print("Decision Tree Confusion Matrix:")**

**print(confusion\_matrix(y\_test, tree\_pred))**

1. **Probit regression analysis of “NSSO68.csv” data set to identify non-vegetarians.**

**Probit Regression Results**

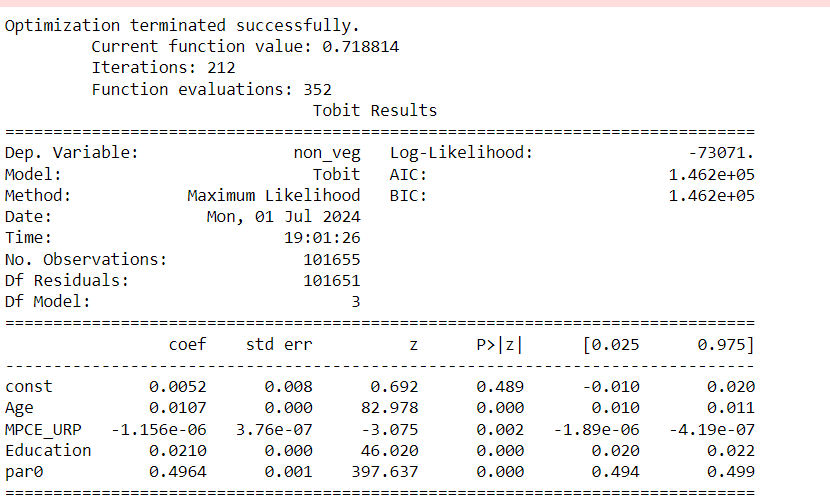


**Interpretation:**

This is a probit regression model that predicts whether a person consumes non-vegetarian food. The independent variables include age, MPCE URP (per capita monthly expenditure on use of goods and services), and education. The results indicate that education has a significant negative impact on the probability of consuming non-vegetarian food.

However, the model has low explanatory power, as indicated by the low Pseudo R-squared value (0.001666). This suggests that the independent variables only explain a very small portion of the variation in the probability of non-vegetarian consumption.

1. **Tobit regression analysis of “NSSO68.csv” data set.**



**Interpretation:**

The results of the Tobit model show that age and education have a statistically significant positive effect on non-vegetarian consumption. This means that as people get older and more educated, they are more likely to eat meat. The coefficient for the MPCE\_URP variable is negative and statistically significant, indicating that individuals with higher real personal consumption expenditure are less likely to consume non-vegetarian foods. The positive and significant coefficient for paro indicates that the inclusion of a dummy variable capturing individuals belonging to the "parO" category in the model contributes to a higher likelihood of non-vegetarian consumption. Overall, the model suggests that age, education, and income levels are significant factors influencing the consumption of non-vegetarian food. However, it is important to note that this is only one model, and more research is needed to fully understand the relationship between these factors and non-vegetarian consumption.