VIRGINIA COMMONWEALTH UNIVERSITY

Statistical analysis and modelling (SCMA 632)

A3c: Limited Dependent Variable Models:

Tobit Regression Analysis

SAYA SANTHOSH

V01101901

Date of Submission: 07-07-2024

CONTENTS

Introduction

The dataset "NSSO68.csv" is used in this report's Tobit regression analysis, which is especially helpful for managing censored data—data whose observations are restricted because of upper or lower bounds. Socioeconomic variables including demographic data, economic indicators, and geographical characteristics are included in the dataset. The link between these variables and the censored dependent variable is investigated using the Tobit regression model. The impact of censoring on the observed variable and the linear relationship between the independent variables and the latent variable are both captured by the model's estimated parameters. The concept finds use in a number of industries, including marketing, healthcare, and economics, where data censorship or truncation is common. In situations where data limitations are inherent, the Tobit regression model offers a strong framework for analyzing censored data, improving the precision of predictions and dependability of statistical judgments.

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 :

There are important implications for policy formulation, market analysis, financial and investment decisions, labor market analysis, social and economic development, and corporate strategy from the Tobit regression study conducted with the "NSSO68.csv" dataset. It assists

policymakers in creating focused interventions to enhance economic circumstances, distribute resources effectively, divide markets profitably, and gauge demand. By comprehending how socioeconomic factors impact demand, it also helps organizations optimize pricing and marketing techniques.

In addition, tobit regression supports labor market analysis, risk assessment, and financial planning, directing policies meant to lower unemployment and raise worker productivity. Furthermore, it facilitates social and economic advancement by highlighting socio-economic inequalities, encouraging inclusive growth, and enhancing living standards.

Understanding the socioeconomic factors of performance gives businesses a competitive edge when formulating their strategies, enhancing their competitiveness and business strategy. Optimization of operations, supply chain management, and resource allocation can be achieved through insights into consumer preferences and behavior. The study's conclusions advance academic understanding while also offering practical advice to companies, organizations, and legislators that will help them make well-informed decisions and develop strategic plans in a challenging socioeconomic environment.

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 association between socio-economic characteristics and MPCE_URP is demonstrated by the substantial findings of the Tobit regression model, fitted with {survreg} in R. The intercept, age, dummy variables, and p-values are important readings. The majority of variables exhibit statistically significant relationships with MPCE_URP, as evidenced by their low p-values. There are statistically significant predictors such as age, sex, education, religion, and hhdsz. The standard deviation of the latent variable in the model is indicated by the scale parameter (7.60e+00). The predictions of the model are impacted by a bigger scale parameter, which indicates increased variability in the unfiltered latent variable. To evaluate model fit, the model's log-likelihood (-920762.9) is contrasted with a null model. The total model significance is tested using the 21-degree-of-freedom chi-square statistic (20398.14).

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 'fun' function value, which reflects the negative log-likelihood of the model, and the'success' status, which indicates the effectiveness of the optimization procedure, are included in the output. The estimated parameters, which include the intercept and coefficients for each explanatory variable, match the Tobit model's coefficients. The 'nit' value indicates the number of evaluations and iterations, and the 'nfev' and 'njev' values reflect the number of function evaluations and Jacobian evaluations that occur during the optimization process. Standard errors and inferential statistics for the estimated coefficients can be computed using the 'hess_inv' value, which gives the inverse of the Hessian matrix. These findings support the interpretation of the relative contributions of each predictor variable to the outcome within the constraints imposed by the Tobit model, as well as the evaluation of the model's fit and predictive ability by comparing the coefficients.

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

After analyzing the Tobit regression analysis on the "NSSO68.csv" dataset, suggestions are made for additional improvement. Further variables, such as demographics, geography, or other socioeconomic indicators, that could have an impact on the dependent variable (MPCE_URP) should be investigated for inclusion in the model. It is important to guarantee data quality and gathering in order to remedy any discrepancies or missing values. The model's output should be interpreted, providing a justification for the practical or economic importance of each variable's influence on MPCE_URP. Sensitivity analysis ought to be done to evaluate the robustness of the model. Discussions of the findings' potential policy ramifications should focus on how they might influence choices about social welfare initiatives, economic policy, and income distribution. To evaluate the relative performance and insights obtained, comparisons with various models ought to be conducted. It is important to identify potential avenues for future study, such as investigating other variables, running longitudinal studies, or using the model in various datasets or situations. It is important to make sure that the results are reported succinctly and clearly, using visual aids and plain language. These suggestions can enhance well-informed decision-making in pertinent sectors by bolstering the empirical findings and provide insightful information about the elements impacting MPCE_URP.

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