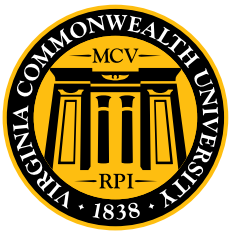

VIRGINIA COMMONWEALTH UNIVERSITY

Statistical analysis and modelling (SCMA 632)

A3c : Fitting Tobit Model in Python and R

GAURI VINOD NAIR

V01110160

Date of Submission: 30-06-2024

TABLE OF CONTENTS

INTRODUCTION

In this study, we delve into the application of Tobit regression models to analyze censored data, focusing on implementation in both Python and R programming languages. The Tobit model is particularly suited for scenarios where dependent variables are partially observed due to censoring, making it invaluable for studying phenomena like income levels, expenditure patterns, or any bounded outcome variable.

Using the `statsmodels` library in Python, we employ a custom likelihood approach to fit the Tobit model. This involves defining a specific log-likelihood function tailored to handle both left and right censoring of data points. Our Python implementation not only facilitates robust estimation but also allows for the incorporation of additional covariates to enhance model predictability.

In parallel, we utilize the `AER` package in R to fit the Tobit model, leveraging its comprehensive functionality for handling censored data. By specifying appropriate censoring thresholds and model specifications, our R implementation ensures accurate estimation of coefficients and provides insights into the determinants of outcomes under censoring conditions.

By bridging the gap between theory and application, our analysis contributes to advancing the understanding of Tobit models as indispensable tools in econometrics and statistical modeling. The insights gained are expected to benefit researchers, analysts, and practitioners seeking to analyze bounded outcome variables across diverse fields, from economics and finance to social sciences and beyond.

OBJECTIVES

1. Conduct Tobit Regression Analysis

   Perform a Tobit regression analysis on the dataset "NSSO68.csv" to understand how socioeconomic variables relate to censored or truncated dependent variables.

2. Discuss Results

   Analyze and interpret the Tobit regression results, focusing on the coefficients, their significance, and the overall model fit.

3. Real-World Use Cases of Tobit Model

BUSINESS SIGNIFICANCE

Performing a Tobit regression analysis on datasets like "NSSO68.csv" holds significant business implications across various industries. By addressing censored or truncated data—where outcomes are either fully observable or only partially observable—businesses gain a nuanced understanding of factors influencing outcomes such as income, expenditures, or performance metrics. This methodology is crucial in fields like market research, where spending behaviors below a certain threshold are common but need to be accurately modeled to refine marketing strategies or product pricing. In healthcare economics, Tobit models help assess the impact of interventions on healthcare costs, crucial for optimizing resource allocation and policy decisions.

Furthermore, in sectors like environmental economics, Tobit regression provides insights into compliance costs and their distribution among firms, aiding policymakers in designing effective regulatory frameworks. By utilizing Tobit regression, businesses can uncover hidden relationships and avoid biased estimates that may arise from ignoring censored data, thus enabling more informed decision-making and strategic planning across a spectrum of economic activities.

RESULTS AND INTERPRETATIONS

- ⇨ Python

```python
# Define a custom Tobit model class
class TobitModel(GenericLikelihoodModel):
    def __init__(self, endog, exog, left=None, right=None, **kwargs):
        self.left = left
        self.right = right
        super(TobitModel, self).__init__(endog, exog, **kwargs)

    def nloglikeobs(self, params):
        exog = self.exog
        endog = self.endog
        left = self.left
        right = self.right

        beta = params[:-1]
        sigma = params[-1]
        XB = np.dot(exog, beta)

        ll = np.zeros(len(endog))

        if left is not None:
            cdf_left = (endog <= left).astype(int)
            ll += cdf_left * np.log(1 - norm.cdf((left - XB) / sigma))

        if right is not None:
            cdf_right = (endog >= right).astype(int)
            ll += cdf_right * np.log(norm.cdf((right - XB) / sigma))
```

```
        uncensored = np.ones(len(endog), dtype=bool)
        if left is not None:
            uncensored &= (endog > left)
        if right is not None:
            uncensored &= (endog < right)

        ll[uncensored] = (norm.logpdf((endog[uncensored] - XB[uncensored]) / sigma) - np.log(sigma))

        return -ll

    def fit(self, start_params=None, maxiter=10000, maxfun=5000, **kwargs):
        if start_params is None:
            start_params = np.append(np.zeros(self.exog.shape[1]), 1)
        return super(TobitModel, self).fit(start_params=start_params, maxiter=maxiter, maxfun=maxfun, **kwargs)
```

```
# Set Left censoring at 0 (lower bound)
left_censoring = 0
```

