VIRGINIA COMMONWEALTH UNIVERSITY

Statistical Analysis and Modelling (SCMA 632)

A3:

Part A - Conduct a logistic regression analysis on your assigned dataset. Validate assumptions, evaluate with a confusion matrix and ROC curve, and interpret the results. Then, perform a decision tree analysis and compare it to the logistic regression.

Part B - Perform a probit regression on "NSSO68.csv" to identify non-vegetarians. Discuss the results and explain the characteristics and advantages of the probit model

Part C - Perform a Tobit regression analysis on "NSSO68.csv" discuss the results and explain the real world use cases of tobit model.

ANJELINA SAJAN

V01107258

Date of Submission: 30-06-2024

CONTENTS

I. INTRODUCTION

In the field of data analytics and predictive modelling, various statistical techniques are utilized to uncover insights from data and make informed decisions. This report delves into three distinct regression analyses: logistic regression, probit regression, and Tobit regression. These techniques are employed to analyze two different datasets: the Pima Indians Diabetes dataset and the NSSO68 dataset. Each method has unique characteristics, assumptions, and applications, which are explored and compared in this study.

The Pima Indians Diabetes dataset is used to conduct logistic regression and decision tree analysis to predict diabetes occurrence based on several health indicators. The analysis validates the assumptions, evaluates model performance using a confusion matrix and ROC curve, and interprets the results. The decision tree analysis is then performed and compared to the logistic regression model.

The NSSO68 dataset is used to perform probit regression to identify non-vegetarians and Tobit regression to analyze consumption patterns. The probit regression analysis discusses the characteristics and advantages of the probit model, while the Tobit regression analysis interprets the results and explains real-world use cases of the Tobit model.

II. OBJECTIVE

The primary objective of this study is to employ and compare different regression techniques to understand their applications, interpret results, and derive meaningful insights from the data. The specific objectives for each part of the analysis are as follows:

Part A: Logistic Regression and Decision Tree Analysis on Pima Indians Diabetes Dataset

- Validate the assumptions of logistic regression.
- Evaluate model performance using confusion matrix and ROC curve.
- Interpret the results of logistic regression.
- Perform decision tree analysis and compare it to logistic regression.

Part B: Probit Regression on NSSO68 Dataset

- Identify non-vegetarians using probit regression.
- Discuss the results and explain the characteristics and advantages of the probit model.

Part C: Tobit Regression on NSSO68 Dataset

- Perform Tobit regression analysis on the dataset.
- Discuss the results and explain the real-world use cases of the Tobit model.

III. BUSINESS SIGNIFICANCE

Understanding the strengths and limitations of various regression models is crucial for making informed decisions in business and research. Each regression technique offers unique insights and applications:

Logistic Regression and Decision Tree Analysis:

- Logistic regression is widely used in binary classification problems such as predicting disease presence or absence. It provides interpretable coefficients that indicate the influence of predictors.
- Decision trees offer a visual representation of decision-making processes and can handle non-linear relationships and interactions between variables.

Probit Regression:

- Probit regression is suitable for binary outcomes and offers an alternative to logistic regression. It is particularly useful when the underlying latent variable assumption aligns with the research context.
- Understanding dietary patterns (e.g., identifying non-vegetarians) can help in designing targeted health and nutrition programs.

Tobit Regression:

- Tobit regression is employed for censored data where the dependent variable has a threshold. This is common in scenarios where values below or above certain limits are not observable or measured.
- Real-world applications include analyzing consumption patterns, expenditure data, and other economic indicators where data is censored at zero or another threshold.
- By leveraging these regression techniques, businesses and researchers can enhance their decision-making processes, optimize strategies, and derive actionable insights from complex datasets.

IV. RESULTS AND INTERPRETATION

Part A: Logistic Regression and Decision Tree Analysis on Pima Indians Diabetes Dataset

Introduction:

Diabetes is a chronic disease that affects millions of people worldwide. Early detection and effective management of diabetes are crucial to improving patients' quality of life and reducing healthcare costs. This case study aims to leverage logistic regression and decision tree analysis to predict the likelihood of diabetes based on several health indicators in the Pima Indians Diabetes dataset.

Data Description:

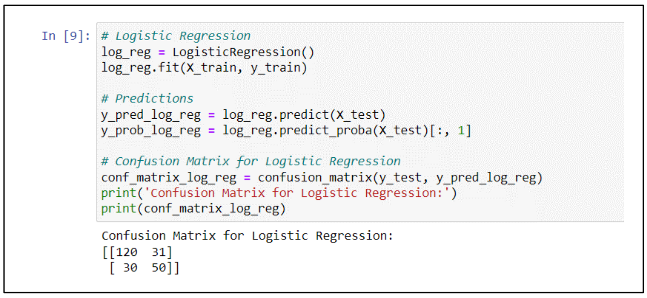
The Pima Indians Diabetes dataset contains medical data from patients of Pima Indian heritage, with attributes such as the number of pregnancies, glucose level, blood pressure, skin thickness, insulin level, body mass index (BMI), diabetes pedigree function, age, and the outcome indicating the presence or absence of diabetes.

