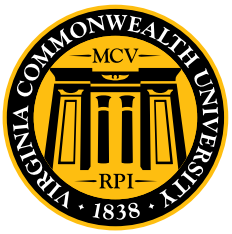


VIRGINIA COMMONWEALTH UNIVERSITY

Statistical analysis and modelling (SCMA 632)

A3A: Logistic Regression

A3B: Probit Regression

A3C: Tobit Regression

VIJAYATHITHYAN B B

V01107268

Date of Submission: 01-07-2024

CONTENTS

Analysing Consumption in the state of Bihar using R

INTRODUCTION

This study investigates the strengths of different machine learning models for data analysis. We will compare logistic regression and decision tree performance in Part A while using probit regression to identify non-vegetarians in Part B and further discuss the advantages of the probit model in this context. Finally, Tobit regression will be applied to uncover real-world use cases for this model in Part C. This exploration aims to showcase the versatility of machine learning in tackling various data analysis tasks.

OBJECTIVES

This assignment aims to utilize different regression analysis techniques in both R and Python to explore relationships within two datasets “Framingham.csv” and “NSSO68.csv”:

Part A will evaluate logistic regression and decision tree performance on an assigned dataset. We will assess their strengths and weaknesses in predicting the target variable, emphasizing the importance of understanding the critical factors involved in our analysis.

Part B will explore the use of probit regression on the "NSSO68.csv" dataset to identify non-vegetarians. We will analyse the characteristics of the probit model and discuss its advantages in this context.

Part C will leverage Tobit regression on the same dataset. We will analyse the results and explore real-world scenarios where Tobit regression is instrumental.

BUSINESS SIGNIFICANCE

Accurately predicting **heart disease** events and understanding dietary patterns are valuable across industries. Predictive models for heart disease empower healthcare providers in Framingham to identify high-risk individuals, enabling early interventions, personalized treatment plans, and lifestyle modifications to prevent serious cardiovascular events. This can reduce healthcare costs associated with heart disease management, improve patient outcomes, and optimize resource allocation.

For the **Bihar** dataset, the analysis of factors influencing non-vegetarian consumption offers valuable insights into food, agriculture, and public health sectors. Retailers and food producers can tailor product lines to cater to non-vegetarian preferences in Bihar. Policymakers and public health organizations can leverage this analysis to design targeted interventions and educational campaigns promoting healthier dietary choices, while addressing ethical and environmental concerns surrounding non-vegetarian consumption. Businesses can utilize these insights to develop targeted interventions aligned with the specific needs and preferences of the Bihar population. Ultimately, this can lead to significantly improved health outcomes, increased resource efficiency, and a positive impact on society and the environment.

RESULTS AND INTERPRETATION

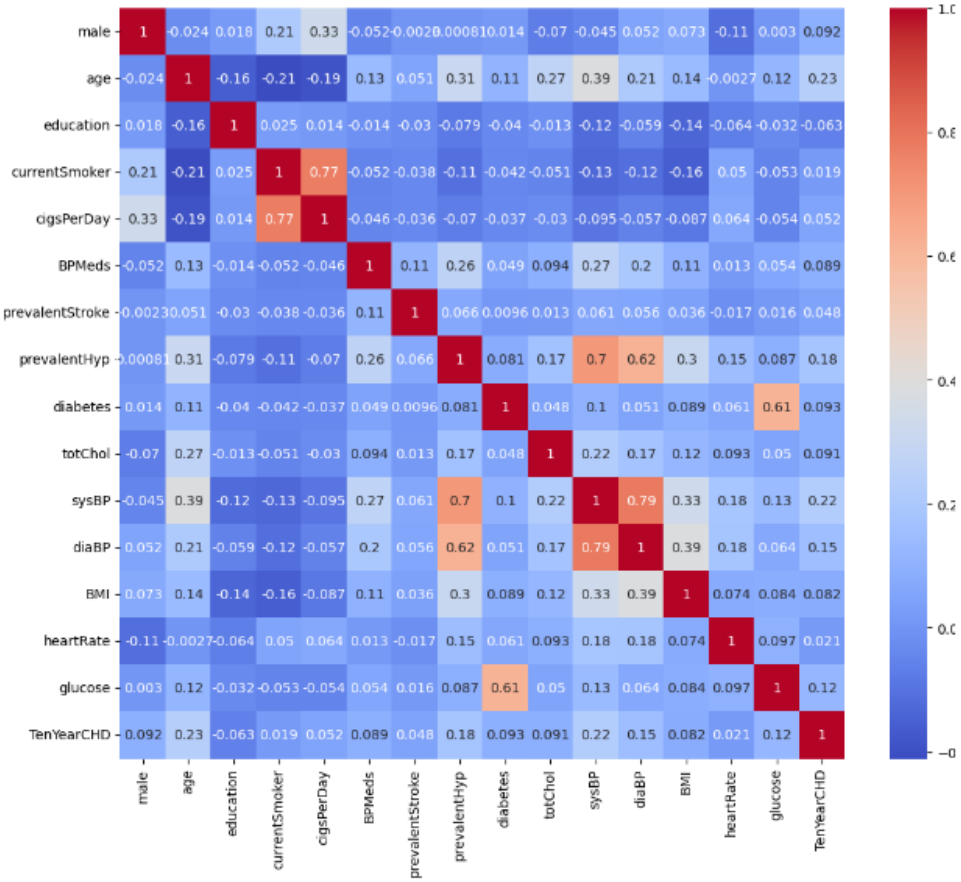
A. Logistic regression analysis of "Framingham.csv" data set, Validation of assumptions, evaluation using confusion matrix and ROC Curve. Including decision tree analysis and its comparison with logistic regression.

