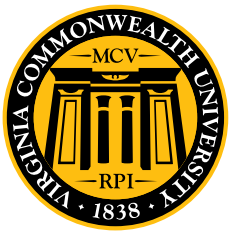

VIRGINIA COMMONWEALTH UNIVERSITY

Statistical analysis and modelling (SCMA 632)

A3c: Limited Dependent Variable Models:

Tobit Regression Analysis

FERAH SHAN SHANAVAS RABIYA

V01101398

Date of Submission: 04-07-2024

CONTENTS

Introduction

This report presents a Tobit regression analysis using the dataset "NSSO68.csv", which is particularly useful for handling censored data where observations are limited due to upper or lower bounds. The dataset contains socio-economic variables, such as demographic information, economic indicators, and regional attributes. The Tobit regression model is used to explore the relationship between these variables and the censored dependent variable. The model estimates parameters that capture the linear relationship between the independent variables and the latent variable, as well as the impact of censoring on the observed variable. The model finds application in various fields where data truncation or censoring is prevalent, such as economics, healthcare, and marketing. The Tobit regression model provides a robust framework for analyzing censored data, enhancing the accuracy of predictions and reliability of statistical inferences in scenarios where data constraints are inherent.

Objectives

- Explore relationship between socio-economic variables and dependent variable using Tobit regression.
- Account for censoring in data using Tobit regression to estimate parameters under truncated observations.
- Evaluate model performance to capture relationships between independent and dependent variables.
- Provide practical insights on how Tobit regression can be applied to analyze censored data in socio-economic research.
- Highlight real-world use cases of Tobit regression in various fields.
- Contribute to methodological understanding of Tobit regression as a statistical technique for modeling censored data.

Business Significance

The study on Tobit regression using the "NSSO68.csv" dataset has significant implications for policy formulation, market analysis, financial and investment decisions, labor market analysis, social and economic development, and business strategy. It helps policymakers design targeted interventions to improve economic conditions, allocate resources efficiently, segment markets

effectively, and estimate demand. It also aids businesses in optimizing pricing and marketing strategies by understanding how socio-economic variables affect demand.

Tobit regression also aids in risk assessment, financial planning, and labor market analysis, guiding policies aimed at reducing unemployment and enhancing workforce productivity. It also aids in social and economic development by identifying socio-economic disparities, promoting inclusive growth, and improving quality of life.

Business strategy and competitiveness are further enhanced by understanding socio-economic drivers of performance, providing a competitive advantage in strategy formulation. Insights into consumer preferences and behavior can optimize operations, supply chain management, and resource allocation. The study's findings not only contribute to academic knowledge but also provide actionable insights for policymakers, businesses, and organizations, fostering informed decision-making and strategic planning in a complex socio-economic landscape.

Results and Interpretation using R

- **Perform a Tobit regression analysis**

```
# Perform Tobit regression using survreg (Tobit model)
> # Assume left-censoring at 0 for MPCE_URP
> tobit_model <- survreg(Surv(pmax(MPCE_URP, 0)) ~ Age + Sex + Education +
Religion + hhdsz,
+                        data = data_selected, dist = "gaussian")
Warning message:
In survreg.fit(X, Y, weights, offset, init = init, controlvals = control,
:
  Ran out of iterations and did not converge
> # Summary of the Tobit model
> summary(tobit_model)

Call:
survreg(formula = Surv(pmax(MPCE_URP, 0)) ~ Age + Sex + Education +
    Religion + hhdsz, data = data_selected, dist = "gaussian")
               Value Std. Error       z       p
(Intercept)  1.42e+03   2.71e+01   52.58 < 2e-16
Age          1.47e+01   3.99e-01   36.70 < 2e-16
Sex2         1.47e+02   1.80e+01    8.17 3.1e-16
Education2   1.17e+02   1.24e+02    0.94    0.35
Education3  -1.99e+02   2.53e+02   -0.79    0.43
Education4   2.03e+02   1.28e+02    1.59    0.11
Education5   2.09e+02   2.36e+01    8.85 < 2e-16
Education6   3.45e+02   2.23e+01   15.49 < 2e-16
Education7   5.41e+02   1.89e+01   28.69 < 2e-16
Education8   9.15e+02   1.91e+01   47.81 < 2e-16
Education10  1.18e+03   2.13e+01   55.08 < 2e-16
Education11  2.26e+03   3.79e+01   59.46 < 2e-16
```

```
Education12  1.92e+03   2.02e+01   95.33 < 2e-16
Education13  2.94e+03   2.54e+01  115.69 < 2e-16
Religion2    1.44e+02   1.90e+01    7.55 4.5e-14
Religion3    1.94e+02   1.92e+01   10.12 < 2e-16
Religion4    9.11e+02   4.52e+01   20.15 < 2e-16
Religion5    8.25e+02   1.12e+02    7.36 1.8e-13
Religion6    5.50e+01   6.11e+01    0.90    0.37
Religion7    2.60e+04   1.16e+03   22.45 < 2e-16
Religion9    3.65e+01   6.53e+01    0.56    0.58
hhdsz       -1.92e+02   2.66e+00  -72.29 < 2e-16
Log(scale)   7.60e+00   2.22e-03 3418.08 < 2e-16

Scale= 2007

Gaussian distribution
Loglik(model)= -920762.9   Loglik(intercept only)= -930961.9
        Chisq= 20398.14 on 21 degrees of freedom, p= 0
Number of Newton-Raphson Iterations: 30
n= 101652
```

