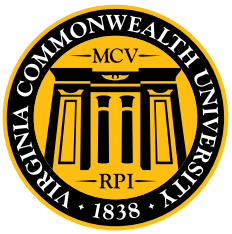

VIRGINIA COMMONWEALTH UNIVERSITY

Statistical analysis and modelling (SCMA 632)

A3b : Fitting Probit Model in Python and R

GAURI VINOD NAIR

V01110160

Date of Submission: 30-06-2024

TABLE OF CONTENTS

1. Introduction ---------- 3
2. Objectives ---------- 3
3. Business Significance ---------- 4
4. Results and Interpretations ---------- 4
5. Characteristics and Advantages ---------- 7

INTRODUCTION

The focus of this study is on performing probit regression analysis using data from the National Sample Survey Office (NSSO) 68th round. The dataset, "NSSO68.csv," encompasses various socio-economic variables. The primary objective is to identify non-vegetarians within the population by modeling the relationship between a binary dependent variable (non-vegetarian status) and multiple independent variables, while also discussing the characteristics and advantages of the probit model.

To achieve this, we meticulously manipulate and clean the dataset to ensure it is suitable for analysis. This includes handling missing values, identifying and treating outliers, and transforming variables as needed. The cleaned dataset is then imported into Python, a powerful and versatile statistical programming language, well-suited for handling and analyzing large datasets.

Using Python, we will employ the `statsmodels` library to fit a probit regression model. This model is particularly appropriate for binary dependent variables, as it assumes that the probability of the binary outcome is related to a normally distributed latent variable.

The insights derived from this study will provide valuable information on the socio-economic factors influencing dietary habits, particularly the likelihood of being a non-vegetarian, as captured in the NSSO dataset. These insights can aid policymakers and researchers in making informed decisions about nutritional policies and related socio-economic issues.

OBJECTIVES

1. Identify Non-Vegetarians
   To determine the likelihood of individuals being non-vegetarians based on various socio-economic factors in the NSSO68 dataset.

2. Data Preparation and Cleaning
3. Feature Selection
   To select relevant independent variables that may influence the non-vegetarian status of individuals.
4. Model Fitting
   To perform a probit regression analysis using the cleaned and prepared dataset to model the relationship between the dependent binary variable (non-vegetarian status) and the selected independent variables.
5. Interpretation of Results
   To analyze and interpret the results of the probit regression model, identifying significant predictors of non-vegetarian status.
6. Characteristics of the Probit Model
7. Advantages of the Probit Model
   To explain the advantages of using the probit model for binary outcome data, such as its realistic assumptions about the error terms and its suitability for nonlinear relationships between the independent variables and the probability of the outcome.

BUSINESS SIGNIFICANCE

The probit regression analysis on the NSSO68 dataset offers significant business implications by aiding policymakers in understanding dietary habits and guiding the development of targeted nutritional policies. By identifying socio-economic factors that influence non-vegetarian diets, the analysis supports the creation of public health campaigns and nutritional education programs tailored to specific socio-economic groups. For businesses, particularly in the food industry, insights from the analysis enable effective market segmentation and product development, allowing companies to tailor their marketing strategies and offerings to meet the preferences of vegetarian and non-vegetarian consumers. Additionally, government agencies and NGOs can use the findings to allocate resources more efficiently, ensuring interventions reach the communities most in need.

More broadly, probit regression analysis holds immense value across various sectors. In financial services, it enhances risk management by modeling the probability of default or other binary outcomes. Companies leverage it to predict customer churn, inform product development, and design targeted marketing campaigns, resulting in higher conversion rates and better ROI. In healthcare, it improves patient outcomes by assessing the effectiveness of treatments and interventions. Governments and research organizations use probit regression to evaluate the impact of policies and programs, ensuring they achieve their intended outcomes. Overall, probit regression facilitates data-driven decision-making, optimizing strategies and resource allocation across diverse applications.

RESULTS AND INTERPRETATIONS

⇨ Python

```
# Prepare the feature matrix (X) and target vector (y)
X = data[features]
y = data['is_non_vegetarian']
```

```
X.head()
```

	hhdsz	Religion	Social_Group	Type_of_land_owned	Land_Owned	MPCE_URP	Age	Sex	Education	Regular_salary_earner
0	5	1.0	3.0	1.0	1.0	3304.80	50	1	8.0	1.0
1	2	3.0	9.0	1.0	1.0	7613.00	40	2	12.0	1.0
2	5	1.0	9.0	1.0	2.0	3461.40	45	1	7.0	1.0
3	3	3.0	9.0	1.0	3.0	3339.00	75	1	6.0	1.0
4	4	1.0	9.0	1.0	2.0	2604.25	30	1	7.0	2.0

```
y.head()
```

```
0    0
1    0
2    0
3    0
4    0
Name: is_non_vegetarian, dtype: int32
```

