Bootstrap LM Tests for the Box Cox Tobit Model

David Vincent
Email: david.vincent@hp.com
Introduction

• This presentation sets out a specification test of the Tobit model against the alternative of a specification described by the Box Cox transformation.

• An LM test is used to test the null hypothesis of no specification error as this requires estimates of the restricted (nested) Tobit model.

• The size and power of the test using asymptotic and bootstrap critical values is estimated by the empirical rejection probabilities for small sample sizes.
1. The Box Cox Tobit Model

- The Tobit model is used to address censoring and corner solution problems.
- When censoring occurs at zero, the model in both applications is written:
  \[ y_i^* = x_i'\beta + \epsilon_i, \quad i = 1, \ldots, N \]  
  \[ y_i = \begin{cases} y_i^* & \text{if } y_i^* \geq 0 \\ 0 & \text{if } y_i^* < 0 \end{cases} \]  
  where \( y_i^* \) is a `latent' variable and \( \epsilon_i \sim NID(0, \sigma^2). \) The observation rule is:

- In censored data problems, we are usually interested in the features of \( y_i^* \) such as \( E[y_i^* | x_i] \). For corner solutions however, it is \( E[y_i | x_i] \) that is of interest.
- Estimation of the parameters \( \beta, \) and \( \sigma \) in (1) is by Maximum Likelihood (ML), with individual contribution to the log-likelihood given by:

\[
\ln L_i = d_i \ln \left[ \frac{1}{\sigma} \phi \left( \frac{y_i-x_i'\beta}{\sigma} \right) \right] + (1 - d_i) \ln \left[ 1 - \Phi \left( \frac{x_i'\beta}{\sigma} \right) \right]
\]
1. The Box Cox Tobit Model

- As Moffat (2003) noted however, there are many instances where \( y_i \) exhibits positive skew that cannot be attributed to the asymmetric censoring.

- In the double hurdle model, Moffat takes the following transformation of \( y_i \) to preserve normality:

\[
y_i^T = \frac{y_i^{\lambda - 1}}{\lambda} \quad 0 \cdot \lambda \cdot 1
\]

- The transformation, originally proposed by Box & Cox (1966) for uncensored data, was designed to ensure that the model for \( y_i^T \) is:

1. Linear in the explanatory variables
2. Has a constant conditional error variance \( E[\epsilon \epsilon' \mid X] = \sigma^2 I_N \)
3. Has a normally distributed error term

- The above properties are essential for the ML-estimators to be consistent for the true parameters in the Tobit model (1):

\[
\hat{\beta} \xrightarrow{p} \beta \quad \hat{\sigma} \xrightarrow{p} \sigma
\]
1. The Box Cox Tobit Model

- Applying the Box Cox Transformation (BCT) to the Tobit model therefore, leads to the following observation rule:

\[
y_i^T = \begin{cases} 
y_i^{T*} & \text{if } y_i^{T*} \geq -1/\lambda \\
-1/\lambda & \text{if } y_i^{T*} < -1/\lambda
\end{cases}
\]

- where \( y_i^{T*} \) is the `transformed' latent variable with specification:

\[
y_i^{*T} = x_i' \beta + \epsilon_i, \quad \epsilon_i \sim NID (0, \sigma^2)
\]

- This should now satisfy (or approximately) the distributional requirements for the ML-estimator to be consistent.

- By a change of variables, the \( i^{th} \) contribution to the log-likelihood is:

\[
\ln L_i = d_i \ln \left[ \frac{y_i^{\lambda-1}}{\sigma} \phi \left( \frac{(y_i^\lambda - 1) / \lambda - x_i' \beta}{\sigma} \right) \right] + (1 - d_i) \ln \left[ 1 - \Phi \left( \frac{1/\lambda + x_i' \beta}{\sigma} \right) \right] \quad (2)
\]
2. LM test of the Tobit specification

• A test of the linearity, homoskedasticity and normality assumptions of the Tobit specification, is therefore equivalent to a test of:

\[ H_0 : \lambda = 1 \]

• against the more general alternative:

\[ H_1 : \lambda \neq 1 \]

• The LM-statistic is the easiest to compute as this requires parameter estimates under the restrictions imposed by the null \( \tilde{\theta} = (\beta, \tilde{\sigma}, 1) \).

