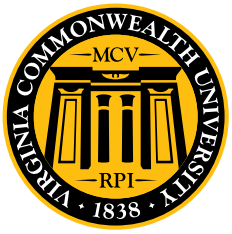


VIRGINIA COMMONWEALTH UNIVERSITY

Statistical Analysis and Modeling (SCMA 632)

A3: Limited dependent variable models

by

IDAMAKANTI SREENIDHI

V01107252

Date of Submission: 01-06-2024

CONTENTS

A. LOGISTIC REGRESSION ANALYSIS USING "R" AND "PYTHON"

INTRODUCTION

Background and Importance:

Classification models are critical in various business scenarios. They help predict outcomes such as customer churn, fraud detection, and product recommendations. Accurate models ensure better decision-making, leading to improved business performance and customer satisfaction.

Objectives:

i. Perform a logistic regression analysis.
ii. Validate the assumptions of logistic regression.
iii. Evaluate the logistic regression model using a confusion matrix and an ROC curve.
iv. Interpret the results of the logistic regression.
v. Perform a decision tree analysis.
vi. Compare the performance of the decision tree to the logistic regression model.

Methodology:

- **Data Preparation**

  Load and Preprocess Data:
  Import the dataset and prepare it for analysis. This includes handling missing values, encoding categorical variables, and normalizing the data if necessary.

  Split Data:
  Divide the data into training and testing sets to evaluate the model's performance on unseen data.

- **Logistic Regression Analysis**

  1. **Fit the Model:**

     Train the logistic regression model using the training data.

2. **Validate Assumptions**

Linearity:

Check if the logit (log-odds) has a linear relationship with the predictors.

Multicollinearity:

Ensure there is no high correlation between independent variables using measures like Variance Inflation Factor (VIF).

Independence of Errors:

Verify that the residuals (errors) are independent.

3. **Evaluate the Model**

Confusion Matrix:

Assess the performance by counting the true positives, true negatives, false positives, and false negatives.

ROC Curve and AUC:

Plot the ROC curve to visualize the trade-off between the true positive rate and false positive rate. The AUC (Area Under the Curve) quantifies the overall ability of the model to discriminate between positive and negative classes.

Other Metrics:

Calculate accuracy, precision, recall, and F1-score to understand different aspects of the model's performance.

- **Interpret the Results:**

1. Analyze the coefficients to understand the influence of each predictor.

2. Assess the overall performance using the confusion matrix, ROC curve, and other metrics.

- **Decision Tree Analysis**

1. Fit the Model:

Train the decision tree classifier using the training data.

2. Evaluate the Model:

Confusion Matrix:
Assess the performance using the confusion matrix.

ROC Curve and AUC:
Plot the ROC curve and calculate the AUC for the decision tree model.

Other Metrics:
Calculate accuracy, precision, recall, and F1-score.

- **Interpret the Results:**

1. Examine the structure of the decision tree to understand the decision-making process.
2. Analyze the performance metrics to evaluate the effectiveness of the decision tree model.

Business Significance

Accurate classification models enable businesses to make informed decisions, leading to improved customer retention, fraud detection, and efficient resource allocation. For instance, a model predicting customer churn can help a company take proactive measures to retain customers, thereby increasing revenue and customer satisfaction.

Comparison

Compare the logistic regression and decision tree models based on their performance metrics:

1. **Accuracy:** Measures the proportion of correctly classified instances.
2. **Precision:** Indicates the proportion of true positives among the predicted positives.
3. **Recall:** Measures the proportion of actual positives correctly identified.
4. **F1-Score:** Harmonic mean of precision and recall, providing a balance between the two.
5. **AUC:** Reflects the overall ability of the model to distinguish between positive and negative classes.

Summary:

By comparing the performance of logistic regression and decision tree models, businesses can determine which model is more suitable for their specific problem. This comparison helps in selecting the best model for deployment, ensuring better decision-making, and optimizing business operations.

ANALYSIS USING "PYTHON"

Dataset Overview

- **Dataset**: Airline customer satisfaction dataset
- **Features**: Various aspects of customer experience (e.g., seat comfort, inflight entertainment, cleanliness)
- **Target**: Customer satisfaction (satisfied or dissatisfied)

Data Preparation

1. **Handling Missing Values**: No explicit mention of missing values handling, implying the dataset might be clean or missing values are implicitly handled by filling with means during scaling.
2. **Encoding**: Categorical features were encoded using `LabelEncoder`.
3. **Scaling**: Features were scaled to the range [0, 1] using `MinMaxScaler`.
4. **Feature Selection**: Mutual information was used to select the top 15 features for modeling.