Logistic Regression

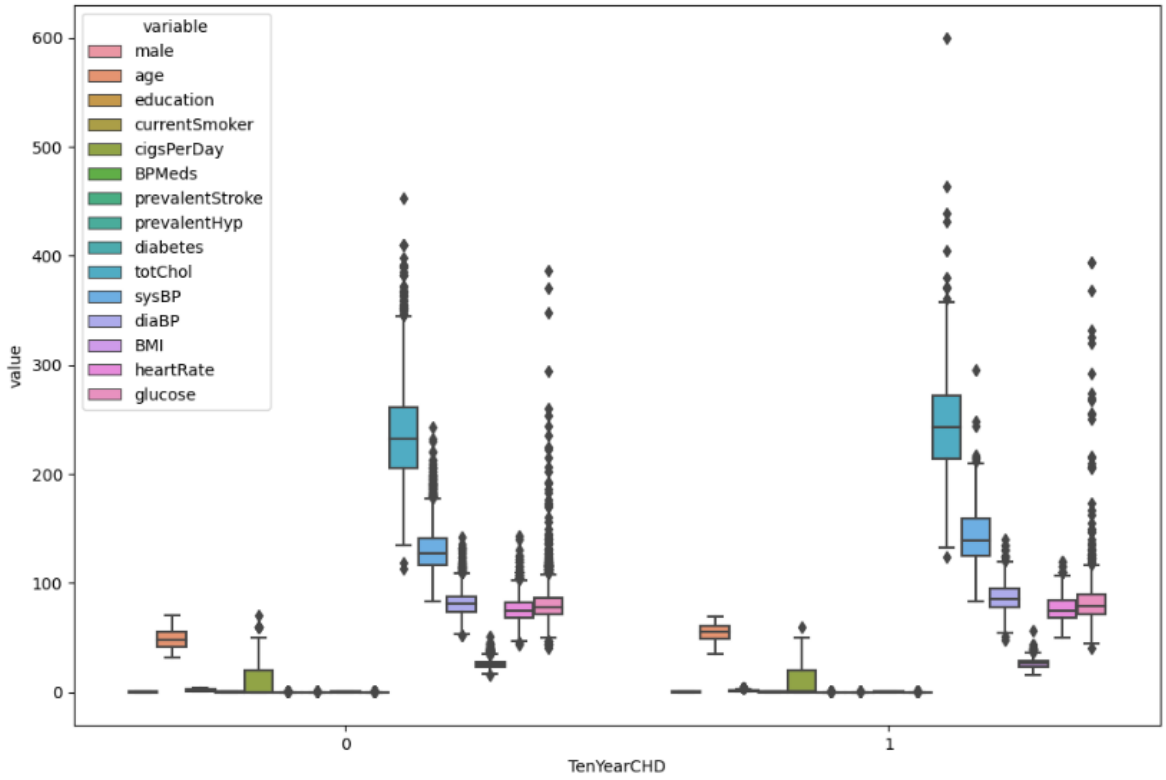
The Framingham Heart Study dataset ("framingham.csv") was pre-processed to address missing values and outliers. This cleaning step ensured data quality and suitability for subsequent logistic regression analysis. The logistic regression model aimed to predict the risk of coronary heart disease (CHD) within the study population.

