****

**VIRGINIA COMMONWEALTH UNIVERSITY**

**Statistical analysis and modeling (SCMA 632)**

**A3- Limited dependent variable Models statistics**

**SAMPREETH S SHETTY**

**V01107634**

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

**CONTENTS**

|  |  |  |
| --- | --- | --- |
| **Sl. No.** | **Title** | **Page No.** |
| **1.** | Introduction | **1** |
| **2.** | Results | **2,3,7,8** |
| **3.** | Interpretations | **5,6,11** |
| **4.** | Illustrations | **3,4,5** |
| **5.** | Codes | **-** |
| **6.** | References | **-** |

**Introduction**

Limited dependent variable models, such as probit and logit models, are essential for analyzing outcomes that are categorical or restricted in range. These models are commonly used when the dependent variable is binary, ordinal, or censored.

**Objectives**

Limited dependent variable models, like probit and logit models, aim to:

* Model binary outcomes accurately by estimating probabilities based on predictor variables.
* Handle the non-linear relationship between predictors and probabilities.
* Ensure predictions are bounded between 0 and 1, avoiding nonsensical results.
* Provide interpretable coefficients that explain the impact of predictors on the outcome probability.

**Probit Model**

The probit model is used for binary outcomes, where the dependent variable can take only two values, typically 0 and 1. It estimates the probability that an observation will fall into one of the categories, assuming the error terms follow a standard normal distribution. This model is helpful in contexts where normality is a reasonable assumption.

**Logit Model**

The logit model is similar to the probit model but assumes that the error terms follow a logistic distribution with slightly heavier tails than the normal distribution. It is popular in economics and social sciences due to its interpretability and ease of use.

**Business Applications**

Both models are widely used in various fields. In marketing, they model consumer choices; in finance, they predict loan defaults; and in medicine, they assess the presence or absence of a disease. The choice between a probit and a logit model depends on theoretical considerations and data characteristics.