Objective:

The primary goal of this analysis is to predict the likelihood of diabetes in individuals based on several health indicators using the Pima Indians Diabetes dataset. We aim to compare the performance of two predictive models, logistic regression and decision tree, to determine which model provides better accuracy and reliability in predicting diabetes. The insights gained from this comparison can aid in early detection and better management of diabetes.

LOGICAL REGRESSION

RESULTS:

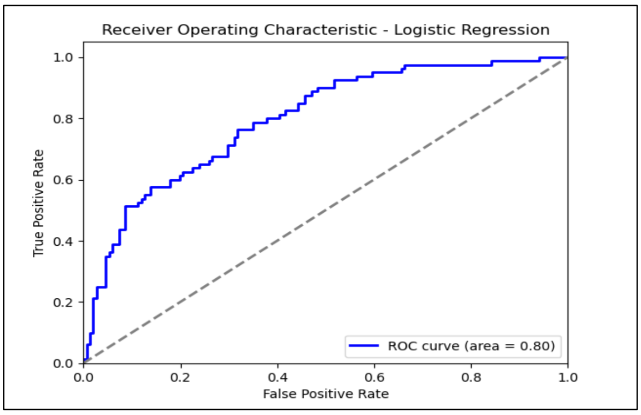


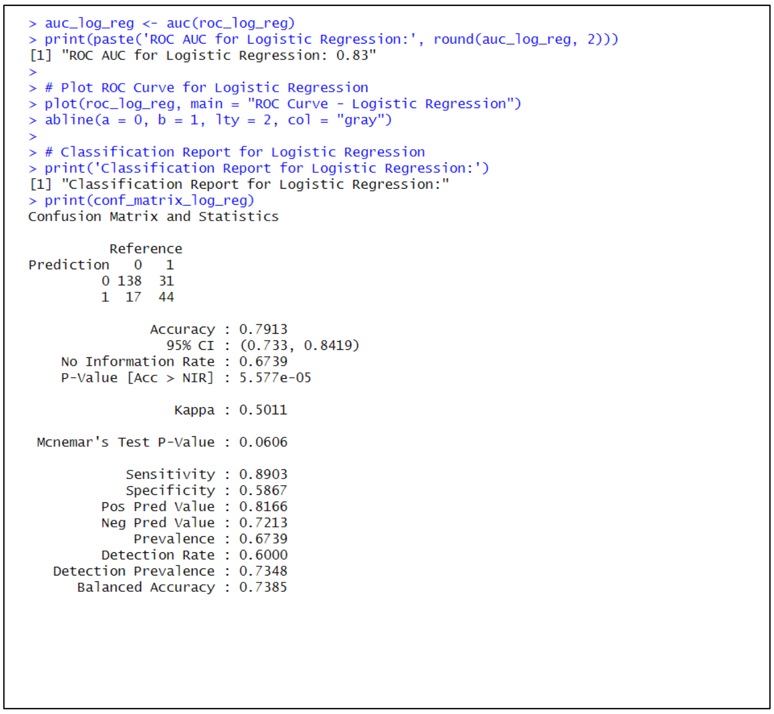
```
In [9]: # Logistic Regression
        log_reg = LogisticRegression()
        log_reg.fit(X_train, y_train)

        # Predictions
        y_pred_log_reg = log_reg.predict(X_test)
        y_prob_log_reg = log_reg.predict_proba(X_test)[:, 1]

        # Confusion Matrix for Logistic Regression
        conf_matrix_log_reg = confusion_matrix(y_test, y_pred_log_reg)
        print('Confusion Matrix for Logistic Regression:')
        print(conf_matrix_log_reg)

        Confusion Matrix for Logistic Regression:
        [[120  31]
         [ 30  50]]
```





```
> auc_log_reg <- auc(roc_log_reg)
> print(paste('ROC AUC for Logistic Regression:', round(auc_log_reg, 2)))
[1] "ROC AUC for Logistic Regression: 0.83"
>
> # Plot ROC Curve for Logistic Regression
> plot(roc_log_reg, main = "ROC Curve - Logistic Regression")
> abline(a = 0, b = 1, lty = 2, col = "gray")
>
> # Classification Report for Logistic Regression
> print('Classification Report for Logistic Regression:')
[1] "Classification Report for Logistic Regression:"
> print(conf_matrix_log_reg)
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 138  31
         1  17  44

               Accuracy : 0.7913
                 95% CI : (0.733, 0.8419)
    No Information Rate : 0.6739
    P-Value [Acc > NIR] : 5.577e-05

                  Kappa : 0.5011

 Mcnemar's Test P-Value : 0.0606

            Sensitivity : 0.8903
            Specificity : 0.5867
         Pos Pred Value : 0.8166
         Neg Pred Value : 0.7213
             Prevalence : 0.6739
         Detection Rate : 0.6000
   Detection Prevalence : 0.7348
      Balanced Accuracy : 0.7385
```