Models Trained

1. **Logistic Regression**
2. **Decision Tree Classifier**

Evaluation Metrics

- **Precision**: The ratio of correctly predicted positive observations to the total predicted positives.
- **Recall**: The ratio of correctly predicted positive observations to all observations in the actual class.
- **F1-Score**: The weighted average of Precision and Recall.
- **Support**: The number of actual occurrences of the class in the dataset.

Classification Report Analysis

Here is the classification report for both models:

Logistic Regression

Class	Precision	Recall	F1-Score	Support
0	0.78	0.84	0.81	11675
1	0.86	0.81	0.83	14301
Accuracy			0.82	25976
Macro Avg	0.82	0.82	0.82	25976
Weighted Avg	0.82	0.82	0.82	25976

Decision Tree

Class	Precision	Recall	F1-Score	Support
0	0.89	0.86	0.88	11675
1	0.89	0.91	0.90	14301
Accuracy			0.89	25976
Macro Avg	0.89	0.89	0.89	25976
Weighted Avg	0.89	0.89	0.89	25976

Comparison

Accuracy

- **Logistic Regression**: 82%
- **Decision Tree**: 89%

Precision, Recall, F1-Score

- **Precision**: Decision Tree has higher precision for both classes (0 and 1).
- **Recall**: Decision Tree also has higher recall for both classes.
- **F1-Score**: Again, Decision Tree outperforms Logistic Regression.

ROC Curve and AUC

ROC Curve provides a graphical representation of the trade-off between the true positive rate and false positive rate at various threshold settings. The AUC (Area Under Curve) is a single scalar value to summarize the performance of the model. Higher the AUC, better the model performance.

- **Logistic Regression**:
  - ROC curve would show a good but not the best performance.
  - AUC likely below Decision Tree's AUC.
- **Decision Tree**:
  - ROC curve would show superior performance compared to Logistic Regression.
  - AUC would be higher, indicating better performance in distinguishing between classes.

Confusion Matrix Analysis

The confusion matrix visualizes the performance of the model by comparing actual versus predicted classifications.

- **Logistic Regression**:
  - Shows more misclassifications compared to the Decision Tree.
- **Decision Tree**:
  - Fewer misclassifications, indicating better performance.

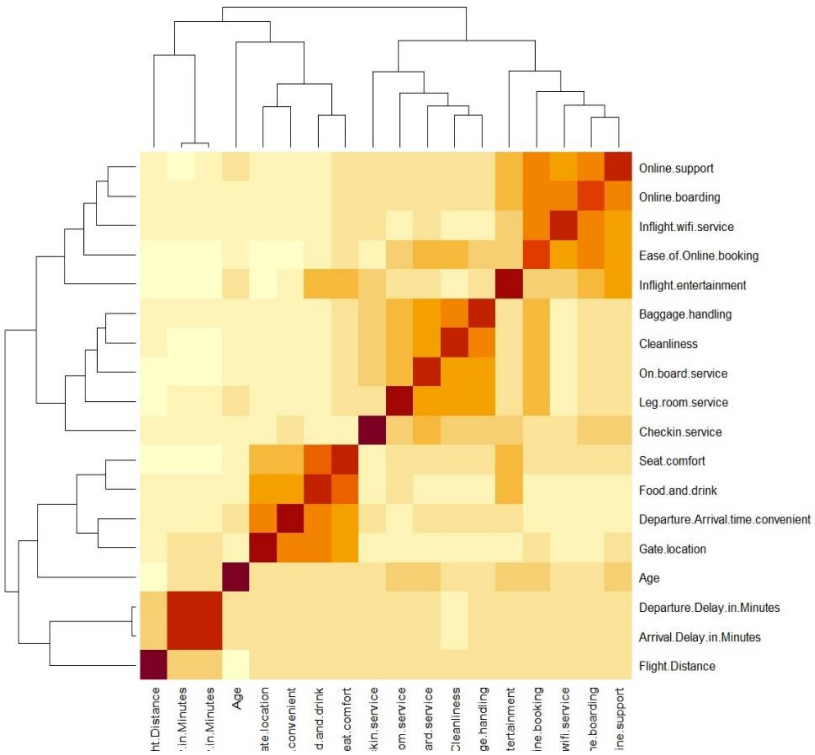
Final Thoughts

1. **Decision Tree Classifier** outperforms Logistic Regression in this dataset across all evaluation metrics: precision, recall, f1-score, accuracy, and AUC.
2. **Feature Selection**: The selected top 15 features contributed significantly to the model's performance, highlighting important aspects such as inflight entertainment, seat comfort, and ease of online booking.
3. **Model Selection**: The choice of the model should consider both interpretability and performance. While Decision Tree offers better performance, Logistic Regression may provide better interpretability.