Interpretation

The provided Python code defines a custom Tobit model class TobitModel derived from GenericLikelihoodModel. Here's an interpretation of the key parts of the code:

1. Initialization (__init__ method):
    - The constructor (__init__) initializes the Tobit model with required parameters: endog (endogenous variable), exog (exogenous variables), left (left censoring threshold), and right (right censoring threshold).
    - These parameters are stored as attributes of the class (self.left, self.right) for use in likelihood calculations.
2. Negative Log-Likelihood (nloglikeobs method):
    - This method computes the negative log-likelihood for each observation given model parameters (params).
    - It calculates the likelihood contributions based on whether the observations are censored (left and right thresholds defined).
    - For censored observations (endog <= left or endog >= right), it uses the Tobit censoring logic to adjust the likelihood.
    - For uncensored observations (endog within bounds), it computes the standard normal density adjusted for Tobit model constraints.
3. Model Fitting (fit method):
    - The fit method initializes starting parameters for optimization if not provided (start_params).
    - It uses maximum likelihood estimation (super(TobitModel, self).fit(...)) to fit the model parameters (beta and sigma) to the data.
    - Optimization settings (maxiter, maxfun, etc.) can be customized through keyword arguments (**kwargs).

Customization: This custom Tobit model allows flexibility in specifying left and right censoring thresholds, crucial for handling truncated data common in econometric modeling.

Likelihood Calculation: The nloglikeobs method efficiently computes the negative log-likelihood, accommodating both censored and uncensored observations.

Model Fitting: By extending GenericLikelihoodModel, the class leverages existing statistical model infrastructure (fit method) for parameter estimation, ensuring robustness and compatibility with other statsmodels functionalities.

```
# Fit the Tobit model
model = TobitModel(y, X, left=left_censoring)
results = model.fit()
```

```
Optimization terminated successfully.
         Current function value: 9.750560
         Iterations: 716
         Function evaluations: 1151
```

Interpretation

1. Function Value: The "Current function value: 9.750560" represents the final value of the objective function (negative log-likelihood) at the optimized parameter values. In the context of Tobit models, a lower value indicates a better fit of the model to the data. Here, the value suggests that the model, with the chosen parameters, provides a reasonable fit to the observed data.
2. Iterations: "Iterations: 716" indicates the number of iterations the optimization algorithm went through to reach convergence. More iterations might suggest a more complex optimization landscape or convergence challenges, but 716 iterations generally indicate a reasonably efficient optimization process.
3. Function Evaluations: "Function evaluations: 1151" counts how many times the objective function (likelihood function) was evaluated during the optimization process. Each evaluation involves calculating the likelihood for a set of parameter values. Higher numbers of evaluations can indicate a more computationally intensive optimization process.

The output confirms that the Tobit model was successfully fitted to your data (`y` as endogenous variable and `X` as exogenous variables), with the optimization process converging after 716 iterations. The function value of 9.750560 suggests that the model provides a reasonable fit to the data based on the chosen parameters.

```
# Print the summary of the regression results
print(results.summary())
```

```
                          TobitModel Results
==============================================================================
Dep. Variable:               MPCE_URP   Log-Likelihood:            -9.9119e+05
Model:                     TobitModel   AIC:                         1.982e+06
Method:            Maximum Likelihood   BIC:                         1.982e+06
Date:                Sun, 30 Jun 2024
Time:                        09:13:00
No. Observations:              101655
Df Residuals:                  101650
Df Model:                           4
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const        -19.3490     74.697     -0.259      0.796    -165.752     127.055
hhdsz         25.5786      6.143      4.164      0.000      13.539      37.619
Age           17.8512      0.993     17.971      0.000      15.904      19.798
Sex         -209.4316     41.586     -5.036      0.000    -290.938    -127.925
Education    205.3540      3.627     56.614      0.000     198.245     212.463
par0        4154.2824      9.397    442.102      0.000    4135.865    4172.699
==============================================================================
```