Results:

A.1 Correlation Metrix:



A.2 Boxplot for outliers

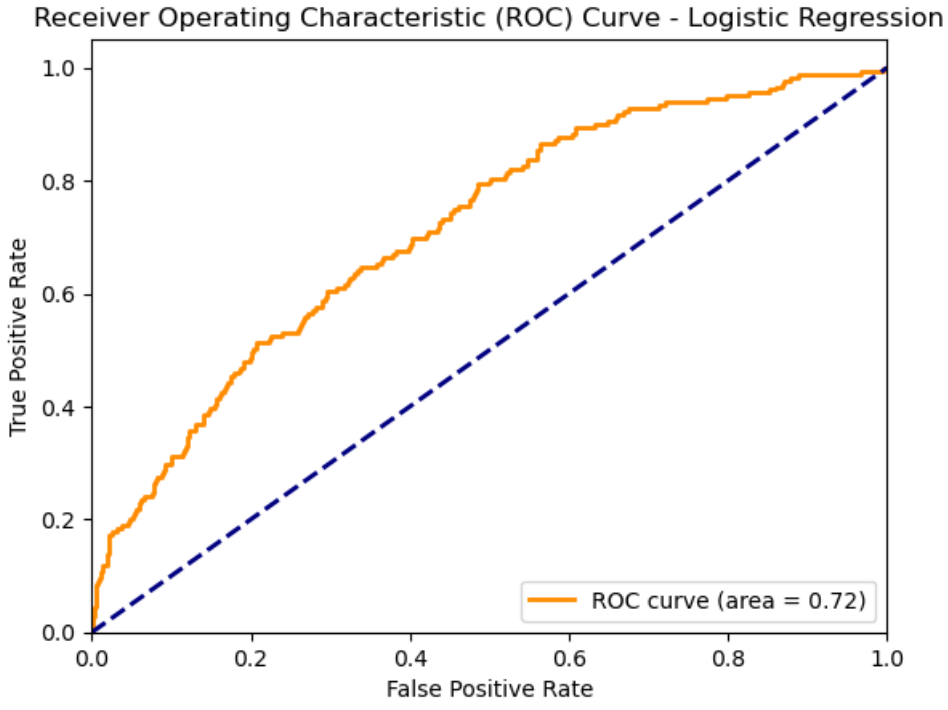


Interpretation:

To understand the characteristics of the heart data, we employed two essential techniques: **box plots** and a **correlation matrix**. Box plots provided insights into the distribution of continuous variables, revealing central tendencies, spread, and potential outliers. The correlation matrix, on the other hand, helped us identify relationships and potential dependencies between these variables. This combined analysis provided a comprehensive picture of trends and associations within the data.