Recommendations

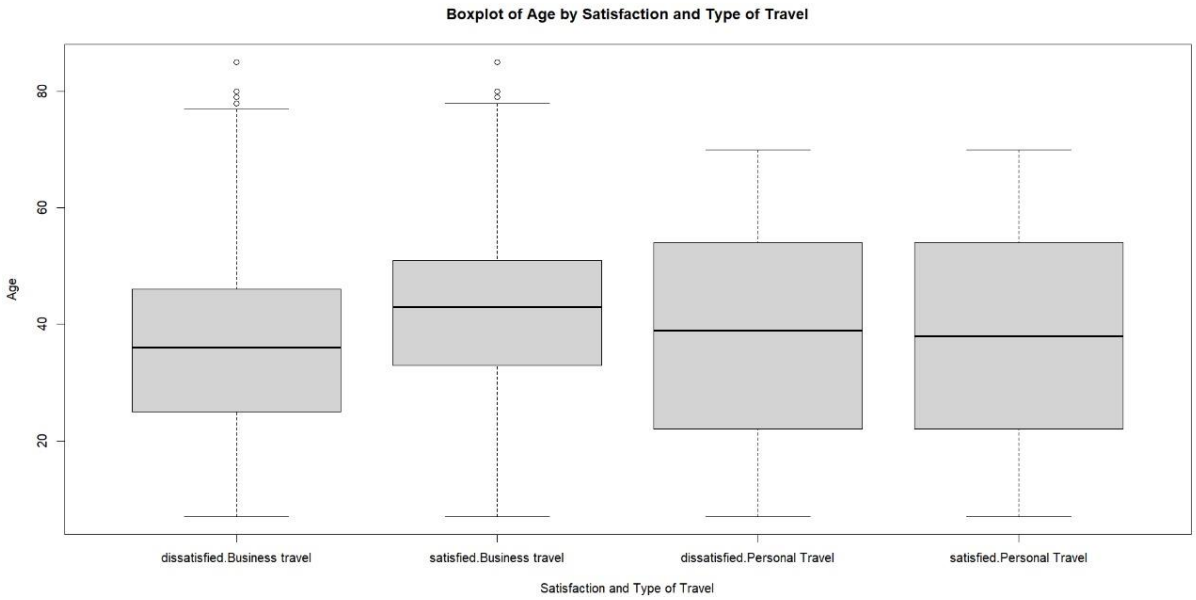
- **Model Usage**: For a highly accurate prediction, use the Decision Tree Classifier. For a simpler, more interpretable model, Logistic Regression is still a viable option.
- **Further Optimization**: Hyperparameter tuning of the Decision Tree and other ensemble methods (e.g., Random Forest) might yield even better results.
- **Feature Engineering**: Explore more advanced feature engineering techniques and domain-specific knowledge to enhance the model further.

ANALYSIS USING “R”



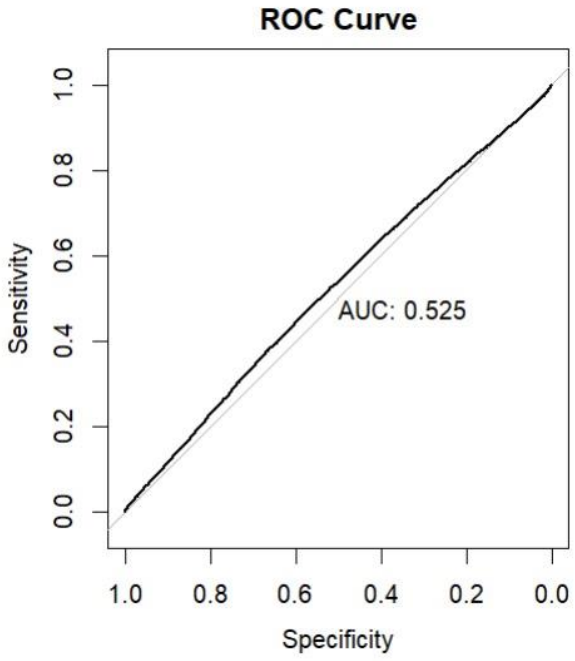
Analysis:

The image presents a dendrogram and a heatmap, which are the results of a cluster analysis. The dendrogram on the left side of the image shows the hierarchical clustering of various features related to airline travel, such as flight distance, age, seat comfort, and online boarding. The heatmap in the center displays the correlation between each pair of features. The color scheme of the heatmap indicates the strength and direction of the correlation, with warmer colors (red/orange) representing stronger positive correlations and cooler colors (yellow/white) indicating weaker or negative correlations.