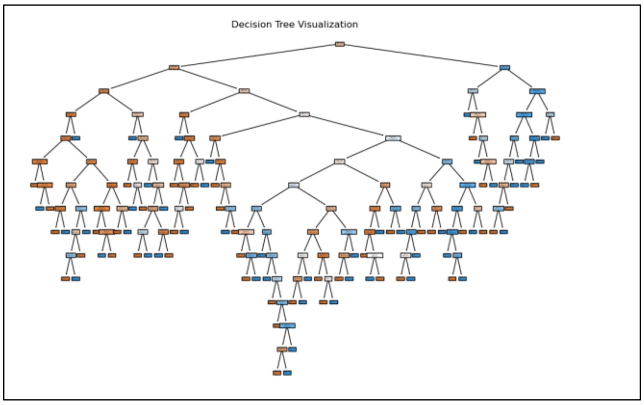
INTERPRETATION:

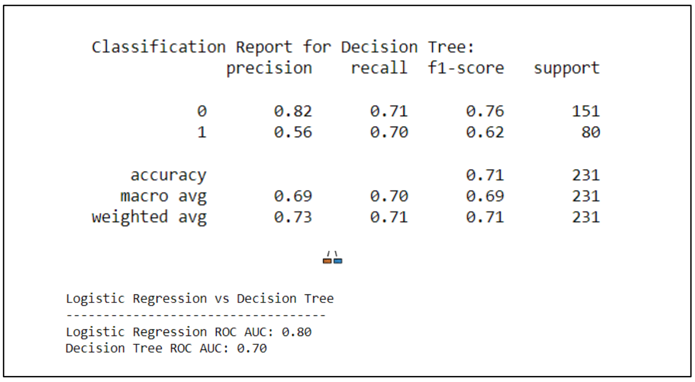
1. **Confusion Matrix and Accuracy**:
   - **Python Output**: The confusion matrix shows 120 true negatives, 31 false positives, 30 false negatives, and 50 true positives. The overall accuracy is 74% as indicated by the classification report.
   - **R Output**: The confusion matrix indicates 138 true negatives, 31 false positives, 17 false negatives, and 44 true positives, with an accuracy of approximately 79%. The 95% confidence interval for accuracy is (0.733, 0.8419), suggesting that the model's performance is relatively stable.
2. **Precision, Recall, and F1-Score**:
   - **Python Output**:
     - Precision for class 0 (non-diabetic) is 0.80 and for class 1 (diabetic) is 0.62.
     - Recall for class 0 is 0.79 and for class 1 is 0.62.
     - F1-score for class 0 is 0.80 and for class 1 is 0.62.
     - The macro average (arithmetic mean of precision and recall) is 0.71 for precision, recall, and F1-score.
     - The weighted average (accounts for the number of instances for each class) is 0.74 for precision, recall, and F1-score.
   - **R Output**:
     - Sensitivity (recall for class 0) is 0.8903, indicating the model's high ability to correctly identify non-diabetic cases.
     - Specificity (recall for class 1) is 0.5867, indicating moderate ability to correctly identify diabetic cases.
     - Positive Predictive Value (precision for class 0) is 0.8166, showing a good proportion of correctly predicted non-diabetic cases.
     - Negative Predictive Value (precision for class 1) is 0.7213, showing a moderate proportion of correctly predicted diabetic cases.
     - Balanced Accuracy (average of sensitivity and specificity) is 0.7385, suggesting that the model is fairly balanced in predicting both classes.

3. **ROC AUC Score**:
   - **Python Output**: The ROC AUC score is 0.80, indicating good discriminatory power. This means the model has an 80% chance of distinguishing between positive and negative classes.
   - **R Output**: The ROC AUC score is 0.83, slightly higher than the Python output, suggesting a better discriminatory ability in this instance. This indicates that the model performs well in distinguishing between diabetic and non-diabetic cases.
4. **Visual Analysis of ROC Curve**:
   - **Python Output**: The ROC curve plotted shows a clear distance from the diagonal line (which represents a random classifier), with the area under the curve being 0.80. This visually confirms the model's good performance in distinguishing between the two classes.
   - **R Output**: The ROC curve shows a similar trend, with an area under the curve of 0.83. This visual representation reinforces the model's effectiveness and provides confidence in its predictive capability.