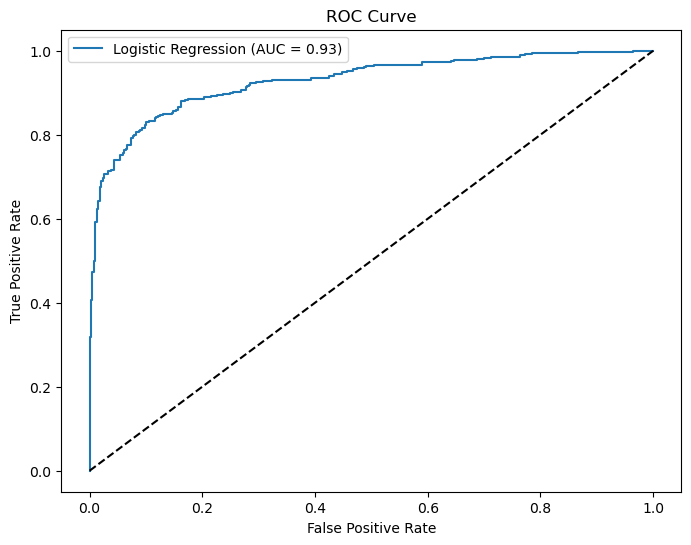
**Part A**

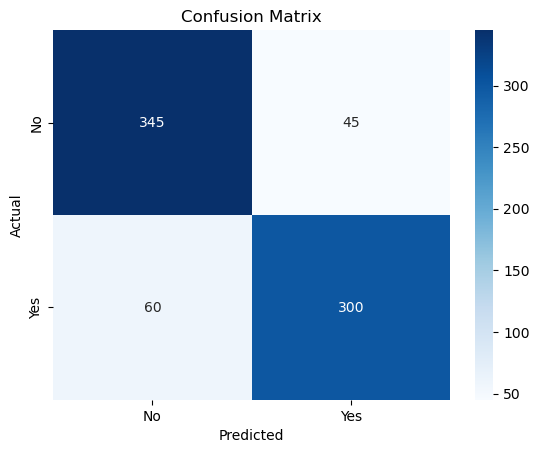
The logistic regression and decision tree analysis data included information on pumpkin seeds and their grading. Pumpkin seeds are frequently consumed as confection worldwide because of their adequate protein, fat, carbohydrate, and mineral contents. This study was carried out on the two most important and quality pumpkin seeds, ‘‘Urgup\_Sivrisi’’ and ‘‘Cercevelik’’, generally grown in Turkey's Urgup and Karacaoren regions. However, morphological measurements of 2500 pumpkin seeds of both varieties were made possible using the gray and binary threshold techniques. Considering morphological features, all the data were modeled with five different machine learning methods: Logistic Regression (LR), Multilayer Perceptrons (MLP), Support Vector Machine (SVM) and Random Forest (RF), and k-nearest Neighbor (k-NN), which further determined the most successful method for classifying pumpkin seed varieties. However, the models' performances were determined with the help of the ten-fold cross-validation method. The accuracy rates of the classifiers were obtained as LR 87.92 percent, MLP 88.52 percent, SVM 88.64 percent, RF 87.56 percent, and k-NN 87.64 percent. The variable ‘Class’ was taken as the dependent variable to fit the model.

**Results and interpretation**

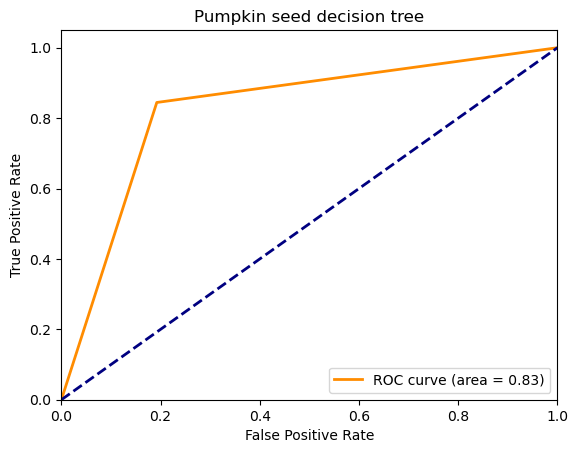
**Regression Model Coefficients**

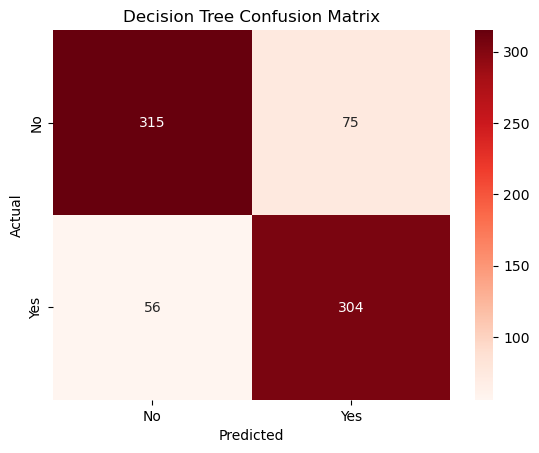
|  |  |
| --- | --- |
| Variable | Coefficient |
| Area | 0.001686 |
| Perimeter | -0.000570 |
| Major Axis Length | -0.000298 |
| Minor Axis Length | -0.073879 |
| Convex Area | -0.001644 |
| Equiv Diameter | -0.033031 |
| Eccentricity | -0.000041 |
| Solidity | -0.000196 |
| Extent | -0.000197 |
| Roundness | -0.000318 |
| Aspect Ration | 0.000529 |
| Compactness | -0.000298 |