• Denoting \( \tau \) as an \( N \times 1 \) vector one 1’s, \( \tilde{G} = (\tilde{g}_1, \ldots, \tilde{g}_N)' \) where \( \tilde{g}_i = \frac{\partial \ln L_i}{\partial \theta}|_{\tilde{\theta}} \) represents the \( i^{th} \) contribution to the unrestricted score evaluated at the restricted \( \tilde{\theta} \), then the OPG-version of the LM-test is:

\[ LM = \tau' \tilde{G} (G' G)^{-1} \tilde{G}' \tau \xrightarrow{d} \chi^2_1 \]
2. LM test of the Tobit specification

- In this form, the LM-statistic is simply $N \times R_u^2$ from artificial regression:
  \[ 1 = \tilde{g}_i \pi + e_i \]

- From (2), the individual elements of $\tilde{g}_i$ are:
  \[ \frac{\partial \ln L_i(\theta)}{\partial \beta} \bigg|_{\tilde{\theta}} = d_i \frac{\tilde{v}_{i1}}{\tilde{\sigma}} x_i + (1 - d_i) \frac{-\phi(\tilde{v}_{i2}/\tilde{\sigma})}{1 - \Phi(\tilde{v}_{i2}/\tilde{\sigma})} \frac{x_i}{\tilde{\sigma}} \] (3)
  \[ \frac{\partial \ln L_i(\theta)}{\partial \sigma} \bigg|_{\tilde{\theta}} = d_i \frac{1}{\tilde{\sigma}} \left[ \frac{v_{i1}^2}{\tilde{\sigma}^2} - 1 \right] + (1 - d_i) \frac{\phi(\tilde{v}_{i2}/\tilde{\sigma})}{1 - \Phi(\tilde{v}_{i2}/\tilde{\sigma})} \frac{\tilde{v}_{i2}}{\tilde{\sigma}} \] (4)
  \[ \frac{\partial \ln L_i(\theta)}{\partial \lambda} \bigg|_{\tilde{\theta}} = d_i \left[ \ln y_i - \frac{v_{i1}}{\tilde{\sigma}} [y_i (\ln y_i - 1) + 1] \right] + (1 - d_i) \frac{\phi(\tilde{v}_{i2}/\tilde{\sigma})}{1 - \Phi(\tilde{v}_{i2}/\tilde{\sigma})} \frac{1}{\tilde{\sigma}} \] (5)

- where $\tilde{v}_{i1} = y_i - (1 + x_i' \tilde{\beta})$ and $\tilde{v}_{i2} = 1 + x_i' \tilde{\beta}$. Under the restrictions imposed by the null, (3) and (4) are the scores of the Tobit model evaluated at the Tobit MLE’s; (5) can therefore be constructed from these estimates.
3. Bootstrap Critical Values

- The critical value for a test of size $\alpha$ is the solution to $G_n(c_{n,\alpha}; F_0) = 1 - \alpha$ where $G_n(c; F_0) = Pr(LM \cdot c)$ and $F_0 = F(x_i, y_i; \theta_0)$ is the distribution of the data.

- Unless $F_0$ is known, $c_{n,\alpha}$ cannot be obtained and we use critical values from the limiting distribution under $H_0$, i.e.: $G_\infty(c_{\infty,\alpha}) = Pr(\chi_1^2 \cdot c_{\infty,\alpha}) = 1 - \alpha$

- The size of the test using $c_{\infty,\alpha}$ is $\alpha + O(n^{-1})$ which can be determined through the asymptotic expansion $G_n(c; F_0) = G_\infty(c) + O(n^{-1})$. This error can be large.

- An alternative approach is to obtain critical values from the bootstrap null distribution $G_n(c; F_n)$ which replaces $F_0$ with a consistent estimator $F_n$. Then:

  $$G_n(c; F_0) = G_n(c; F_n) + O(n^{-3/2}) \quad (6)$$

- which has a smaller error of order $O(n^{-3/2})$. The critical value $c_{n,\alpha}^\dagger$ solving $G_n(c_{n,\alpha}^\dagger; F_n) = 1 - \alpha$ be found by Monte Carlo simulation as the $1 - \alpha$ quantile of the B ordered bootstrap statistics $LM_1^\dagger, \ldots, LM_B^\dagger$. 
4. The Parametric Bootstrap Algorithm

- The null \( H_0 : \lambda = 1 \) is rejected if \( LM > c_{n,\alpha}^\dagger \)
- In the \( B \)-simulations, each bootstrap sample is generated by re-sampling \( x_i \) from the EDF, while generating \( y_i \) from \( F(y_i, | x_i; \theta) \). The algorithm is:

1. Estimate the Tobit model parameters: \( \hat{\beta}, \hat{\sigma} \). This imposes the constraint \( \lambda = 1 \)
2. Draw a random sample of size \( N \) from the EDF of \( x_i \) and denote these \( x_i^\dagger, \ldots, x_n^\dagger \)
3. Generate \( N \) errors from \( N(0, \hat{\sigma}^2) \) and denote these \( \epsilon_1^\dagger, \ldots, \epsilon_n^\dagger \)
4. Use the values in steps 2 and 3 to generate a bootstrap sample of size \( N \)
   \( y_i^\dagger = x_i^\dagger' \hat{\beta} + \epsilon_i^\dagger \) and compute \( y_i^\dagger = \max(0, y_i^\dagger) \)
5. Estimate the Tobit model using the bootstrap sample and compute the contributions to the scores \( g_i^\dagger, \ldots, g_N^\dagger \)
6. Estimate the artificial regression \( 1 = g_i^\dagger \delta + u_i \) and compute \( LM_b^\dagger = N \times R_u^2 \)
7. Repeat steps 2 – 6 a total of \( B \)-times and compute the critical value \( c_{n,\alpha}^\dagger \) as the \( 1 - \alpha \) percentile of the \( B \) ordered bootstrap LM-test statistics.
5. Monte-Carlo Design

- The size and power of the LM-test using bootstrap and first-order asymptotic critical values can be estimated from the empirical rejection probabilities.
- The data for the Monte-Carlo experiments is generated from the DGP:

\[ y_i^* T = x_i' \beta + \epsilon_i, \quad y_i^T = \begin{cases} y_i^{T*} & \text{if } y_i^{T*} \geq -1/\lambda \\ -1/\lambda & \text{if } y_i^{T*} < -1/\lambda \end{cases} \]

\[ y_i = (\lambda y_i^{T*} + 1)^{1/\lambda} \]

The experiments consist of the following steps:

1. Generate \( N \) values for \( \epsilon_i \) and \( x_i \) from a specified DGP and compute \( y_i^* T, y_i^T, y_i \)
2. Estimate the LM statistic for testing \( H_0 : \lambda = 1 \) as detailed earlier
3. Compute the bootstrap critical value at the \( \alpha \)-level for testing \( H_0 : \lambda = 1 \)
4. Repeat steps 1-3, \( T \) - times and count the rejections \( R \). The empirical rejection probability \( R/T \), is an estimate of the true rejection probability \( p \).

- As \( R \sim B(T, p) \), then \( \sqrt{T} (R/T - p) \xrightarrow{d} N[0, p(1-p)] \). Thus for \( p = 0.05 \) and \( T = 2000 \), \( Pr(0.04 \cdot R/T \cdot 0.06 | p = 0.05) \approx 0.95 \)
5.1 Size Estimates

- Under $H_0: \lambda = 1$, the empirical rejection probability is an estimate of the size of the LM-test using bootstrap & asymptotic critical values.

- For these experiments $N = 25$, $\alpha = 0.05$, $B = 499$, $T = 2000$, $\epsilon_i \sim NID(0, 1)$ and $x_i'\beta = \beta_0 + \beta_1 x_i$ where: $\ln x_i \sim N(1, 0.5)$, $\beta_0 = 1$ and $\beta_1 \in \{-0.5, -0.55, -0.6, -0.65, -0.7, -0.75, -0.8, -0.85, -0.9, -0.95\}$. The size estimates are:

- Using bootstrap critical values there is no size distortion. This is not the case using asymptotic critical values which result in large size distortions.
5.2 Power Estimates (1)

- Under $H_1 : \lambda = \lambda_1$, the empirical rejection probability is an estimate of the power of the LM-test against the alternative.

- For these experiments, $N = 25$, $\alpha = 0.05$, $B = 499$, $T = 2000$, $\epsilon_i \sim NID(0, 1)$ and $x_i' \beta = \beta_0 + \beta_1 x_i$ where $\ln x_i \sim N(1, 0.5)$, $\beta_0 = 1$, $\beta_1 = -0.5$ and $\lambda = \lambda_1 \in \{.1, .15, .2, ..., 1.3\}$. The power estimates are:

- With the exception of $\lambda = 0.5$, the LM-test using bootstrap critical values at the 5% level of significance seems reasonably powerful for $N = 25$.
5.3 Power Estimates (2)

• Whilst the LM-test exhibits reasonable power for \( \lambda \neq 1 \), it is worth examining the power against DGP’s where a \( \lambda \neq 1 \) would necessary for consistency.

• For these experiments, \( N = 100, \alpha = 0.05, B = 499, T = 2000 \), and the data are generated using similar DGP’s to those used by Drukker(2002):

\[
y_i^* = 1 + x_{i1} + x_{i2} + x_{i3} + \varepsilon_i \sqrt{h(z_i' \alpha)},
\]
\[
x_{i1} \sim N(0, 1) \quad x_{i2} = .3x_{1i} + u_{i2}, \ u_{i2} \sim N(0, 1)
\]
\[
x_{i3} = .3x_{1i} + u_{i3}, \ u_{i3} \sim N(0, 1)
\]

• The \( \varepsilon_i \) are generated from, \( N(0, 1), t_4 \), and \( \chi_5^2 \), distributions and the function \( h(z_i' \alpha) = 1 \) for homoskedastic and \( h(z_i' \alpha) = e^{2x_{i1}} \) for hetroskedastic errors.