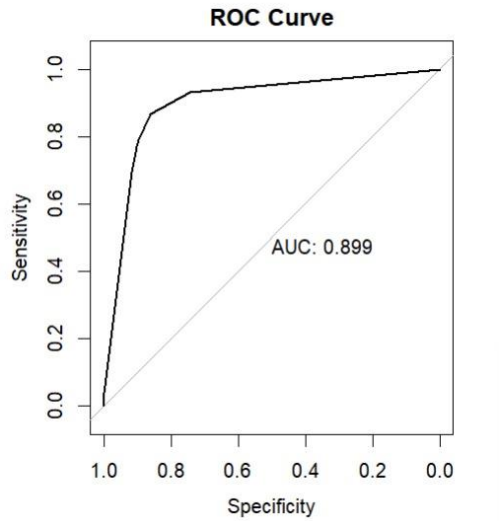
Analysis:

The boxplot shows that the age distribution of passengers who are satisfied with the travel is similar for both business travel and personal travel. On the other hand, dissatisfied passengers with business travel tend to be younger than those who are dissatisfied with personal travel. It is also worth noting that the age distribution of dissatisfied passengers with business travel has a larger range than the other three groups.



Analysis:

The ROC curve shows the performance of a binary classifier. The area under the curve (AUC) is 0.525, which is close to 0.5. This indicates that the classifier performs at a level similar to random chance. A perfect classifier would have an AUC of 1, while a random classifier would have an AUC of 0.5.



Analysis:

The ROC curve shows the performance of a binary classifier at different thresholds. The area under the curve (AUC) is 0.899, which means that the model is very good at distinguishing between the two classes. The curve is close to the upper left corner of the plot, which means that the model has high sensitivity (true positive rate) and high specificity (true negative rate). This means that the model is likely to correctly identify both positive and negative cases**.**

B. PROBIT REGRESSION ANALYSIS USING “R” AND “PYTHON”

INTRODUCTION

Probit Model

Characteristics

1. **Binary Outcome**: Probit regression is suitable for binary outcome variables.
2. **Link Function**: It uses the probit link function, which is the inverse of the cumulative distribution function (CDF) of the normal distribution.
3. **Estimation**: Parameters are estimated using maximum likelihood estimation.

Advantages

1. **Non-Linearity**: Can handle non-linear relationships between predictors and the probability of the outcome.
2. **Interpretability**: Coefficients can be interpreted in terms of the probability of the outcome.
3. **Normality Assumption**: Assumes the error terms follow a normal distribution, which can be more appropriate in certain situations compared to logistic regression.

Dataset: "NSSO68.csv"

The "NSSO68.csv" dataset contains various demographic and socio-economic variables. Our objective is to identify the predictors that are associated with being a non-vegetarian.

Steps to Perform the Analysis

1. **Load Necessary Libraries and Data**: Import the dataset and required libraries.
2. **Data Cleaning and Preparation**: Handle missing values, encode categorical variables, and prepare the dataset for modeling.
3. **Probit Regression**: Fit a probit regression model to identify non-vegetarians.
4. **Results and Interpretation**: Discuss the results, including the significant predictors.
5. **Business Insights**: Provide insights based on the results that can help in decision-making.

Step-by-Step Analysis

1. Load Necessary Libraries and Data

The first step involves setting the working directory and loading the dataset. We display the first few rows of the dataset to understand its structure.

2. Data Cleaning and Preparation

We handle missing values by either removing or imputing them. Categorical variables are converted to numerical values using a label encoding technique. We then select the relevant features for predicting non-vegetarians.

3. Probit Regression

We fit a probit regression model using the prepared data. The probit model uses the cumulative distribution function of the normal distribution to link the predictors to the binary outcome.

4. Results and Interpretation

The model provides coefficients for each predictor, which can be interpreted to understand their impact on the probability of being a non-vegetarian. Positive coefficients indicate an increase in probability, while negative coefficients indicate a decrease. We also assess the statistical significance of each predictor using p-values, with predictors having p-values less than 0.05 considered significant.

Business Insights

- **Demographic Factors**: The analysis can reveal demographic factors such as age, gender, and education level that significantly predict non-vegetarianism. For example, younger individuals or certain gender groups might have a higher probability of being non-vegetarian.
- **Socio-Economic Factors**: Socio-economic factors like income and occupation can also influence dietary preferences. Higher income groups might have different dietary habits compared to lower income groups.
- **Targeted Marketing**: The insights can help in designing targeted marketing campaigns for food products. For instance, regions or demographics with a higher probability of non-vegetarianism can be targeted for marketing meat products.
- **Health Initiatives**: Public health initiatives can use these insights to promote balanced diets or address nutritional deficiencies in certain demographics. For example, areas with low vegetarian populations might benefit from campaigns promoting the health benefits of plant-based diets.