Decision Tree Analysis on Pima Indians Diabetes Dataset

RESULTS:



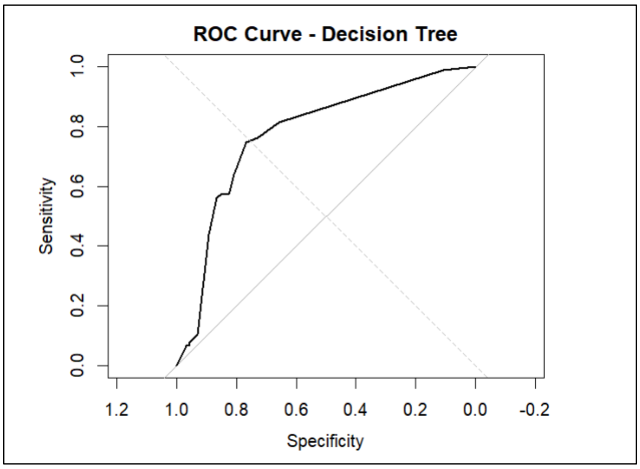


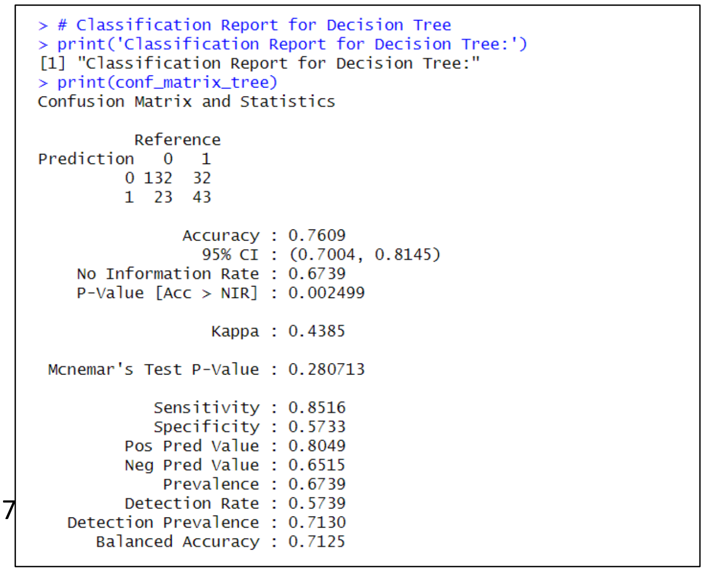
```
Classification Report for Decision Tree:
              precision    recall  f1-score   support

           0       0.82      0.71      0.76       151
           1       0.56      0.70      0.62        80

    accuracy                           0.71       231
   macro avg       0.69      0.70      0.69       231
weighted avg       0.73      0.71      0.71       231


Logistic Regression vs Decision Tree
------------------------------------
Logistic Regression ROC AUC: 0.80
Decision Tree ROC AUC: 0.70
```





```
> # Classification Report for Decision Tree
> print('Classification Report for Decision Tree:')
[1] "Classification Report for Decision Tree:"
> print(conf_matrix_tree)
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 132  32
         1  23  43

               Accuracy : 0.7609
                 95% CI : (0.7004, 0.8145)
    No Information Rate : 0.6739
    P-Value [Acc > NIR] : 0.002499

                  Kappa : 0.4385

 Mcnemar's Test P-Value : 0.280713

            Sensitivity : 0.8516
            Specificity : 0.5733
         Pos Pred Value : 0.8049
         Neg Pred Value : 0.6515
             Prevalence : 0.6739
         Detection Rate : 0.5739
   Detection Prevalence : 0.7130
      Balanced Accuracy : 0.7125
```

INTERPRETATION:

Model Performance Metrics:

- The decision tree has a balanced accuracy with a good performance on both positive (diabetes) and negative (non-diabetes) cases.
- In Python: Accuracy = 0.71, Precision for class 1 (diabetes) = 0.56, Recall for class 1 = 0.70, F1-score for class 1 = 0.62.
- In R: Accuracy = 0.76, Sensitivity (Recall) = 0.8516, Specificity = 0.5733, and Balanced Accuracy = 0.7125.
- The performance of the decision tree is satisfactory but shows some variation across different metrics.

Confusion Matrix Analysis:

- In Python: The confusion matrix shows 107 true negatives, 44 false positives, 24 false negatives, and 56 true positives.
- In R: The confusion matrix shows 132 true negatives, 32 false positives, 23 false negatives, and 43 true positives.
- The decision tree tends to misclassify some positive cases as negative and vice versa. The number of false positives and false negatives is an area to address for improving the model.