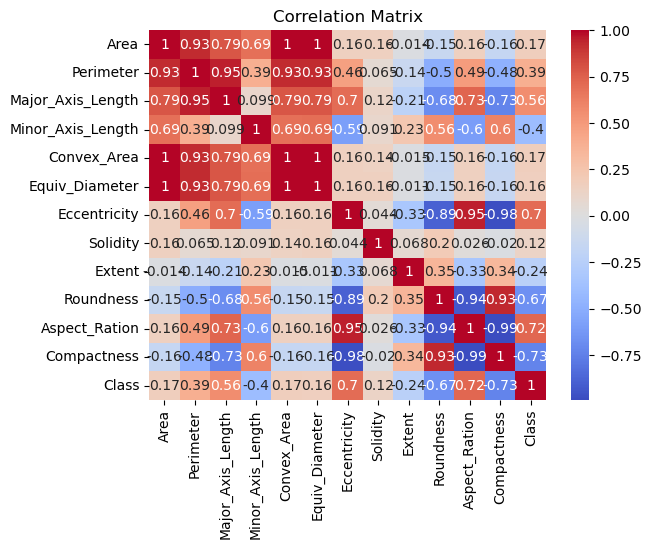


**Decision Tree**





**Correlation Matrix for the Pumpkin seed dataset**



**Area:** A one-unit increase in area corresponds to a rise of approximately 0.0017 in the outcome.

**Perimeter**: A one-unit increase in perimeter leads to a decrease of about 0.0006 in the outcome.

**Major Axis Length**: A one-unit increase in major axis length results in an increase of approximately 0.0568 in the outcome.

**Minor Axis Length**: A one-unit increase in minor axis length corresponds to a decrease of around 0.0739 in the outcome.

**Convex Area**: A one-unit increase in convex area leads to a decrease of about 0.0016.

**Equiv Diameter**: A one-unit increase in equivalent diameter results in a decrease of approximately 0.0330 in the outcome.

**Eccentricity:** A one-unit increase in eccentricity has a negligible impact (around -0.00004) on the outcome.

**Solidity:** A one-unit increase in solidity leads to a decrease of about 0.0002 in the outcome.

**Extent**: A one-unit increase in extent corresponds to a decrease of approximately 0.0002 in the outcome.

**Roundness**: A one-unit increase in roundness results in a decrease of around 0.0003 in the outcome.

**Aspect Ratio:** A one-unit increase in aspect ratio corresponds to a rise of approximately 0.0005 in the outcome.

**Compactness:** A one-unit increase in compactness leads to a decrease of about 0.0003 in the outcome.

**AUC:** At 0.93, the AUC of the ROC of the regression model indicates a better discrimination power than the decision tree model with an AUC of 0.83. Hence, it can be concluded that the regression model is a better fit.

**Part B**

**Probit Regression**

Probit regression is a statistical model used for binary classification, similar to logistic regression. It predicts the probability of an event (e.g., success/failure) based on predictor variables. Unlike logistic regression, it uses the cumulative distribution function (CDF) of the standard normal distribution. Probit models find applications in credit risk assessment, drug effectiveness studies, and marketing response modeling. Their business significance lies in making informed decisions by understanding the impact of various factors on binary outcomes, such as customer conversions, loan defaults, or product adoption**.**

The NSSO data was used to fit a probit regression model to determine if certain socio-economic factors influence the consumption of chicken across various states in the country.

**Results**

Probit Regression Results

==============================================================================

Dep. Variable: chicken\_q No. Observations: 101662

Model: Probit Df Residuals: 101658

Method: MLE Df Model: 3

Date: Mon, 01 Jul 2024 Pseudo R-squ.: 0.01405

Time: 23:06:23 Log-Likelihood: -11752.

converged: True LL-Null: -11920.

Covariance Type: nonrobust LLR p-value: 2.615e-72

==================================================================================

coef std err z P>|z| [0.025 0.975]

----------------------------------------------------------------------------------

const -2.2494 0.054 -41.604 0.000 -2.355 -2.143

Age 0.0015 0.001 2.205 0.027 0.000 0.003

Marital\_Status -0.0337 0.023 -1.483 0.138 -0.078 0.011

Education 0.0420 0.002 17.326 0.000 0.037 0.047

==================================================================================