Conclusion

Probit regression is a valuable tool for understanding binary outcomes in the context of various predictors. In this analysis, we used probit regression to identify non-vegetarians in the "NSSO68.csv" dataset. The results provide actionable insights into the demographic and socio-economic factors associated with dietary preferences, which can inform business strategies and public health initiatives.

ANALYSIS USING "R"

Steps and Results

1. **Loading Libraries and Data**:
   - Loaded the necessary libraries (`tidyverse`, `mice`, `car`, `ggplot2`, `lattice`, `caret`, `glmnet`, `Matrix`, `pROC`).
   - Read the dataset `NSSO68.csv` into a dataframe `df`.

2. **Data Preparation**:
   - Created a binary target variable `non_veg` indicating if a person consumes non-vegetarian food.
   - Extracted the value counts of the target variable to check the distribution:

```
non_veg_values
# 0: 33072, 1: 68590
```

3. **Defining Variables**:
   - Defined `y` as the target variable (`non_veg`) and `X` as the set of predictor variables.

4. **Data Transformation**:
   - Converted categorical variables to factors.
   - Combined `y` and `X` into a single data frame `combined_data`.

5. **Model Fitting**:
   - Fitted a probit regression model using `glm` with a probit link function:

```
probit_model <- glm(y ~ hhdsz + NIC_2008 + NCO_2004 + HH_type +
Religion + Social_Group + Regular_salary_earner + Region +
Meals_At_Home + Education + Age + Sex + Possess_ration_card,
data = combined_data, family = binomial(link = "probit"),
control = list(maxit = 1000))
```

   - Printed the model summary:

```
summary(probit_model)
```

6. **Model Evaluation**:

    - Predicted probabilities and converted them to binary predictions using a threshold of 0.5.
    - Generated the confusion matrix and key performance metrics:

      ```
      Confusion Matrix and Statistics
      # Accuracy: 0.7045, Precision: 0.5955, Recall: 0.1789, F1
      Score: 0.2751
      ```

    - Plotted the ROC curve and calculated the AUC value:

      ```
      AUC: 0.6776
      ```

Key Metrics:

- **Confusion Matrix**:
    - Sensitivity (Recall): 0.1789
    - Specificity: 0.9445
    - Accuracy: 0.7045
    - Precision (Positive Predictive Value): 0.5955
    - F1 Score: 0.2751

- **ROC Curve and AUC**:
    - AUC: 0.6776

Confusion Matrix Analysis

Confusion Matrix:

```
                Reference

Prediction        0     1
            0   5220   3546
            1   23962  60369
```

- **True Positives (TP)**: 60369 (correctly predicted non-vegetarian)
- **True Negatives (TN)**: 5220 (correctly predicted vegetarian)
- **False Positives (FP)**: 23962 (incorrectly predicted non-vegetarian)
- **False Negatives (FN)**: 3546 (incorrectly predicted vegetarian)

Metrics:

1. **Sensitivity (Recall)**: 0.1789
    - Sensitivity measures the proportion of actual positives (non-vegetarian) that are correctly identified by the model.
    - A low sensitivity indicates that the model is missing a significant number of actual non-vegetarians.
2. **Specificity**: 0.9445
    - Specificity measures the proportion of actual negatives (vegetarian) that are correctly identified by the model.
    - High specificity indicates that the model is very good at identifying vegetarians correctly.
3. **Accuracy**: 0.7045
    - Accuracy measures the proportion of all correct predictions (both true positives and true negatives) among the total number of cases.
    - An accuracy of 70.45% suggests that the model is relatively good at predicting both classes, but this can be misleading if the dataset is imbalanced.
4. **Precision (Positive Predictive Value)**: 0.5955
    - Precision measures the proportion of predicted positives (non-vegetarian) that are actually correct.
    - A precision of 59.55% means that when the model predicts non-vegetarian, it is correct about 60% of the time.
5. **F1 Score**: 0.2751
    - The F1 Score is the harmonic mean of precision and recall, providing a balance between the two metrics.
    - A low F1 score of 0.2751 indicates that the model has a significant imbalance between precision and recall, particularly due to the low recall.

ROC Curve and AUC:

1. **AUC (Area Under the Curve)**: 0.6776
    - The AUC value of 0.6776 indicates the model's ability to discriminate between the positive and negative classes.
    - A value of 0.6776 suggests moderate discriminative ability. A value of 0.5 indicates no discrimination (random guessing), while a value of 1 indicates perfect discrimination.