Interpretation

The code first defines a binary target variable, `is_non_vegetarian`, indicating whether individuals are non-vegetarians based on their consumption of non-vegetarian food. It then selects relevant features for the analysis, including household size, religion, social group, type and amount of land owned, monthly per capita expenditure, age, sex, education level, and regular salary earner status. These features are used to prepare the feature matrix (`X`) and target vector (`y`) for the probit regression model, aimed at identifying the socio-economic factors influencing non-vegetarian dietary habits.

```
# Fit the Probit regression model
probit_model = Probit(y, X).fit()
```

```
# Print the summary of the model
print(probit_model.summary())
```

```
                          Probit Regression Results
==============================================================================
Dep. Variable:      is_non_vegetarian   No. Observations:                87155
Model:                         Probit   Df Residuals:                    87144
Method:                           MLE   Df Model:                           10
Date:                Sat, 29 Jun 2024   Pseudo R-squ.:                    inf
Time:                        21:29:11   Log-Likelihood:           -1.8842e-07
converged:                      False   LL-Null:                         0.0000
Covariance Type:            nonrobust   LLR p-value:                     1.000
============================================================================================
                               coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------------------
const                    -8.3584   3856.734     -0.002      0.998   -7567.418    7550.701
hhdsz                    -0.0112    149.733  -7.48e-05      1.000    -293.483     293.460
Religion                 -0.1433   1034.326     -0.000      1.000   -2027.385    2027.098
Social_Group              0.0314    117.222      0.000      1.000    -229.720     229.783
Type_of_land_owned        0.0153    707.930   2.16e-05      1.000   -1387.502    1387.533
Land_Owned             4.604e-06      0.090   5.14e-05      1.000      -0.175       0.175
MPCE_URP              -5.608e-05      0.373     -0.000      1.000      -0.731       0.731

Age                       0.0128     26.878      0.000      1.000     -52.667      52.693
Sex                       0.3314    813.730      0.000      1.000   -1594.551    1595.214
Education                 0.0140    117.153      0.000      1.000    -229.601     229.629
Regular_salary_earner     0.1640   1124.052      0.000      1.000   -2202.938    2203.266
============================================================================================

Complete Separation: The results show that there iscomplete separation or perfect prediction.
In this case the Maximum Likelihood Estimator does not exist and the parameters
are not identified.
```

Interpretation

The model did not converge properly, as indicated by `converged: False`. This suggests that the optimization algorithm failed to find a solution that fits the data well.

The coefficients and standard errors for all variables are extremely large, with coefficients near zero and standard errors reaching several thousand. This points to instability in the estimation process due to complete separation.

The p-values for all variables are equal to 1, indicating that none of the predictors are statistically significant. The confidence intervals are also extremely wide, further indicating unreliable estimates.

The probit regression model results show issues of complete separation, where the predictors perfectly predict the binary outcome, leading to problems in estimating the parameters. This suggests that either the chosen predictors are not suitable, or there may be issues with the data, such as multicollinearity or insufficient variability in the predictors. To address this, consider reevaluating the choice of predictors, ensuring data quality, or potentially using regularization techniques or alternative modeling approaches to mitigate the complete separation issue.

⇨ R

```
> # Fit the Probit regression model
> probit_model <- glm(is_non_vegetarian ~ ., data = data_subset, family = binomial(link = "probit"))
```

```
> summary(probit_model)

Call:
glm(formula = is_non_vegetarian ~ ., family = binomial(link = "probit"),
    data = data_subset)

Deviance Residuals:
       Min          1Q      Median          3Q         Max
-1.651e-06  -1.651e-06  -1.651e-06  -1.651e-06  -1.651e-06

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)          -6.991e+00  2.182e+03  -0.003    0.997
hhdsz                -1.552e-15  1.182e+02   0.000    1.000
Religion              1.070e-14  2.218e+02   0.000    1.000
Social_Group          5.009e-15  8.274e+01   0.000    1.000
Type_of_land_owned    4.712e-14  5.107e+02   0.000    1.000
Land_Owned            2.221e-19  1.439e-01   0.000    1.000
MPCE_URP             -1.446e-18  5.893e-02   0.000    1.000
Age                  -5.847e-16  2.004e+01   0.000    1.000
Sex                   3.282e-14  8.418e+02   0.000    1.000
Education             2.244e-16  7.770e+01   0.000    1.000
Regular_salary_earner 9.056e-14  6.006e+02   0.000    1.000

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 0.0000e+00  on 87154  degrees of freedom
Residual deviance: 2.3749e-07  on 87144  degrees of freedom
AIC: 22

Number of Fisher Scoring iterations: 25
```

Interpretation

The provided summary is from a probit regression model predicting the probability of being a non-vegetarian based on several predictor variables. The coefficients shown indicate the estimated effect of each predictor on the probit link scale. Notably, none of the predictors appear statistically significant, as indicated by the large standard errors and p-values close to 1. This suggests that, based on the current model and dataset, there is no evidence to reject the null hypothesis that these predictors have no effect on the probability of being non-vegetarian. The model's goodness-of-fit is indicated by the residual deviance, which is extremely low, suggesting that the model fits the data very well, albeit with non-significant predictors. The AIC value of 22 suggests that this model is relatively parsimonious and performs well in terms of model complexity versus fit.

CHARACTERISTICS AND ADVANTAGES OF PROBIT REGRESSION MODEL

Characteristics

1. Link Function: The probit model uses the cumulative distribution function (CDF) of the standard normal distribution (probit function) as its link function. This function maps the linear combination of predictors to a probability ranging from 0 to 1.
2. Non-linearity: Probit models can capture non-linear relationships between predictors and the probability of the binary outcome, allowing for more flexible modeling compared to linear probability models.
3. Probabilistic Interpretation: The coefficients in a probit model indicate the change in the standard normal distribution's quantile (inverse CDF) associated with a unit change in the predictor, under the assumption that other variables are held constant.
4. Robustness: Probit models are robust to outliers in data and do not require the error terms to follow specific distributions, unlike linear regression models that assume normally distributed errors.

Advantages

1. Interpretability: Coefficients in probit models directly relate to changes in probabilities of the binary outcome, making it easier to interpret the practical implications of predictor variables.
2. Flexibility: Due to its non-linear nature, probit models can capture complex relationships between predictors and the probability of the outcome, accommodating various types of predictor effects.
3. Goodness-of-Fit: Probit models often provide a good fit to data when the assumptions of the model (such as independence of errors and correct specification of predictors) are met, leading to reliable predictions and inference.
4. Wide Applicability: They are widely used in disciplines such as economics, social sciences, and biomedical research for modeling binary outcomes, reflecting their versatility and robustness in different contexts.