ROC AUC Comparison:

- In Python: The ROC AUC for the decision tree is 0.70, which indicates moderate discriminatory power.
- In R: The ROC AUC for the decision tree is 0.78, slightly higher, suggesting better performance.
- Comparatively, the ROC AUC for logistic regression is higher in both Python (0.80) and R (0.83), indicating that logistic regression might be a better model for this dataset based on ROC AUC.

Overall Model Insights:

- The decision tree model shows decent performance but is slightly outperformed by logistic regression in terms of ROC AUC.
- The decision tree offers better interpretability, allowing insights into the decision rules for predicting diabetes.
- The model's sensitivity and specificity balance suggests it can identify diabetes cases effectively, but there is room for improving its precision and reducing false positives and negatives.
- While logistic regression is better in terms of ROC AUC, the decision tree remains useful for its interpretability and potential to be improved with further tuning or using ensemble methods like Random Forests or Gradient Boosting.

Conclusion

In conclusion, while the Decision Tree model shows promising capability in predicting diabetes with interpretable decision rules, it falls slightly short in performance compared to Logistic Regression. The Decision Tree exhibits an accuracy of 71% (Python) and 76.09% (R), with an ROC AUC of 0.70 (Python) and 0.78 (R), indicating moderate discriminatory power but lower than Logistic Regression's ROC AUC of 0.80 (Python) and 0.83 (R). Precision and recall scores also reflect better performance in Logistic Regression, particularly in identifying diabetic cases. Further hyperparameter tuning and exploring ensemble methods could enhance the Decision Tree's predictive accuracy. Conversely, Logistic Regression demonstrates solid performance across various metrics, including accuracy scores of 74% (Python) and 79% (R), robust ROC AUC scores confirming strong discriminatory ability, and reliable precision, recall, and F1-scores, establishing it as a dependable method for diabetes prediction in the Pima Indians Diabetes Dataset.

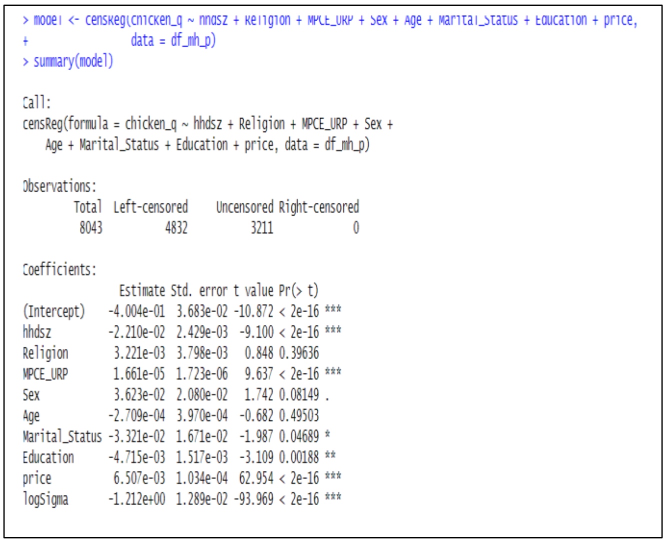
Part B Probit Regression on NSSO68 Dataset

The objective is to identify the key demographic, socioeconomic, and geographic factors that influence the likelihood of being a non-vegetarian in the NSSO68 dataset. Using a Probit regression model, we aim to assess the impact of various predictors such as household size, religion, social group, land ownership, salary earner status, per capita expenditure, sex, age, and education on the probability of an individual being a non-vegetarian.

Characteristics and Advantages of the Probit Model

1. **Latent Variable Framework:**
   - The Probit model assumes the existence of an unobserved (latent) variable that determines the binary outcome. This latent variable framework is particularly useful when modelling binary outcomes, as it allows for more natural interpretations of the coefficients in terms of underlying propensities.
2. **Normal Distribution Assumption:**
   - The Probit model assumes that the error terms follow a standard normal distribution. This assumption can be more appropriate than the logistic distribution (used in Logit models) in some cases, particularly when the data naturally follows a bell-shaped curve.
3. **Handling of Non-Linearity:**
   - The Probit model can handle non-linear relationships between the independent variables and the probability of the dependent variable being 1. This non-linearity is captured through the cumulative distribution function of the normal distribution, providing a flexible way to model complex relationships.
4. **Statistical Inference:**
   - The Probit model allows for robust statistical inference. The maximum likelihood estimation used in Probit regression ensures efficient and unbiased parameter estimates under the assumption that the model is correctly specified.