slno grp Round\_Centre FSU\_number Round fds

0 1 40999999999999992652495293775872.0 1 41000 68

1 2 40999999999999992652495293775872.0 1 41000 68

2 3 40999999999999992652495293775872.0 1 41000 68

3 4 40999999999999992652495293775872.0 1 41000 68

4 5 40999999999999992652495293775872.0 1 41000 68

Schedule\_Number Sample Sector state State\_Region ... pickle\_v fds

0 10 1 2 24 242 ... 0.0

1 10 1 2 24 242 ... 0.0

2 10 1 2 24 242 ... 0.0

3 10 1 2 24 242 ... 0.0

4 10 1 2 24 242 ... 0.0

sauce\_jam\_v Othrprocessed\_v Beveragestotal\_v foodtotal\_v foodtotal\_q \

0 0.0 0.0 0.000000 1141.492400 30.942394

1 0.0 0.0 17.500000 1244.553500 29.286153

2 0.0 0.0 0.000000 1050.315400 31.527046

3 0.0 0.0 33.333333 1142.591667 27.834607

4 0.0 0.0 75.000000 945.249500 27.600713

state\_1 Region fruits\_df\_tt\_v fv\_tot

0 GUJ 2 12.000000 154.18

1 GUJ 2 333.000000 484.95

2 GUJ 2 35.000000 214.84

3 GUJ 2 168.333333 302.30

4 GUJ 2 15.000000 148.00

**Interpretation**

**Model Summary:**

* The model successfully converged after seven iterations.
* The dependent variable is chicken\_q.
* The model includes three predictor variables: Age, marital status, and Education.
* The pseudo-R-squared value (a measure of model fit) is approximately 0.014, indicating limited explanatory power.

**Coefficients:**

* The coefficients represent each predictor's impact on the event's probability (chicken consumption in this case).
* For example:
* A one-unit increase in Age corresponds to a small positive change in the probability.
* Marital status has a negligible impact.
* A one-unit increase in Education leads to an increase in the probability.

**Part C**

**Tobit Regression**

Tobit regression models are used when dealing with censored or truncated data. They combine a linear regression component (for observed data) with a variable latent element (for unobserved data). These models are valuable in understanding factors affecting censored outcomes, such as income or expenditure. Businesses can use Tobit models to optimize resource allocation and improve decision-making**.**

The NSSO data was used to fit a Tobit regression model with the variable ‘MPCE\_URP’ as the dependent variable and socio-economic factors as the independent variables.

**Results**

Tobit Results

==============================================================================

Dep. Variable: MPCE\_URP Log-Likelihood: 333.38

Model: Tobit AIC: -652.8

Method: Maximum Likelihood BIC: -586.1

Date: Tue, 02 Jul 2024

Time: 00:21:00

No. Observations: 101624

Df Residuals: 101618

Df Model: 5

=========================================================================================

coef std err z P>|z| [0.025 0.975]

-----------------------------------------------------------------------------------------

const -0.0023 0.057 -0.041 0.967 -0.114 0.110

Whether\_owns\_any\_land -0.0016 0.018 -0.088 0.930 -0.036 0.033

hhdsz -0.0016 0.002 -0.862 0.389 -0.005 0.002

Religion 0.0026 0.004 0.688 0.491 -0.005 0.010

Social\_Group 0.0050 0.002 2.204 0.027 0.001 0.009

Regular\_salary\_earner 0.0018 0.024 0.072 0.943 -0.046 0.050

par0 0.0803 0.004 20.898 0.000 0.073 0.088

**Interpretation**

Dependent Variable: MPCE\_URP

**Model Information:**

Log-Likelihood: Indicates the maximized log-likelihood function at the estimated coefficients.

**Coefficients Interpretation**

**Constant (const):** Not significantly different from zero (p-value = 0.967), suggesting it may not contribute significantly to the model.

**Whether\_owns\_any\_land:** Not significant (p-value = 0.930), indicating it does not have a statistically significant effect on the dependent variable.

**hhdsz (Household size)**: Not significant (p-value = 0.389), suggesting no significant relationship with the dependent variable.

**Religion**: Not significant (p-value = 0.491), implying no significant impact on the dependent variable.

**Social\_Group:** Significant (p-value = 0.027), suggesting a statistically significant effect on the dependent variable.

**Regular\_salary\_earner:** Not significant (p-value = 0.943), indicating no significant impact on the dependent variable.

par0: Significant (p-value < 0.001), suggesting a substantial and statistically significant impact on the dependent variable.

**Conclusion**

The Tobit model results suggest that Social\_Group and par0 are statistically significant predictors of MPCE\_URP. At the same time, other variables such as household characteristics (hhdsz), ownership of land (Whether\_owns\_any\_land), religion, and employment status (Regular\_salary\_earner) do not show significant effects. These interpretations are based on the provided coefficients, standard errors, and p-values.