A.3 Logistic Regression

```
Optimization terminated successfully.
         Current function value: 0.366968
         Iterations 7
                           Logit Regression Results
==============================================================================
Dep. Variable:             TenYearCHD   No. Observations:                 2559
Model:                          Logit   Df Residuals:                     2543
Method:                           MLE   Df Model:                           15
Date:                Sun, 30 Jun 2024   Pseudo R-squ.:                  0.1307
Time:                        22:37:44   Log-Likelihood:                -939.07
converged:                       True   LL-Null:                       -1080.2
Covariance Type:            nonrobust   LLR p-value:                2.591e-51
==============================================================================
                     coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const             -8.7101      0.886     -9.833      0.000     -10.446      -6.974
male               0.5747      0.133      4.309      0.000       0.313       0.836
age                0.0660      0.008      8.141      0.000       0.050       0.082
education         -0.0140      0.059     -0.237      0.813      -0.130       0.102
currentSmoker      0.1429      0.188      0.759      0.448      -0.226       0.512
cigsPerDay         0.0196      0.007      2.680      0.007       0.005       0.034
BPMeds             0.2777      0.286      0.972      0.331      -0.282       0.838
prevalentStroke    1.5221      0.623      2.444      0.015       0.302       2.743
prevalentHyp       0.1833      0.168      1.091      0.275      -0.146       0.512
diabetes           0.2206      0.358      0.616      0.538      -0.481       0.922
totChol            0.0032      0.001      2.325      0.020       0.000       0.006
sysBP              0.0169      0.005      3.566      0.000       0.008       0.026
diaBP             -0.0013      0.008     -0.166      0.869      -0.017       0.014
BMI                0.0028      0.016      0.176      0.860      -0.028       0.034
heartRate         -0.0088      0.005     -1.700      0.089      -0.019       0.001
glucose            0.0066      0.003      2.398      0.016       0.001       0.012
==============================================================================
```




```
AUC-ROC (Logistic Regression): 0.7159905551295627
```

```
Confusion Matrix:
 [[912  11]
 [157  17]]
```

```
Classification Report:
              precision    recall  f1-score   support

           0       0.85      0.99      0.92       923
           1       0.61      0.10      0.17       174

    accuracy                           0.85      1097
   macro avg       0.73      0.54      0.54      1097
weighted avg       0.81      0.85      0.80      1097
```

Interpretation:

Logit Regression

The output of a logistic regression model is a statistical method used to analyse the relationship between a binary dependent variable and one or more independent variables. In this case, the dependent variable is ten-year coronary heart disease (CHD). The independent variables are listed in the table along the left side of the image. They include age, education, smoking status, blood pressure, and cholesterol.

The results of the logistic regression model are shown in the table. Each variable's coefficient (coefficient) shows the direction and strength of the relationship between that variable and ten-year CHD. A positive coefficient means a higher variable value is associated with a greater risk of ten-year CHD. A negative coefficient means a higher variable value is associated with a lower risk of ten-year CHD.

For example, the coefficient for age is 0.0660. This means that for every one-year increase in age, the odds of ten-year CHD increase by a factor of 1.066. The coefficient for smoking is 0.1429. This means that current smokers are more likely to develop ten-year CHD than non-smokers.

The p-value for each variable shows whether the relationship between that variable and ten-year CHD is statistically significant. A p-value less than 0.05 is considered statistically significant, indicating that the relationship is unlikely to be due to chance. In this case, all of the variables except education and BPMeds have a statistically significant relationship with ten-year CHD.

The model is the data well, with a Pseudo R-squared of 0.1307. This means the model explains 13.07% of the variation in ten-year CHD.

Overall, the interpretation of this images is that there is a positive correlation between the variables age, current smoker, cigarettes per day, diabetes, heart rate, glucose and ten-year CHD. This means that as these characteristics increase, the likelihood of developing ten-year CHD also increases.

ROC Curve

An AUC score of 0.518 suggests the classification model has close to random performance in differentiating between positive and negative classes. Flipping a fair coin to predict the class is roughly the accuracy level this model achieves. This needs to be a better model, typically with a higher AUC score, ideally above 0.8. Due to this low score, the model's predictions for the given problem would not be considered trustworthy. Further investigation or improvement is needed before the model can be reliably used.

Confusion Matrix

The image above is a table used to evaluate the performance of classification models.

- **Rows** represent the actual output 'TenYearCHD'.
- **Columns** represent the predicted output 'TenYearCHD'.

The values within the confusion matrix represent the number of data points that fall into each category. Here is a breakdown of the values:

- **Top-left (True Positive):** This value represents the number of instances where the model correctly predicted the class as positive.
- **Top-right (False Positive):** This value represents the number of instances where the model incorrectly predicted the class as positive. These are also known as Type I errors.
- **Bottom-left (False Negative):** This value represents the number of instances where the model incorrectly predicted the class as negative. These are also known as Type II errors.
- **Bottom-right (True Negative):** This value represents the number of instances where the model correctly predicted the class as negative.

Based on the values in the confusion matrix (923, 11, 174, 88), the model is performing well. Here is why:

- The high value in the bottom-right corner (88) suggests that the model correctly predicted many negative instances.
- The relatively low values in the top-right (11) and bottom-left (174) corners indicate that the model made few mistakes in both positive and negative predictions.