• The following table sets out the power estimates:

<table>
<thead>
<tr>
<th>Distribution</th>
<th>( h(z_i' \alpha) = 1 )</th>
<th>( h(z_i' \alpha) = e^{2x_{i1}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( N(0, 1) )</td>
<td>N/A</td>
<td>0.734</td>
</tr>
<tr>
<td>( t_4 )</td>
<td>0.085</td>
<td>0.795</td>
</tr>
<tr>
<td>( \chi_5^2 )</td>
<td>0.140</td>
<td>0.872</td>
</tr>
</tbody>
</table>
6. Description of `bctobit’ Program

bctobit [, Fixed Nodots bfile(string) reps(integer 499)]

**Description**

- **bctobit** computes the LM-statistic for testing $H_0 : \lambda = 1$ against $H_1 : \lambda \neq 1$ in the Box Cox Tobit model. This is equivalent to testing the linearity, normality and homoskedasticity assumptions of the Tobit specification.

- The regressors are assumed to be random, and critical values are obtained from the bootstrap null distribution of the LM test statistic by repeated sampling from the (parametric) bootstrap DGP.

**Options**

- **Fixed** - specifies that the regressors are fixed in the bootstrap null distribution
- **Nodots** – suppresses the replication dots
- **bfile(name)** – the name of the saved file which contains the LM-statistics computed from the bootstrap samples
- **reps(#)** - the number of samples to be drawn from the bootstrap DGP to estimate the percentiles of the bootstrap null distribution. Default is 499
6. Description of `bctobit’ Program

Tobit regression

Number of obs = 100
LR chi2(3) = 139.54
Prob > chi2 = 0.0000
Pseudo R2 = 0.3734

Log likelihood = -117.08451

| y   | Coef.  | Std. Err. | t     | P>|t| | [95% Conf. Interval] |
|-----|--------|-----------|-------|------|----------------------|
| x1  | .8808724 | .1447619 | 6.08  | 0.000 | .5935602 1.168185    |
| x2  | .9554311 | .1253373 | 7.62  | 0.000 | .7066713 1.204191    |
| x3  | .9387104 | .1204485 | 7.79  | 0.000 | .6996535 1.177767    |
| _cons | 1.200638 | .1305344 | 9.20  | 0.000 | .9415631 1.459712    |
| /sigma | 1.05923 | .0898169 |       |       | .8809688 1.237492    |

Obs. summary: 29 left-censored observations at y<=0
71 uncensored observations
0 right-censored observations

.bctobit, reps(299)
Bootstrap replications (299)

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>150</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>200</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>250</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

LM test of Tobit specification
Bootstrap critical values

<table>
<thead>
<tr>
<th>1m</th>
<th>%10</th>
<th>%5</th>
<th>%1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.4669</td>
<td>2.86527</td>
<td>4.1014972</td>
<td>10.135839</td>
</tr>
</tbody>
</table>
7. Further Research....

• A natural extension would be to consider the alternative of a Box Cox transformation with an error term that is hetroskedastic

\[ y_{i}^{T*} = x_{i}'\beta + \epsilon_{i}\sqrt{h(z_{i}\alpha)}, \]

where \( h \) is an unknown function , with \( h'(.) \neq 0, h(0) = 1 \) and \( h'(0) = \kappa \)

• A test of the joint hypothesis: \( H_1 : \lambda = 1, \eta = 0 \) against the alternative of \( H_1 : \lambda \neq 1, \eta \neq 0 \) is equivalent to testing the validity of the Tobit specification.

• The LM statistic would now be based on the additional components of the score vector, evaluated at the restrictions given by the null. These are:

\[
\frac{\partial \ln L_{i}(\theta)}{\partial \alpha} \bigg|_{\tilde{\theta}} = d_{i} \frac{1}{2} \left[ \frac{\tilde{v}_{i1}^{2}}{\sigma^{2}} - 1 \right] \kappa z_{i} + (1 - d_{i}) \frac{-\phi(\tilde{v}_{i2}/\sigma)}{1 - \Phi(\tilde{v}_{i2}/\hat{\sigma})} \frac{\kappa z_{i}}{2\hat{\sigma}}
\]

• As such \( LM \xrightarrow{d} \chi^2_{1+\text{dim}(z)} \). The size and power using bootstrap critical values can be estimated from empirical rejection probabilities as before.
8. References


- Moffatt, P. G. (2003) “Hurdle models of loan default”, *School of Economic and Social Studies, University of East Anglia, Norwich, UK*