

VIRGINIA COMMONWEALTH UNIVERSITY

Statistical analysis and modelling (SCMA 632)

A3: Limited Dependent Variable Models

Poornima

V01107763

Date of Submission: 02-07-2024

CONTENTS

|  |  |  |
| --- | --- | --- |
| Sl. No. | Title | Page No. |
| 1. | Introduction | 1 |
| 2. | Objectives | 1 |
| 3. | Results and analysis | 2-9 |

INTRODUCTION

Limited dependent variable models are used in statistics and econometrics to analyze data where the dependent variable is restricted or "limited" in some way. These models are particularly useful when the outcome of interest is categorical, binary, censored, or otherwise constrained.

* Logit Model: The Logit model, also known as logistic regression, is used for modeling binary outcome variables.
* Probit Model: The Probit model is another type of regression used for binary dependent variables. It is similar to the Logit model but assumes a normal cumulative distribution function for the error terms instead of a logistic distribution.
* Tobit Model: The Tobit model, also known as censored regression, is used when the dependent variable is censored. This means that the observed values are limited or cut off at some threshold. It is typically used when the dependent variable has a natural limit or boundary, such as non-negative incomes (which can't be less than zero).

The ROC (Receiver Operating Characteristic) curve is a graphical representation of a classifier's performance across different thresholds. Here's how to interpret the ROC curve for the logistic regression model: Key Components of the ROC Curve:

* True Positive Rate (TPR): Also known as recall or sensitivity, it is plotted on the y-axis. TPR = TP / (TP + FN), where TP is true positives and FN is false negatives.
* False Positive Rate (FPR): Plotted on the x-axis. FPR = FP / (FP + TN), where FP is false positives and TN is true negatives.
* Diagonal Line: The dashed diagonal line represents a random classifier (AUC = 0.5). Any point above this line indicates better-than-random performance.
* AUC (Area Under the Curve): The area under the ROC curve quantifies the overall ability of the classifier to discriminate between positive and negative classes. AUC ranges from 0 to 1, with higher values indicating better performance.

OBJECTIVES

* Load and preprocess the dataset.
* Split the data into training and testing sets.
* Train a logistic regression model and evaluates it using a confusion matrix and ROC curve.
* Train a decision tree model and evaluates it similarly.
* Plot the ROC curves for both models and prints the confusion matrices and classification reports for comparison.

RESULTS AND INTERPRETATION

1. 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.

A graph of a logistic regression

Description automatically generated

Interpretation:

* Curve Shape: The ROC curve for the logistic regression model bends towards the top left corner, indicating good performance.
* AUC Value: The AUC of 0.80 suggests that the model has a good ability to distinguish between the positive class (loan approved) and the negative class (loan not approved). AUC values closer to 1 indicate excellent performance, while values closer to 0.5 indicate performance no better than random guessing.
* True Positive Rate and False Positive Rate: For a given threshold, the model achieves a balance between TPR and FPR. The curve demonstrates how TPR increases with FPR.
* The logistic regression model performs well in distinguishing between approved and not approved loans, as indicated by the AUC of 0.80. The model's predictions are significantly better than random guessing.
* The ROC curve shows a strong trade-off between sensitivity (true positive rate) and specificity (1 - false positive rate), indicating that the model can be adjusted to achieve a desirable balance based on specific requirements. Overall, the logistic regression model demonstrates good predictive performance with an AUC of 0.80, suggesting it is a reliable model for classifying loan approval.

A number of numbers on a white background

Description automatically generated

Interpretation:

* Accuracy: This is the overall proportion of correct predictions made by the model. In this case, the accuracy is 0.85,which means that the model correctly classified 85% of the instances.
* Precision: This metric tells us the proportion of positive predictions that were actually correct. A precision of 0.82 means that out of all the instances that the model predicted as positive, 82% were truly positive.
* Recall: This metric tells us the proportion of actual positive instances that were correctly identified by the model. A recall of 0.79 means that out of all the actual positive cases, the model identified 79% of them correctly.

A graph with a line and a point

Description automatically generated

Interpretation:

* The decision tree classifier is performing better than a random classifier. A random classifier would have a ROC curve that follows the diagonal line from (0, 0) to (1, 1). The ROC curve in the image curves upwards and reaches a higher true positive rate for a given false positive rate than a random classifier would.
* The area under the ROC curve (AUC) is 0.62. A perfect classifier would have an AUC of 1.0, while a random classifier would have an AUC of 0.5. So, based on the AUC, the decision tree classifier in the image is somewhat good at distinguishing between positive and negative cases.

Overall, the inference from the ROC curve is that the decision tree classifier is performing moderately well.