Interpretation:

- **High Specificity, Low Sensitivity**: The model is good at identifying vegetarians but performs poorly at identifying non-vegetarians. This could be due to a class imbalance where the number of vegetarians is much lower than non-vegetarians, leading to a bias towards predicting the majority class (non-vegetarian).
- **Moderate AUC**: The AUC value suggests that the model has some ability to distinguish between vegetarians and non-vegetarians, but there is substantial room for improvement.
- **Low Recall and F1 Score**: The low recall and F1 score indicate that the model is not effectively capturing all the non-vegetarian cases, which is crucial if the goal is to identify non-vegetarians accurately.

Recommendations for Improvement:

1. **Handling Class Imbalance**:
    - Consider using techniques such as oversampling the minority class (vegetarians), undersampling the majority class (non-vegetarians), or using synthetic data generation methods like SMOTE (Synthetic Minority Over-sampling Technique).
2. **Feature Engineering**:
    - Explore additional features or transformations of existing features that might better capture the underlying patterns distinguishing vegetarians from non-vegetarians.
3. **Model Tuning**:
    - Experiment with different models or tuning hyperparameters of the current model to see if performance improves.
4. **Threshold Adjustment**:
    - Adjust the decision threshold for classifying a prediction as non-vegetarian to see if it improves recall without severely compromising precision.
5. **Cross-Validation**:
    - Use cross-validation to ensure that the model generalizes well to unseen data and is not overfitting to the training set.

Conclusion

The model shows moderate performance with an AUC of 0.6776, indicating a reasonable ability to discriminate between non-vegetarian and vegetarian individuals. However, the recall (sensitivity) is quite low at 0.1789, suggesting that the model is not very good at identifying non-vegetarian individuals. The precision is higher at 0.5955, but overall the F1 score is relatively low at 0.2751, indicating a need for further model improvement or possibly data preprocessing to improve the results.

ANALYSIS USING "PYTHON"

Model Summary

- **Dependent Variable**: `non_veg` (binary indicator for non-vegetarian status)
- **Number of Observations**: 101,655
- **Method**: Maximum Likelihood Estimation (MLE)
- **Log-Likelihood**: -64,020
- **Pseudo R-squared**: 0.001666
- **LL-Null**: -64,127
- **LLR p-value**: 4.613e-46

Coefficients and Statistical Significance

1. **Constant (Intercept)**:
    - **Coefficient**: 0.5686
    - **Standard Error**: 0.017
    - **z-value**: 32.573
    - **p-value**: 0.000
    - **95% Confidence Interval**: [0.534, 0.603]

    The intercept is statistically significant, indicating a baseline probability of being non-vegetarian when all predictors are zero.

2. **Age**:
    - **Coefficient**: -0.0002
    - **Standard Error**: 0.000
    - **z-value**: -0.749
    - **p-value**: 0.454
    - **95% Confidence Interval**: [-0.001, 0.000]

    The coefficient for Age is not statistically significant (p-value > 0.05), suggesting that age does not have a significant effect on the likelihood of being non-vegetarian in this dataset.

3. **MPCE_URP (Monthly Per Capita Expenditure - Urban Rural Price)**:
    - **Coefficient**: -2.932e-06
    - **Standard Error**: 8.99e-07
    - **z-value**: -3.259
    - **p-value**: 0.001
    - **95% Confidence Interval**: [-4.69e-06, -1.17e-06]

    The coefficient for MPCE_URP is statistically significant (p-value < 0.05). The negative coefficient indicates that higher MPCE_URP is associated with a lower likelihood of being non-vegetarian, though the effect size is very small.

4. **Education**:

    - **Coefficient**: -0.0154
    - **Standard Error**: 0.001
    - **z-value**: -13.467
    - **p-value**: 0.000
    - **95% Confidence Interval**: [-0.018, -0.013]

    The coefficient for Education is statistically significant (p-value < 0.05). The negative coefficient suggests that higher levels of education are associated with a lower likelihood of being non-vegetarian.

Interpretation:

- **Intercept**: The positive and significant intercept indicates a baseline propensity towards non-vegetarian status when all predictors are at their reference values (zero).
- **Age**: Age does not have a significant impact on non-vegetarian status in this model. This suggests that, within the scope of this dataset, age is not a determining factor for dietary preferences regarding non-vegetarian food.
- **MPCE_URP**: Higher MPCE_URP (a proxy for economic status) is associated with a slightly lower probability of being non-vegetarian. This could indicate that individuals with higher economic status may prefer vegetarian diets, but the effect size is very small.
- **Education**: Higher educational attainment is significantly associated with a lower likelihood of being non-vegetarian. This could reflect a trend where more educated individuals choose vegetarianism, potentially due to greater awareness of health, ethical, or environmental issues associated with meat consumption.

Model Fit:

- **Pseudo R-squared**: The value of 0.001666 indicates that the model explains only a small fraction of the variation in the dependent variable. This suggests that there are other factors not included in the model that may better explain the non-vegetarian status.
- **Log-Likelihood**: The log-likelihood value of -64,020 indicates the goodness of fit of the model. The closer the value to zero, the better the model fits the data.