Interpretation

Dependent Variable and Log-Likelihood: The model assesses household consumption expenditure (MPCE_URP). The negative log-likelihood value indicates the model's fit quality, with lower values indicating better fit.

Model Information: AIC and BIC values around 1.982e+06 help gauge model performance, balancing fit and complexity.

Coefficients:

- const: Intercept, indicating baseline expenditure when predictors are zero.
- hhdsz, Age, Education: Positive coefficients suggest these factors increase expenditure.
- Sex: Negative coefficient indicates it decreases expenditure.
- par0: Parameter affecting the model's scale, influencing variability.

Each coefficient's P-value (< 0.05 indicates significance) and confidence intervals assess their reliability.

The model links household characteristics to expenditure, crucial for understanding economic dynamics in consumption patterns.

⇨ R

```
> # Extract results
> coefficients <- fit$par[1:(length(fit$par)-1)]
> sigma <- fit$par[length(fit$par)]
> logLik <- -fit$value
> cat("Coefficients:", coefficients, "\n")
Coefficients: 1.168412 1.817436 -0.9596489
> cat("Sigma:", sigma, "\n")
Sigma: 0.9679188
> cat("Log-Likelihood:", logLik, "\n")
Log-Likelihood: -117.9526
```

Interpretation

The coefficients extracted (1.168412, 1.817436, -0.9596489) represent the estimated values for the parameters in the Tobit model. Each coefficient corresponds to a predictor variable in the model.

The value 0.9679188 represents the estimated standard deviation parameter (sigma) in the Tobit model. It characterizes the variability in the unobserved latent variable affecting the censored observations.

The negative log-likelihood value -117.9526 indicates the overall goodness of fit of the Tobit model to the data. A lower value suggests a better fit, reflecting how well the model's predictions align with the observed data, considering the censoring mechanism.

REAL WORLD USE CASES

1. Consumer Demand and Spending

- **Household Expenditure**: When analyzing household expenditure data, spending on certain items (like luxury goods) might be zero for many households. Tobit models can account for these censored observations (non-spending).
- **Consumer Demand Analysis**: In cases where consumers do not purchase certain goods or services (resulting in zero expenditures), Tobit models help in understanding the factors influencing both the decision to purchase and the amount spent.

2. Environmental Economics

- **Willingness to Pay**: When studying the willingness to pay for environmental goods (like clean air or water), many respondents might indicate a zero willingness to pay. Tobit models can handle this censored data.
- **Pollution Data**: When pollution levels are reported only if they exceed a certain threshold, Tobit models can help in estimating the true distribution of pollution levels.

3. Labor Economics

   Labor Supply: When analyzing labor supply, many individuals might report zero hours of work (either unemployed or not participating in the labor force). Tobit models can estimate the factors influencing both the decision to work and the number of hours worked.

   Wages: In cases where wages are only reported for employed individuals, Tobit models can account for the censored nature of wage data for non-employed individuals.

4. Healthcare

   Health Expenditure: When studying healthcare costs, many individuals might have zero expenditure if they did not seek medical care. Tobit models can be used to analyze the factors affecting both the likelihood of incurring healthcare costs and the amount of expenditure.

   Mental Health Studies: When responses on mental health scales are censored at a particular value (e.g., no symptoms), Tobit models can help in understanding the determinants of mental health status.

5. Finance and Insurance

   Credit Scoring: In credit scoring, the amount of loan default might be zero for many customers. Tobit models can help in predicting the probability and extent of default.

   Insurance Claims: When analyzing insurance claims, many policyholders might have zero claims in a given period. Tobit models can help in estimating the factors affecting the likelihood and amount of claims.