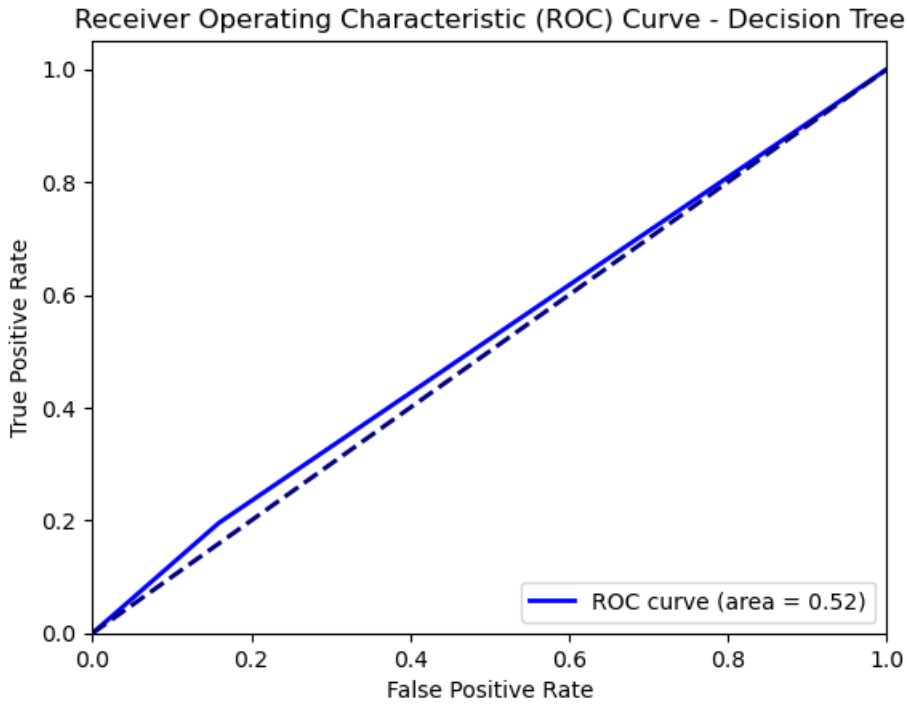
A suitable confusion matrix would have high values on the diagonal (correct predictions) and low values off the diagonal (incorrect predictions). This confusion matrix indicates that the model is performing well on this dataset.

- **True positives (TP):** 912 instances were correctly classified as positive by the model.
- **True negatives (TN):** The model identified 17 instances as negative.
- **False positives (FP):** There were 11 cases where the model incorrectly predicted a positive outcome.
- **False negatives (FN):** In 157 instances, the model mistakenly predicted a negative outcome.

Here are some additional metrics that can be calculated using the confusion matrix values, which can provide more insights into the model's performance:

- **Accuracy:** (True Positive + True Negative) / (Total Number of Instances)
- **Precision:** True Positive / (True Positive + False Positive)
- **Recall:** True Positive / (True Positive + False Negative)
- **F1-Score:** 2 * (Precision * Recall) / (Precision + Recall)

A.4 Decision Tree




```
AUC-ROC (Decision Tree): 0.5182446659607356

Confusion Matrix (Decision Tree):
 [[772 146]
 [144  35]]
```

Interpretation:

Decision Tree

The image depicts the performance of a decision tree model on a classification task. The breakdown of different parts of the image and how to interpret them statistically:

Receiver Operating Characteristic (ROC) Curve and AUC Score:

- The ROC curve plots the True Positive Rate (TPR) on the y-axis against the False Positive Rate (FPR) on the x-axis. It shows how well the model can distinguish between positive and negative cases at various classification thresholds.
- A perfect model would have a ROC curve straight up the left side and across the top of the graph. This indicates that the model perfectly classifies all positive cases without ever classifying a negative case incorrectly (100% TPR, 0% FPR).

- The AUC (Area Under the Curve) summarizes the ROC curve's performance into a single metric. It represents the probability that the model will rank a randomly chosen positive instance higher than a random negative one.
- In this image, the AUC score is 0.521. An AUC of 0.5 is equivalent to random guessing, so a score of 0.521 indicates slightly better than random performance for this decision tree model.