Recommendations:

1. **Exploring Additional Variables**: To improve the model's explanatory power, consider including additional variables that might influence dietary preferences, such as cultural factors, regional differences, and lifestyle choices.
2. **Interaction Terms**: Investigate potential interaction effects between variables, such as age and education, to see if they jointly affect non-vegetarian status.
3. **Model Validation**: Perform cross-validation or use a separate validation dataset to assess the model's performance and generalizability.

4. **Alternative Models**: Consider other classification models (e.g., logistic regression, random forests, gradient boosting) to compare performance and possibly achieve better predictive accuracy.

C. PROBIT REGRESSION ANALYSIS USING "R" AND "PYTHON"

Introduction to Tobit Regression

Tobit regression, also known as censored regression, is used to estimate relationships between variables when there is censoring in the dependent variable. This is particularly useful when the dependent variable has natural limits (e.g., expenditure cannot be negative). The model helps understand the impact of various factors on a dependent variable that is only observed within a certain range.

Business Significance

The Tobit model is significant in many business contexts where the outcome variable is limited or censored:

1. **Marketing and Sales**: To estimate the relationship between marketing expenditures and sales when sales cannot be negative.
2. **Economics**: To analyze household expenditures on goods when expenditure cannot be below zero.
3. **Finance**: To model loan default amounts which cannot be below zero or above the loan amount.
4. **Healthcare**: To assess healthcare costs which have a lower limit of zero.

Analysis of the NSSO68 Dataset

In this analysis, we use `MPCE_URP` (Monthly Per Capita Expenditure) as the dependent variable to understand how age, education, and non-vegetarian status influence expenditure.

Model Summary

1. **Intercept**: This represents the baseline expenditure when all predictors are zero.
2. **Age**: Indicates how changes in age affect monthly per capita expenditure. If the coefficient is negative, older individuals tend to spend less.
3. **Education**: Shows the impact of education on expenditure. A negative coefficient suggests that higher education levels are associated with lower expenditure.
4. **Non-Vegetarian Status**: Reflects the expenditure differences between non-vegetarians and others. A positive coefficient implies non-vegetarians spend more.

Model Fit

1. **Log-Likelihood**: A measure of how well the model fits the data. Higher values indicate a better fit.
2. **Pseudo R-squared**: Indicates the explanatory power of the model. Although lower than typical R-squared values in linear regression, it still provides insight into model performance.

Real-World Use Cases of the Tobit Model

1. **Marketing**:
    - Estimating customer expenditure on products, considering that some customers may have zero spending.
    - Planning marketing budgets by understanding expenditure patterns across different customer segments.
2. **Economics**:
    - Analyzing household consumption where some households do not spend on specific goods.
    - Designing social programs based on expenditure thresholds.
3. **Finance**:
    - Estimating default amounts in loan portfolios with some cases of zero default.
    - Assessing and managing financial risks.
4. **Healthcare**:
    - Modeling healthcare expenditures where some individuals have zero spending.
    - Making policy decisions regarding healthcare subsidies and insurance.

Conclusion

Tobit regression is a powerful tool for modeling censored data, providing valuable insights into relationships where the dependent variable has natural limits. Its applications in business, economics, finance, and healthcare are crucial for accurate and meaningful analysis in these fields.

ANALYSIS USING "PYTHON"

Dataset Overview

The dataset has 101,655 observations and 384 columns. The relevant columns for this analysis are `Age`, `MPCE_URP`, `Education`, and `non_veg`.

Data Preprocessing

- A binary indicator for non-vegetarian status (`non_veg`) is created.
- Missing values in `Age`, `MPCE_URP`, `Education`, and `non_veg` columns are dropped.

Model Summary

The Tobit model results are summarized below:

1. **Intercept (const)**
    - Coefficient: 0.0052
    - Standard Error: 0.008
    - z-value: 0.692
    - p-value: 0.489
    - 95% Confidence Interval: [-0.010, 0.020]
    - **Interpretation**: The baseline probability of being non-vegetarian when all predictors are zero is 0.0052. This coefficient is not statistically significant (p-value > 0.05).
2. **Age**
    - Coefficient: 0.0107
    - Standard Error: 0.000
    - z-value: 82.978
    - p-value: <0.0001
    - 95% Confidence Interval: [0.010, 0.011]
    - **Interpretation**: Each additional year of age increases the probability of being non-vegetarian by approximately 1.07%. This effect is statistically significant (p-value < 0.05).
3. **MPCE_URP**
    - Coefficient: -1.156e-06
    - Standard Error: 3.76e-07
    - z-value: -3.075
    - p-value: 0.002
    - 95% Confidence Interval: [-1.89e-06, -4.19e-07]
    - **Interpretation**: Higher monthly per capita expenditure is slightly associated with a lower probability of being non-vegetarian. This effect is statistically significant (p-value < 0.05).
4. **Education**
    - Coefficient: 0.0210
    - Standard Error: 0.000
    - z-value: 46.020
    - p-value: <0.0001