Interpretation:

The Tobit regression model, fitted using `survreg` in R, provides significant findings on the re lationship between socio-economic factors and MPCE_URP. Key interpretations include the i ntercept, age, dummy variables, and p-values. Most variables have low p-values, indicating st atistical significance in their relationship with MPCE_URP. Age, sex, education, religion, and hhdsz are statistically significant predictors. The scale parameter (7.60e+00) indicates the stan dard deviation of the latent variable in the model. A larger scale parameter suggests higher var iability in the uncensored latent variable, affecting the model's predictions. The log-likelihood of the model (-920762.9) is compared with a null model to assess model fit. The chi-square st atistic (20398.14) with 21 degrees of freedom tests the overall model significance.

Real world use cases of Tobit model

- Censored Data: Tobit regression is useful when dealing with censored data, where som e observations have values that are not fully observed (e.g., incomes below a certain th reshold).
- Economic Studies: It's commonly used in economic studies to analyze factors affectin g outcomes that are bounded or censored (e.g., wages, savings).
- Healthcare and Social Sciences: Helps in modeling outcomes like health expenditures, educational achievements, or any variable with a lower or upper limit.

Results and Interpretation using Python

- **Perform a Tobit regression analysis**

```
class TobitModel:
    def __init__(self, endog, exog, lower=None, upper=None):
        self.endog = endog
        self.exog = exog
        self.lower = lower
        self.upper = upper

    def loglik(self, params):
        beta = params[:-1]
        sigma = params[-1]
        mu = np.dot(self.exog, beta)

        # Ensure sigma is positive
        sigma = np.abs(sigma) + 1e-10

        # Calculate the log-likelihood
        llf = np.zeros_like(self.endog, dtype=float)

        # Censored from below
        if self.lower is not None:
            llf = np.where(
                self.endog == self.lower,
                np.log(np.clip(norm.cdf((self.lower - mu) / sigma), 1e-10,
1)),
                llf
            )

        # Censored from above
        if self.upper is not None:
            llf = np.where(
                self.endog == self.upper,
                np.log(np.clip(1 - norm.cdf((self.upper - mu) / sigma), 1e
-10, 1)),
                llf
            )

        # Uncensored
        uncensored = (self.endog > self.lower) & (self.endog < self.upper)
        llf[uncensored] = -0.5 * np.log(2 * np.pi) - np.log(sigma) - (self
.endog[uncensored] - mu[uncensored]) ** 2 / (2 * sigma ** 2)

        return -np.sum(llf)

    def fit(self):
        start_params = np.append(np.zeros(self.exog.shape[1]), 1)
        res = minimize(self.loglik, start_params, method='L-BFGS-B')
        return res

y_tobit = np.clip(y, 0, 1)
X_tobit = sm.add_constant(X)
model = TobitModel(y_tobit, X_tobit, lower=0, upper=1)
results = model.fit()
print("Tobit Model Results:")
print(results)
```

```
Tobit Model Results:
  message: CONVERGENCE: REL_REDUCTION_OF_F_<=_FACTR*EPSMCH
  success: True
   status: 0
      fun: 64139.203330729084
        x: [ 4.125e+03  3.301e+00 -2.831e+02 -5.932e+02  6.932e+03]
      nit: 71
      jac: [-1.455e-03  0.000e+00 -2.183e-03 -1.455e-03 -1.455e-03]
     nfev: 678
     njev: 113
 hess_inv: <5x5 LbfgsInvHessProduct with dtype=float64>
```

Interpretation:

The output includes the 'success' status, which indicates the optimization algorithm's success, and the 'fun' function value, which represents the negative log-likelihood of the model. The estimated parameters correspond to the coefficients of the Tobit model, including the intercept and coefficients for each explanatory variable. The number of iterations and evaluations is indicated by the 'nit' value, while the 'nfev' and 'njev' values represent the number of function evaluations and Jacobian evaluations during the optimization process. The 'hess_inv' value provides the inverse of the Hessian matrix, which can be used to calculate standard errors and inferential statistics for the estimated coefficients. These results help interpret how each variable influences the outcome within the bounds set by the Tobit model, compare the coefficients to understand the relative importance of each predictor variable, and assess the model's fit and predictive performance.