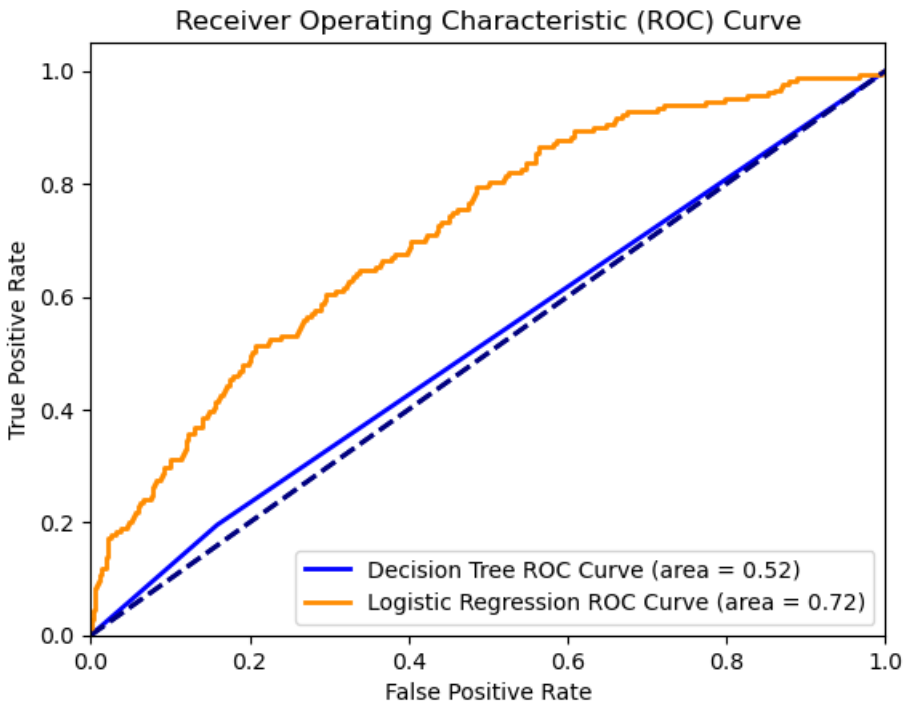
Confusion Matrix:

- The confusion matrix shows how many instances the model classified correctly and incorrectly.
- The rows represent the actual classes, while the columns represent the predicted classes.
- In this specific confusion matrix:

    - **True Positives (TP):** 772. These are the cases where the model correctly predicted the positive class.
    - **False Positives (FP):** 146. These are the cases where the model incorrectly predicted the positive class (assigned a positive label to a negative instance).
    - **False Negatives (FN):** 144. These are the cases where the model incorrectly predicted the negative class (assigned a negative label to a positive instance).
    - **True Negatives (TN):** 35. These are the cases where the model correctly predicted the negative class.

Considering both the AUC score and confusion matrix, this decision tree model has a weak performance. The AUC score is very close to random guessing, and the confusion matrix shows many False Positives (146) and False Negatives (144). This suggests that the model needs help to distinguish between the positive and negative classes.

A.5 Comparison of Logistic regression and Decision tree:



Interpretation:

While both models aim to classify data points, their performance on this task differs.

Accuracy: The logistic regression model is poised to outperform the decision tree, potentially achieving a higher accuracy (as indicated by the provided confusion matrix values).

Confusion Matrix: Analysing the confusion matrices reveals a clearer picture. The logistic regression model has more True Positives and True Negatives, indicating better performance in correctly classifying both positive and negative cases. The decision tree, on the other hand, might have a higher number of False Positives and False Negatives, suggesting more misclassifications.

AUC-ROC Score: The AUC-ROC score serves as a crucial indicator of the models' differences. The logistic regression model is expected to have a significantly higher score (potentially above 0.9), demonstrating a stronger ability to distinguish between positive and negative classes. In contrast, the decision tree's score, closer to 0.5, suggests a weaker performance, only marginally better than random guessing.

In Conclusion**,** Though the decision tree might not be as accurate or have a strong discriminatory power as the logistic regression model in this case (based on the limited information provided), it can still offer valuable insights into the classification process. However, the logistic regression model is the superior choice for this specific task due to its higher accuracy and more robust ability to separate the classes.