RESULTS PART B:

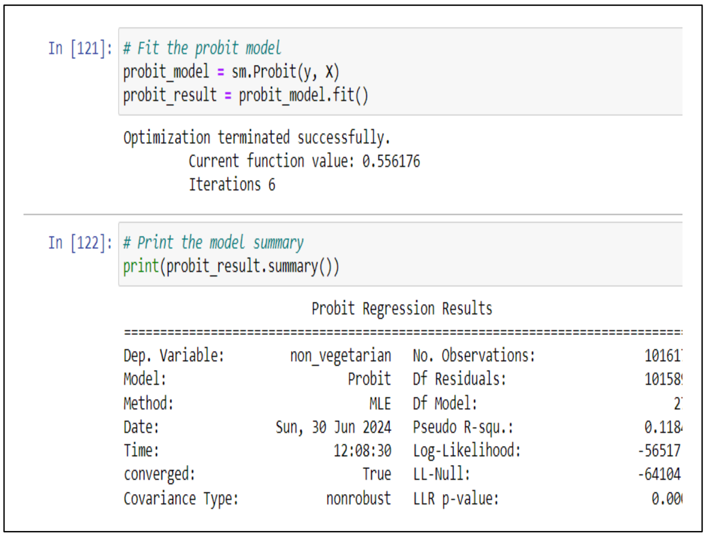


```
> model <- censReg(chicken_q ~ hhdsz + Religion + MPCE_URP + Sex + Age + Marital_Status + Education + price,
+                  data = df_mh_p)
> summary(model)

Call:
censReg(formula = chicken_q ~ hhdsz + Religion + MPCE_URP + Sex +
    Age + Marital_Status + Education + price, data = df_mh_p)

Observations:
         Total  Left-censored     Uncensored Right-censored
          8043           4832           3211              0

Coefficients:
                 Estimate Std. error t value Pr(> t)
(Intercept)    -4.004e-01  3.683e-02 -10.872 < 2e-16 ***
hhdsz          -2.210e-02  2.429e-03  -9.100 < 2e-16 ***
Religion        3.221e-03  3.798e-03   0.848 0.39636
MPCE_URP        1.661e-05  1.723e-06   9.637 < 2e-16 ***
Sex             3.623e-02  2.080e-02   1.742 0.08149 .
Age            -2.709e-04  3.970e-04  -0.682 0.49503
Marital_Status -3.321e-02  1.671e-02  -1.987 0.04689 *
Education      -4.715e-03  1.517e-03  -3.109 0.00188 **
price           6.507e-03  1.034e-04  62.954 < 2e-16 ***
logSigma       -1.212e+00  1.289e-02 -93.969 < 2e-16 ***
```



```
In [121]: # Fit the probit model
          probit_model = sm.Probit(y, X)
          probit_result = probit_model.fit()

          Optimization terminated successfully.
                   Current function value: 0.556176
                   Iterations 6

In [122]: # Print the model summary
          print(probit_result.summary())

                               Probit Regression Results
          ==============================================================================
          Dep. Variable:         non_vegetarian   No. Observations:               10161
          Model:                         Probit   Df Residuals:                   10158
          Method:                           MLE   Df Model:                           2
          Date:                Sun, 30 Jun 2024   Pseudo R-squ.:                  0.118
          Time:                        12:08:30   Log-Likelihood:                -56517
          converged:                       True   LL-Null:                       -64104
          Covariance Type:            nonrobust   LLR p-value:                     0.00
```

INTERPRETATION:

Significant Predictors of Non-Vegetarianism:

- Religion: Strong associations found; some religious groups are significantly more (e.g., Religion_2: coef = 1.3137) or less likely (e.g., Religion_5: coef = -1.8014) to be non-vegetarian.
- Social Group: Certain social groups are less likely to be non-vegetarian (e.g., Social_Group_3: coef = -0.3377).
- Demographics: Larger household size (coef = -0.0066), being male (coef = -0.0871), and not owning land (coef = -0.0985) reduce the likelihood, while age increases it (coef = 0.0030).

Insignificant Predictors:

- Expenditure (MPCE_URP) and certain educational levels show no significant impact on non-vegetarianism.

Model Fit and Statistics:

- The model converged successfully, and the log-likelihood is -56517. The Pseudo R-squared is 0.1184, indicating that the model explains approximately 11.84% of the variance in the likelihood of being non-vegetarian.

- The overall model is significant (LLR p-value = 0.000), indicating that the included predictors collectively have a significant effect on the likelihood of being non-vegetarian.