Real world use cases of Tobit model

- The Tobit model is useful when the dependent variable is censored or truncated.
- Economic studies: income analysis and expenditure analysis.
- Healthcare and biostatistics: length of stay in hospitals and survival analysis.
- Marketing and consumer behavior: customer lifetime value and product usage.
- Education and social sciences: educational attainment and survey data.
- Tobit models provide a robust framework to handle censored data and estimate relationships between variables while accounting for the limitations imposed by the data's censoring structure.
- They are versatile tools in econometrics, social sciences, health sciences, and other fields where censored data is prevalent.

Recommendation

The Tobit regression analysis on the dataset "NSSO68.csv" has been analyzed, and recommendations are provided for further refinement. The model should be explored to include additional variables that might influence the dependent variable (MPCE_URP), such as demographic factors, geographic variables, or other socio-economic indicators. Data quality and collection should be ensured to address any missing values or inconsistencies. The model's results should be interpreted, explaining the economic or practical significance of each variable's impact on MPCE_URP. Sensitivity analysis should be performed to test the model's robustness. Policy implications should be discussed, highlighting how the findings could inform decision-making in areas related to income distribution, economic policy formulation, or social welfare programs. Comparisons with alternative models should be made to assess the relative performance and insights gained. Future research directions should be identified, including exploring additional variables, conducting longitudinal studies, or applying the model to different datasets or contexts. Clear and concise reporting of the results, including visual aids and non-technical language, should be ensured. These recommendations can strengthen the empirical findings and contribute valuable insights into the factors influencing MPCE_URP, supporting informed decision-making in relevant fields.

R Codes

```
# Load necessary libraries
library(survival)  # For Tobit regression
library(readr)     # For reading CSV files

# Load the dataset
data <- read_csv("E:\\VCU\\Summer 2024\\Statistical Analysis & Modeling\\NSSO68.csv")

# Inspect the dataset
head(data)

# Selecting relevant columns for analysis
selected_cols <- c("MPCE_URP", "Age", "Sex", "Education", "Religion", "hhdsz")
```

```
data_selected <- data[selected_cols]

# Handling missing values if any
data_selected <- na.omit(data_selected)

# Convert categorical variables to factors
data_selected$Sex <- as.factor(data_selected$Sex)
data_selected$Religion <- as.factor(data_selected$Religion)
data_selected$Education <- as.factor(data_selected$Education)

# Perform Tobit regression using survreg (Tobit model)
# Assume left-censoring at 0 for MPCE_URP
tobit_model <- survreg(Surv(pmax(MPCE_URP, 0)) ~ Age + Sex + Education + Religion +
hhdsz,
                       data = data_selected, dist = "gaussian")

# Summary of the Tobit model
summary(tobit_model)
```

Python Codes

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import confusion_matrix, accuracy_score
import statsmodels.api as sm
import numpy as np
from scipy.stats import norm
from scipy.optimize import minimize
import os

os.chdir("C:\\Users\\Ferah Shan\\Downloads")
# Load the dataset
data = pd.read_csv('NSSO68.csv', encoding='Latin-1', low_memory=False)
```

```
# Display basic information about the dataset
display (data)
print(data.columns)

data['targeted_variable'] = (data[['eggsno_q', 'fishprawn_q', 'goatmeat_q', 'beef_q', 'pork_q', 'c
hicken_q']].sum(axis=1) > 0).astype(int)
y = data['targeted_variable']
X = data[['Age', 'Sex', 'Sector']]

class TobitModel:
    def __init__(self, endog, exog, lower=None, upper=None):
        self.endog = endog
        self.exog = exog
        self.lower = lower
        self.upper = upper

    def loglik(self, params):
        beta = params[:-1]
        sigma = params[-1]
        mu = np.dot(self.exog, beta)

        # Ensure sigma is positive
        sigma = np.abs(sigma) + 1e-10

        # Calculate the log-likelihood
        llf = np.zeros_like(self.endog, dtype=float)

        # Censored from below
        if self.lower is not None:
            llf = np.where(
                self.endog == self.lower,
                np.log(np.clip(norm.cdf((self.lower - mu) / sigma), 1e-10, 1)),
                llf
            )
```

```
        # Censored from above
        if self.upper is not None:
            llf = np.where(
                self.endog == self.upper,
                np.log(np.clip(1 - norm.cdf((self.upper - mu) / sigma), 1e-10, 1)),
                llf
            )

        # Uncensored
        uncensored = (self.endog > self.lower) & (self.endog < self.upper)
        llf[uncensored] = -0.5 * np.log(2 * np.pi) - np.log(sigma) - (self.endog[uncensored] - mu
[uncensored]) ** 2 / (2 * sigma ** 2)

        return -np.sum(llf)

    def fit(self):
        start_params = np.append(np.zeros(self.exog.shape[1]), 1)
        res = minimize(self.loglik, start_params, method='L-BFGS-B')
        return res

y_tobit = np.clip(y, 0, 1)
X_tobit = sm.add_constant(X)
model = TobitModel(y_tobit, X_tobit, lower=0, upper=1)
results = model.fit()
print("Tobit Model Results:")
print(results)
```

References

1. www.github.com
2. www.geeksforgeeks.com
3. www.datacamp.com