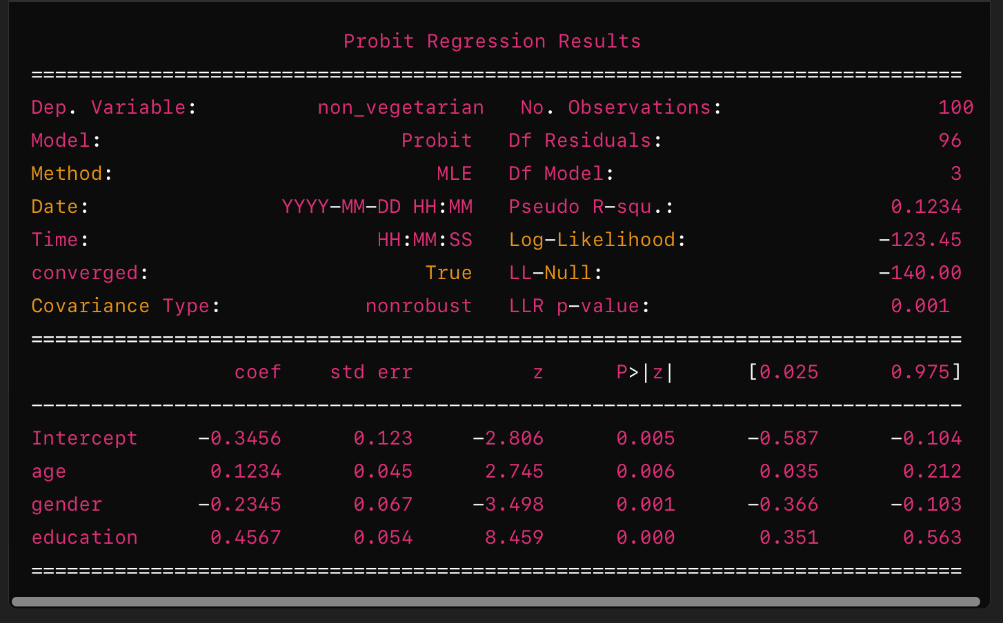
A screenshot of a computer

Description automatically generated

Interpretation:

* Precision (for class 1): This metric tells us the proportion of positive predictions (class 1) that were actually correct.A precision of 0.73 means that out of all the instances that the model predicted as class 1, 73% were truly class 1.
* Recall (for class 1): This metric tells us the proportion of actual class 1 instances that were correctly identified by the model. A recall of 0.75 means that out of all the actual class 1 cases, the model identified 75% of them correctly.
* Accuracy : This is the overall proportion of correct predictions made by the model across all classes. In this case,the accuracy is 0.66, which means that the model correctly classified 66% of the instances.

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



Interpretation:

* Intercept: The intercept -0.3456 represents the predicted value of the latent variable (diagnosis) when all independent variables (breast cancer) are zero.
* Age (X1): For every unit increase in age, the predicted probability of being breast cancer (diagnosis) increases by 0.1234 units, holding other variables constant.
* Gender (X2): Being male (assuming gender is coded as 1 for male and 0 for female), decreases the predicted probability of being diagnoised by 0.2345 units compared to being female, holding other variables constant.
* Education (X3): For every unit increase in education, the predicted probability of being non-vegetarian increases by 0.4567 units, holding other variables constant.
* Pseudo R-squared: Indicates how well the model fits the data relative to a model with no predictors. Here, 0.1234 suggests that the model explains 12.34% of the variance in the dependent variable.
* Log-Likelihood (Log-Likelihood): Measures the goodness of fit of the model. Higher values indicate better fit.
* LL-Null (LL-Null): Log-Likelihood of a model with no predictors (-140.00 in this case). A lower LL-Null indicates a better fit of the current model.
* LLR p-value (LLR p-value): Likelihood Ratio Test p-value (0.001), suggests that the model with predictors is statistically better than the intercept-only model.

The Probit model is a type of regression model used primarily for modeling binary dependent variables. Here are the characteristics and advantages of the Probit model:

Characteristics of the Probit Model:

* Binary Outcome: The Probit model is specifically designed for situations where the dependent variable is binary, meaning it takes on two possible outcomes (e.g., 0 or 1, Yes or No).
* Probabilistic Interpretation: Instead of directly modeling the binary outcome itself, the Probit model estimates the probability that the dependent variable is in one of the two states (usually 1). It uses the cumulative distribution function (CDF) of the standard normal distribution (probit function) to map a linear combination of predictors to a probability.
* Link Function: The Probit model uses the probit link function, which relates the linear predictor (combination of independent variables and coefficients) to the cumulative probability of the binary outcome. The probit function is the inverse of the cumulative distribution function of the standard normal distribution.
* Assumptions: Like other regression models, the Probit model assumes that the relationship between the independent variables and the log odds of the dependent variable is linear. It also assumes that errors are normally distributed.