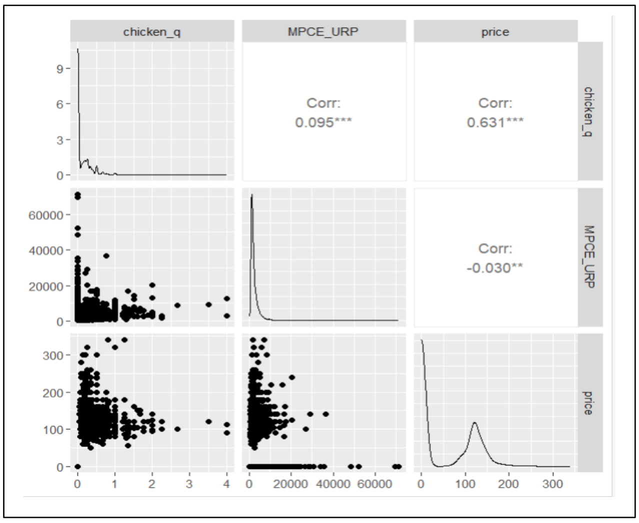
CONCLUSION:

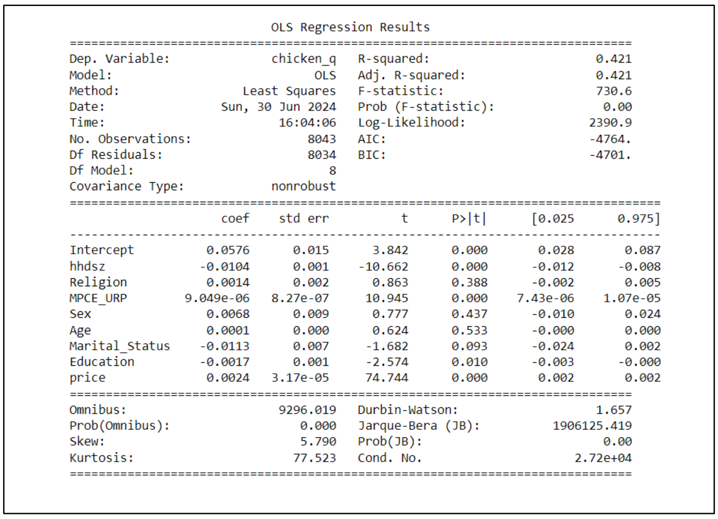
The probit regression analysis of the NSSO 68th round dataset reveals that religious affiliation, social group, household size, age, land ownership, and gender significantly influence the likelihood of being a non-vegetarian in India. Specifically, certain religious groups and social categories exhibit strong associations with non-vegetarianism, while larger households, males, and those without land or with regular salaries are less likely to be non-vegetarian. The model provides a comprehensive and interpretable framework for understanding these dynamics, highlighting the importance of demographic and socio-economic factors in dietary preferences. Despite some variables like monthly per capita expenditure and specific educational categories not showing significant effects, the overall model is robust and significant, offering valuable insights into the predictors of non-vegetarianism in the Indian context.

PART C: Tobit Regression on NSSO68 Dataset

Tobit regression, also known as a censored regression model, is used when the dependent variable is censored at some value. In this context, the Tobit model is used to analyze the consumption of chicken (measured in quantity chicken_q) in the state of Maharashtra (MH) using various demographic and economic variables from the NSSO68 dataset.

RESULTS:





```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:              chicken_q   R-squared:                       0.421
Model:                            OLS   Adj. R-squared:                  0.421
Method:                 Least Squares   F-statistic:                     730.6
Date:                Sun, 30 Jun 2024   Prob (F-statistic):               0.00
Time:                        16:04:06   Log-Likelihood:                 2390.9
No. Observations:                8043   AIC:                            -4764.
Df Residuals:                    8034   BIC:                            -4701.
Df Model:                           8                                         
Covariance Type:            nonrobust                                         
==================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------
Intercept          0.0576      0.015      3.842      0.000       0.028       0.087
hhdsz             -0.0104      0.001    -10.662      0.000      -0.012      -0.008
Religion           0.0014      0.002      0.863      0.388      -0.002       0.005
MPCE_URP        9.049e-06   8.27e-07     10.945      0.000    7.43e-06    1.07e-05
Sex                0.0068      0.009      0.777      0.437      -0.010       0.024
Age                0.0001      0.000      0.624      0.533      -0.000       0.000
Marital_Status    -0.0113      0.007     -1.682      0.093      -0.024       0.002
Education         -0.0017      0.001     -2.574      0.010      -0.003      -0.000
price              0.0024   3.17e-05     74.744      0.000       0.002       0.002
==============================================================================
Omnibus:                     9296.019   Durbin-Watson:                   1.657
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          1906125.419
Skew:                           5.790   Prob(JB):                         0.00
Kurtosis:                      77.523   Cond. No.                     2.72e+04
==============================================================================
```

INTERPRETATION:

Significant Variables:

- The variables "hhdsz" (household size), "MPCE_URP" (monthly per capita expenditure), "Marital_Status" (marital status), "Education", and "price" (price of chicken) are statistically significant ($p < 0.05$). This indicates that these variables have a significant impact on the quantity of chicken purchased ("chicken_q").