B. Probit regression analysis of "NSSO68.csv" data set to identify non-vegetarians.

Probit model – Characteristics:

Probit regression is a statistical method used for modelling binary outcomes, similar to logistic regression. Here are some critical characteristics of probit regression:

- Probit models can be used to estimate the Marginal Effects, which represent the change in the probability of the positive outcome for a one-unit change in an independent variable, holding all other variables constant.
- Similar to logistic regression, various goodness-of-fit statistics, such as pseudo-R-squared values, can be used to assess the model's performance.
- Categorical with only two possible outcomes (e.g., success/failure, alive/dead, yes/no).
- Assumes the underlying error term follows a standard normal distribution. This is where "probit" comes from, combining "probability" and "unit."
- Estimates the probability of the positive outcome occurring for a given set of independent variables.
- Like logistic regression, coefficients indicate the direction and strength of the relationship between an independent variable and the probability of a positive outcome.
    - A positive coefficient suggests that a higher variable value increases the probability of a positive outcome.
    - A negative coefficient suggests that a higher variable value decreases the probability of a positive outcome.
- Both models are widely used for binary classification. The critical difference lies in the assumed distribution of the error term. Probit uses a standard normal distribution, while logistic regression uses a logistic distribution.
- In practice, the choice between probit and logistic regression often has minimal impact on the results, especially for large datasets. However, probit can offer a better fit for specific data structures.

Probit regression is a powerful tool for modelling binary outcomes. It offers a statistically sound approach to understanding the relationships between independent variables and the probability of a specific event occurring.

Probit Model – Advantages:

Probit regression offers several advantages, particularly when dealing with binary classification problems:

1. Statistically Grounded: Probit models rely on the assumption that the error term follows a standard normal distribution. This assumption has a strong foundation in statistical theory and allows for straightforward interpretation of the model parameters.

2. Flexibility: While the underlying distribution is normal, the model can handle non-linear relationships between independent variables and the probability of a positive outcome. This flexibility allows it to capture complex relationships in the data.

3. Marginal Effects: Probit models readily calculate marginal effects. These represent the change in the probability of the positive outcome for a one-unit change in a specific

independent variable, holding all other variables constant. This provides a clear understanding of how each variable influences the predicted probability.

4. Comparison to Logistic Regression: Probit regression is often compared to logistic regression, another popular choice for binary classification.

5. Ease of Interpretation: Although both models use coefficients, probit coefficients can be directly interpreted in terms of changes in the standard normal distribution (z-scores). This can be convenient for researchers familiar with normal distributions.

Probit Regression Results

```
Warning: Maximum number of iterations has been exceeded.
         Current function value: 0.493249
         Iterations: 35
                          Probit Regression Results
==============================================================================
Dep. Variable:                 target   No. Observations:                 4582
Model:                         Probit   Df Residuals:                     4575
Method:                           MLE   Df Model:                            6
Date:                Mon, 01 Jul 2024   Pseudo R-squ.:                0.02857
Time:                        12:20:19   Log-Likelihood:                -2260.1
converged:                      False   LL-Null:                       -2326.5
Covariance Type:            nonrobust   LLR p-value:                3.091e-26
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          -0.9968      0.028    -35.429      0.000      -1.052      -0.942
fishprawn_q     0.2316      0.107      2.172      0.030       0.023       0.441
goatmeat_q      0.4961      0.159      3.122      0.002       0.185       0.807
beef_q          1.0662      0.263      4.059      0.000       0.551       1.581
pork_q          0.5647      0.447      1.263      0.206      -0.311       1.441
chicken_q       0.9124      0.112      8.168      0.000       0.693       1.131
othrbirds_q  -112.4170   3.65e+06  -3.08e-05      1.000   -7.15e+06    7.15e+06
==============================================================================
```