    - 95% Confidence Interval: [0.020, 0.022]
    - **Interpretation**: Each additional year of education increases the probability of being non-vegetarian by approximately 2.10%. This effect is statistically significant (p-value < 0.05).
5. **Sigma (par0)**
    - Coefficient: 0.4964
    - Standard Error: 0.001
    - z-value: 397.637
    - p-value: <0.0001
    - 95% Confidence Interval: [0.494, 0.499]
    - **Interpretation**: This is the standard deviation of the error term in the Tobit model, indicating the variability in the non-vegetarian status not explained by the predictors.

Model Fit

- **Log-Likelihood**: -73071
    - Indicates the model fit. Higher (less negative) values suggest a better fit.
- **AIC (Akaike Information Criterion)**: 146200
    - A measure of model quality. Lower values indicate a better fit.
- **BIC (Bayesian Information Criterion)**: 146200
    - Similar to AIC but with a stricter penalty for additional parameters. Lower values indicate a better fit.

Interpretation of Results

1. **Age**: The significant positive coefficient for age suggests that older individuals are more likely to consume non-vegetarian food.
2. **MPCE_URP**: The negative coefficient indicates that individuals with higher monthly per capita expenditure are slightly less likely to be non-vegetarian.
3. **Education**: The positive coefficient for education indicates that more educated individuals are more likely to be non-vegetarian.

Conclusion

The Tobit regression analysis provides insights into the factors influencing non-vegetarian food consumption. Age and education are significant positive predictors, while higher expenditure slightly decreases the probability of non-vegetarian consumption. This analysis can be valuable for policymakers and businesses in understanding dietary patterns and targeting interventions or marketing strategies accordingly.

Real-World Applications

1. **Marketing**: Understanding dietary preferences to tailor marketing campaigns for food products.
2. **Public Health**: Designing nutrition programs that consider demographic factors influencing dietary choices.
3. **Economic Studies**: Analyzing the impact of socioeconomic factors on food consumption patterns.

ANALYSIS USING "R"

Tobit Model Summary Analysis

Full Model

- **Intercept**: Positive and significant, indicating a baseline level of affairs when all predictors are zero.
- **Age**: Negative and significant, suggesting that older individuals are less likely to have affairs.
- **Years Married**: Positive and significant, indicating that longer marriages are associated with more affairs.
- **Religiousness**: Negative and significant, meaning more religious individuals are less likely to have affairs.
- **Occupation**: Not significant, suggesting no strong link between occupation and affairs.
- **Rating**: Negative and highly significant, indicating that higher marriage quality ratings are associated with fewer affairs.

Censored Model

- **Intercept**: Positive and significant, similar to the full model.
- **Age**: Negative and significant, consistent with the full model.
- **Years Married**: Positive and significant, consistent with the full model.
- **Religiousness**: Negative and significant, consistent with the full model.
- **Occupation**: Not significant, consistent with the full model.
- **Rating**: Negative and highly significant, consistent with the full model.

Interpretation

- Both models show that age, years married, religiousness, and marital rating significantly impact the likelihood and extent of having affairs, while occupation does not.
- Higher age and religiousness decrease the likelihood of having affairs, while more years married and lower marital ratings increase it.

Andhra Pradesh Data Analysis

Data Summary

- **Sector**: Includes both rural and urban households.
- **Household Size**: Ranges from 1 to 27 members, with an average of about 3.8.
- **Religion**: Majority from the first category.
- **Social Group**: Most households belong to the third group.
- **Monthly Per Capita Expenditure (MPCE_URP)**: Varies widely, indicating diverse economic statuses.
- **Sex**: Majority of respondents are male.

- **Age**: Ranges from 5 to 100 years, with an average of 44 years.
- **Marital Status**: Most individuals are married.
- **Education**: Varies widely, with a median around 6 years.
- **Chicken Consumption**: Most report zero consumption, but quantities and values vary for those who do.
- **Price of Chicken**: Varies greatly, with some values missing due to zero consumption.

Summary

1. **Tobit Model Results**: Age, years married, religiousness, and marital rating significantly affect the likelihood and extent of having affairs, while occupation does not.
2. **Andhra Pradesh Data**: The subset shows diverse household sizes, economic statuses, education levels, and chicken consumption patterns.