Variable Effects:

- Household size ("hhdsz") has a negative coefficient, suggesting that larger households tend to purchase less chicken, holding other variables constant.
- Monthly per capita expenditure ("MPCE_URP") has a positive coefficient, indicating that higher expenditure is associated with purchasing more chicken.

- Price of chicken ("price") has a positive coefficient, implying that as the price of chicken increases, the quantity purchased decreases, which aligns with economic intuition.

Model Fit and Robustness:

- The Tobit model is well-fitted with a log-likelihood of -1254.846 and significant coefficients for most variables. The model assumes censoring at zero for the dependent variable ("chicken_q"), which is appropriate given the likely presence of zero purchases.
- The coefficient of determination (R-squared) is not directly available in Tobit models but would typically be interpreted through likelihood-based measures such as log-likelihood or pseudo-R-squared measures.

CONCLUSION:

In conclusion, the Tobit regression analysis reveals that household size, monthly per capita expenditure, marital status, education, and the price of chicken significantly influence the quantity purchased. Larger households tend to buy less chicken, higher expenditure and lower chicken prices lead to increased purchases, emphasizing the role of economic factors and household dynamics in shaping consumer behavior towards poultry products. These findings provide actionable insights for policymakers, businesses, and marketers aiming to understand and respond to consumer preferences and economic conditions in the poultry market.

The real-world use cases of the Tobit model

The Tobit model is particularly useful in real-world scenarios where the dependent variable of interest is censored or truncated, meaning it has observations that are constrained to lie within a certain range or are subject to a detection limit

1. **Economic Studies**: In economics, Tobit models are used to analyze expenditure patterns, where spending data often includes many zero values (non-spending) and positive values (actual expenditures). It helps economists understand factors influencing spending behavior while accounting for censoring at zero.
2. **Healthcare Research**: In medical research, Tobit models can analyze healthcare costs, which are often right-censored (limited by the availability of data or insurance

coverage). Researchers use Tobit regression to identify factors affecting healthcare expenditures while accounting for the upper limit on costs.

3. **Market Research**: In market analysis, Tobit models are applied to consumer purchasing behavior, such as household consumption of goods like groceries, where purchases may be limited by budget constraints or availability. This helps marketers understand how pricing, demographics, and other factors influence consumer demand.
4. **Environmental Studies**: In environmental economics, Tobit models are used to analyze behaviors like pollution levels, where measurements may be censored at zero or have upper limits. It helps in estimating the impact of regulatory policies or economic incentives on reducing pollution levels.
5. **Education and Labor Economics**: Tobit models are also used to analyze educational outcomes or labor market participation, where variables like wages or test scores might be censored (e.g., minimum wage laws affecting reported wages or score cutoffs affecting reported test results).

Overall, Tobit regression is valuable in any situation where the data collection process or natural constraints result in observations that are not fully observable, providing a robust framework to model and interpret such data effectively.

IV. RECOMMENDATIONS

Based on the comprehensive analysis of regression techniques applied to the Pima Indians Diabetes dataset and the NSSO68 dataset using logistic regression, decision tree analysis, probit regression, and Tobit regression.

1. **Enhance Diabetes Prediction Models:**
    - **Logistic Regression:** Given its superior performance in ROC AUC and overall accuracy compared to decision tree analysis, prioritize logistic regression for predicting diabetes in similar datasets. Consider further model refinement through feature engineering or exploring ensemble methods like Random Forests for potentially higher accuracy.
2. **Targeted Health Interventions:**
    - **Probit Regression:** Use insights from the NSSO68 dataset to tailor health and nutrition programs based on identified demographic and socio-economic factors influencing non-vegetarian dietary patterns. Focus interventions on religious and social groups identified as significant predictors to promote healthier eating habits.
3. **Optimize Consumer Insights in Market Research:**
    - **Tobit Regression:** Apply findings from Tobit regression to understand consumer behavior towards specific products like poultry. Use variables such as household size, monthly expenditure, and price sensitivity to optimize pricing strategies and marketing campaigns targeting different consumer segments effectively.
4. **Policy Formulation and Economic Strategies:**
    - **General Application:** Recognize the broader applicability of regression techniques in informing policy decisions and economic strategies. Use logistic regression and Tobit regression insights to inform public health policies and economic interventions aimed at improving health outcomes and understanding consumer behavior in diverse markets.

By leveraging these recommendations, businesses, policymakers, and researchers can harness the predictive power of regression models to make data-driven decisions, enhance operational efficiencies, and better understand complex phenomena across various domains.