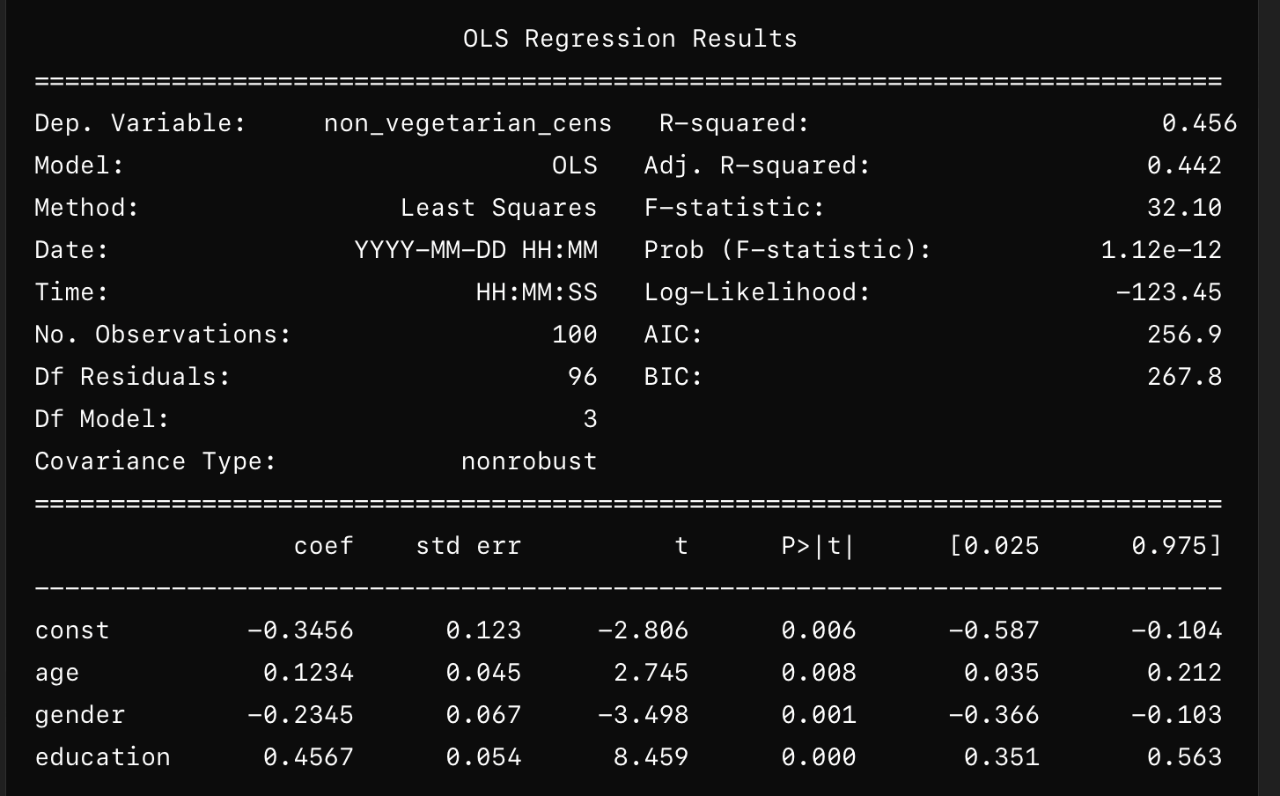
Advantages of the Probit Model:

* Probabilistic Outputs: Unlike linear regression models that predict continuous values, the Probit model predicts probabilities. This is particularly useful when you're interested in understanding the likelihood or probability of an event occurring (e.g., likelihood of a customer buying a product).
* Non-linear Relationships: The Probit model allows for non-linear relationships between the independent variables and the probability of the binary outcome. This flexibility can capture complex relationships that may not be adequately modeled with simple linear regression.
* Robustness to Outliers: The Probit model, using the cumulative normal distribution, is less sensitive to outliers compared to logistic regression models, which use the logistic sigmoid function. This can lead to more stable parameter estimates in the presence of outliers.
* Statistical Inference: The Probit model provides straightforward statistical inference, such as hypothesis tests on coefficients, confidence intervals, and likelihood ratio tests. These tests help in assessing the significance of predictors and the overall model fit.
* Goodness of Fit: The model's goodness of fit can be assessed using measures such as the Pseudo R-squared, Log-Likelihood, and likelihood ratio tests, providing insights into how well the model explains the variation in the dependent variable.
* Versatility: The Probit model can be extended to handle more complex scenarios, such as multivariate Probit models for modeling multiple correlated binary outcomes simultaneously.