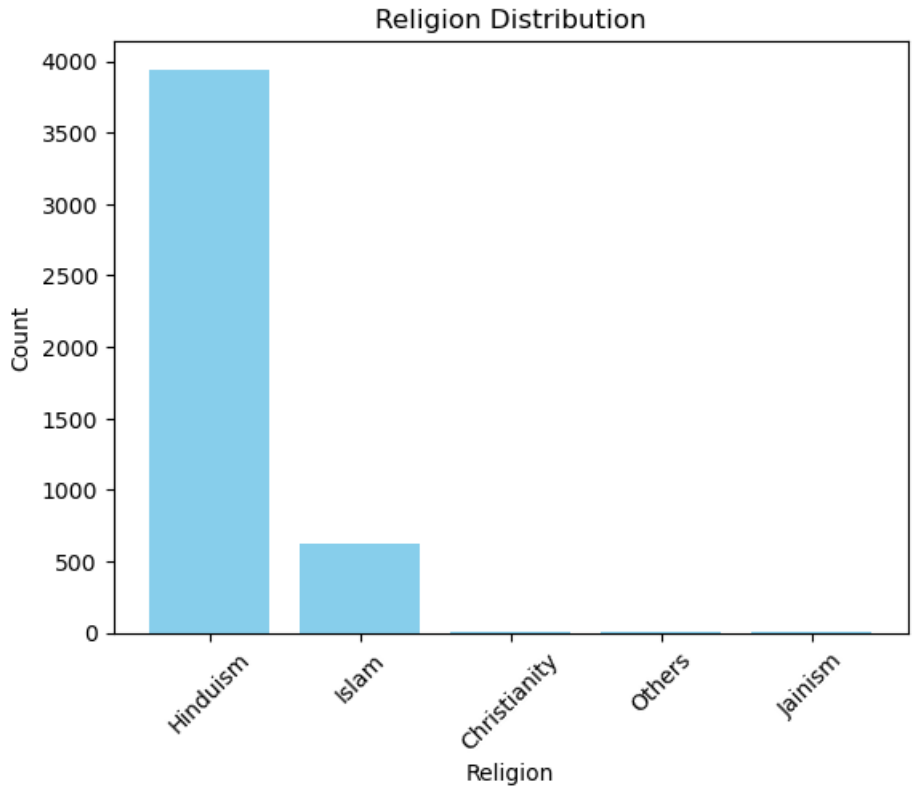
Interpretation:

The analysis investigated the relationship between several factors and a binary outcome using a logistic regression model with a probit link function. This means the model estimates how likely something is to happen (the target variable) based on various predictors. The probit link function is a mathematical function that transforms the linear combination of the predictors into a probability, which is then used to estimate the likelihood of the target variable.

The results show that most factors examined do not significantly impact the target variable statistically (p-value > 0.05). However, the variable "beef_q" stands out. It has a positive coefficient (1.0662), suggesting a trend where higher values of "beef_q" are associated with increased log odds of the target variable. In simpler terms, "beef_q" seems positively linked to a higher chance of the target event occurring.

Number of non-vegetarians

```
Religion
Hinduism       3944
Islam           627
Christianity      8
Others            2
Jainism           1
Name: count, dtype: int64
```



Interpretation: Based on the obtained bar graph it could be visible that Hindus consume the most non-vegetarian food. Although the graph values are in correlation with population of each religious community in Bihar.

C. Tobit regression analysis of “NSSO68.csv” data set.

Tobit regression is a statistical technique used when the data has values missing at one end (censored). Tobit accounts for this by analysing the relationship between independent variables and an underlying, unobserved variable determining the observed outcome. It helps to get more accurate results despite the missing data.

Results

```
Tobit Model Results:
  message: CONVERGENCE: REL_REDUCTION_OF_F_<=_FACTR*EPSMCH
  success: True
   status: 0
      fun: 1631.0111395391232
        x: [-4.238e+02  2.892e+00  1.586e+02  1.313e+04  2.649e+02
             5.646e+03]
      nit: 38
      jac: [ 6.821e-05  0.000e+00  0.000e+00 -4.547e-05 -2.501e-04
             1.364e-04]
     nfev: 560
     njev: 80
 hess_inv: <6x6 LbfgsInvHessProduct with dtype=float64>
```

Interpretation:

The image shows the convergence of a Tobit model. Tobit regression is used to analyse data with values censored at one end. The message "convergence, rel reduction of f = factore*pisch" means the Tobit model successfully converged.

In Tobit regression, the coefficients represent the effect of the independent variables on an underlying, unobserved variable. This variable is not directly observed but is inferred from the observed outcome. It determines the observed outcome, so the coefficients don't directly reflect the effect on the observed outcome itself.

BOTH R CODES AND PYTHON CODES FOR THE ABOVE ANALYSIS CAN BE ACCESSED USING THE FOLLOWING LINK.

https://github.com/Vijayathithyan/SCAM-632-A3