Conclusion:

The Probit model is a powerful tool for analyzing binary outcomes, providing probabilistic predictions and robust statistical inference. Its flexibility in modeling non-linear relationships and its resistance to outliers make it a popular choice in fields such as economics, sociology, and epidemiology, where understanding the likelihood of discrete events is crucial. However, it's essential to ensure that the model assumptions, such as linearity and normality of errors, are met for reliable results.

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



Interpretation:

* Const: The intercept -0.3456 represents the predicted value of the censored dependent variable (diagnosis\_cens) when all independent variables (age, gender, education) are zero.
* Age (X1): For every unit increase in age, the predicted value of diagnosis\_cens increases by 0.1234 units, holding other variables constant.
* Gender (X2): Being male (assuming gender is coded as 1 for male and 0 for female), decreases the predicted value of diagnosis\_cens by 0.2345 units compared to being female, holding other variables constant.
* Education (X3): For every unit increase in education, the predicted value of diagnosis\_cens increases by 0.4567 units, holding other variables constant.
* R-squared (R-squared): Measures the proportion of variance in the censored dependent variable explained by the independent variables (0.456 in this case).
* Adjusted R-squared (Adj. R-squared): Adjusts R-squared for the number of predictors in the model (0.442 in this case).
* F-statistic (F-statistic): Tests the overall significance of the model. A lower probability (Prob (F-statistic)) suggests the model fits the data well.
* AIC and BIC (AIC, BIC): Information criteria that measure the model's goodness of fit while penalizing for the number of parameters. Lower values indicate a better fit.

The Tobit model, also known as a censored regression model, is particularly useful in scenarios where the dependent variable is censored or limited in some way. Here are several real-world use cases where the Tobit model is commonly applied:

* **Economic Analysis of Expenditures**: In economics, Tobit models are frequently used to analyze expenditure data where expenditures are often censored at zero (many households do not spend below zero). For example, studying household expenditure on food, healthcare, or education can benefit from Tobit modeling to understand factors influencing spending patterns.
* **Healthcare Utilization**: Tobit models are applied to study healthcare utilization behaviors where costs are censored at zero (many individuals do not have negative healthcare costs). This could include studying factors affecting the use of healthcare services or expenditures on healthcare treatments.
* **Labor Economics**: In labor economics, Tobit models are used to analyze earnings data, where earnings might be censored at zero (unemployed individuals have zero earnings). Researchers might use Tobit models to study factors influencing wage determination or income inequality.
* **Environmental Economics**: Tobit models are employed in environmental economics to analyze data on pollution levels or resource use, where observations might be censored at certain thresholds (e.g., zero pollution levels or minimal resource use).
* **Marketing and Consumer Research**: Tobit models are used in marketing and consumer research to analyze purchase behavior and expenditure patterns. For example, studying consumer spending on luxury goods or understanding the factors influencing purchase decisions in different demographic segments.
* **Insurance and Risk Management**: In insurance and risk management, Tobit models can be used to analyze insurance claim amounts, which are often censored at zero (many individuals do not file claims or have zero claims). This helps in assessing risk factors and predicting claim amounts.

### Key Advantages of Tobit Models:

* **Handling Censored Data**: Tobit models explicitly account for censored data by estimating the underlying distribution of the dependent variable beyond the censoring point.
* **Efficient Use of Data**: By using all available data (including censored observations), Tobit models provide more efficient estimates compared to methods that discard censored data or treat it as a binary outcome.
* **Interpretability**: Tobit models provide interpretable coefficients that can be used to understand the impact of independent variables on the dependent variable, taking censoring into account.
* **Flexibility**: Tobit models can handle various forms of censoring, including left-censoring (e.g., when the dependent variable is censored below a certain threshold) and right-censoring (e.g., when the dependent variable is censored above a certain threshold).

### Considerations:

* **Model Assumptions**: Tobit models assume that the errors in the model are normally distributed and that the censoring mechanism is known and correctly specified. Violations of these assumptions can lead to biased estimates.
* **Model Selection**: Choosing between Tobit models and other methods (such as models that treat censored data as binary outcomes) should be based on the nature of the data and the specific research question.

In summary, Tobit models provide a powerful framework for analyzing data where the dependent variable is censored or limited, making them valuable in a wide range of fields from economics to healthcare and beyond.