Statistics Innovation

METHOD

Open Access

https://doi.org/10.48130/stati-0025-0004

Statistics Innovation **2025**, 2: e004

Limit predictions in the Tobit model: an alternative derivation approach

Spyros Missiakoulis*

School of Economics and Management, Philips University, 4-6 Lamias Street, 2001 Nicosia, Cyprus

* Corresponding author, E-mail: s.missiakoulis@philipsuni.ac.cy

Abstract

This method considers the problem of predicting already censored values of the dependent variable in the Tobit model. Assuming normality, it suggests a prediction formula that is equivalent to the classical correction using the inverse Mills ratio (IMR). The main contribution is providing an alternative derivation linking Ordinary Least Squares (OLS) and Maximum Likelihood Estimation (MLE) approaches, providing additional practical implementation insights. An extensive experimental investigation and robustness checks demonstrate the method's performance.

Citation: Missiakoulis S. 2025. Limit predictions in the Tobit model: an alternative derivation approach. *Statistics Innovation* 2: e004 https://doi.org/10.48130/stati-0025-0004

Introduction

Tobit models (censored regression) can be categorized into two classes according to whether or not observations on the censored part of the dependent variable are observable. In the classical example of a household's expenditure on major durable goods, the observation of zero expenditure represents the actual expenditure since there is no way to observe expenditure that is less than zero. In unemployment duration problems, however, the maximum value of the duration period is always the monitoring period. For those persons with the maximum period of duration, the recorded value is not the actual one. The actual duration could be observed only by extending the monitoring period. In both examples, the Tobit model, as the appropriate econometric model, is used to estimate the parameters of the model and to generate predictions from the model when new data are available. Furthermore, it may be used to predict the values of the dependent variable for the censored observations.

This note focuses on using the Tobit model not only for estimation but for predicting censored observations. Although the correction formula obtained is equivalent to the classical prediction using the inverse Mills ratio (IMR), the contribution of this note lies in offering an alternative derivation linking Ordinary Least Squares (OLS) and Maximum Likelihood Estimation (MLE) approaches, thus providing additional implementation insights.

The Tobit model

Tobit model assumes the following relationship between a latent variable, y^*, and an observed variable, y:

$$\begin{aligned} y_i^* &= x_i\beta + u_i,\ \ i = 1, \ldots, n, \\ y_i &= y_i^*, \quad \text{if } y_i^* > 0 \\ &= 0, \text{ if } y_i^* \le 0 \end{aligned} \tag{1}$$

where, the notation is obvious. If u_i is normally distributed, the maximum likelihood estimator of β, β_{ML}, say, is given in matrix notation by:

$$\boldsymbol{\beta_{ML}} = \boldsymbol{\beta_1} - \sigma(X_1^T X_1)^{-1} X_0^T \boldsymbol{\gamma_0} \tag{2}$$

where, $\boldsymbol{\beta_1} = (X_1^T X_1)^{-1} X_1^T \boldsymbol{y_1}$ is the OLS estimator based on the non-limit observations, $\boldsymbol{\gamma_0} = \dfrac{\boldsymbol{f_0}}{(1-\boldsymbol{F_0})}$, $\sigma^2 = \dfrac{\boldsymbol{y_1^T}(\boldsymbol{y_1} - \boldsymbol{X_1}\boldsymbol{\beta_{ML}})}{n_1}$, $n_1 + n_0 = n$, $\boldsymbol{f_0}$ and $\boldsymbol{F_0}$ standardized normal density and distribution functions evaluated at $\dfrac{\boldsymbol{x_i}\boldsymbol{\beta_{ML}}}{\sigma}$, and the subscripts 0 and 1 indicate limit and non-limit observations[1].

Limit predictions

Following Maddala[2], the censored part of the dependent variable is predicted by the expected mean of the latent variable:

$$E(\boldsymbol{y_0^*}) = \boldsymbol{X_0}\boldsymbol{\beta_{ML}} \tag{3}$$

and the uncensored part by:

$$E\left(\boldsymbol{y_1^*}\right) = E\left(\boldsymbol{y_1}|\boldsymbol{y^*} > \boldsymbol{0}\right) = \boldsymbol{X_1}\boldsymbol{\beta_{ML}} + \sigma\frac{\boldsymbol{f_1}}{\boldsymbol{F_1}} \tag{4}$$

The ratio $\dfrac{\boldsymbol{f_1}}{\boldsymbol{F_1}}$ is the IMR that corrects for the bias introduced by censoring, and adjusting the predicted latent variable relative to the simple linear prediction[2,3].

The prediction in Eq. (3) gives the mean of potential y's, and is not appropriate when the goal is to predict values already known to have been censored. In such a case, an equivalent to Eq. (4) is needed. Assuming normality, an appropriate prediction formula is sought, such that:

$$E(\boldsymbol{y_0^*}) = E\left(\boldsymbol{y_0}|\boldsymbol{y^*} \leqslant \boldsymbol{0}\right) = \boldsymbol{X_0}\boldsymbol{\beta_{ML}} + \boldsymbol{\delta} \tag{5}$$

where, $\boldsymbol{\delta}$ is an unknown but necessary correction, equivalent to $\dfrac{\sigma \boldsymbol{f_1}}{\boldsymbol{F_1}}$ in Eq. (4). Hence the problem reduces to find a valid estimate of $\boldsymbol{\delta}$.

Assume that *all* observations were available, i.e. no censoring. In this case, OLS will be fully applicable and appropriate. Without loss of generality, assume that the first n_1 observations are the non-limit ones, and the last n_0 are the limit observations. The OLS estimator, $\boldsymbol{\beta_{OLS}}$, can then be written in the following recursive form[4].

$$\boldsymbol{\beta_{OLS}} = \boldsymbol{\beta_1} + \left(\boldsymbol{X_1^T X_1}\right)^{-1}\boldsymbol{X_0^T}\left[\boldsymbol{I_0} + \boldsymbol{X_0}\left(\boldsymbol{X_1^T X_1}\right)^{-1}\boldsymbol{X_0^T}\right]^{-1}\left(\boldsymbol{y_0} - \boldsymbol{X_0}\boldsymbol{\beta_1}\right) \tag{6}$$

If we could find a value $\hat{\boldsymbol{y}}_{\boldsymbol{0}}$ such that the second term on the right-hand side (RHS) of Eq. (6) equals the second term on the RHS of Eq. (2), then the estimators $\boldsymbol{\beta_{ML}}$ and $\boldsymbol{\beta_{OLS}}$ — based on $\hat{\boldsymbol{y}}_{\boldsymbol{0}}$ — would coincide, and $\hat{\boldsymbol{y}}_{\boldsymbol{0}}$ would exhibit the desired predictive properties.

$$-\sigma\left(\boldsymbol{X_1^T X_1}\right)^{-1}\boldsymbol{X_0^T}\boldsymbol{\gamma_0} = \left(\boldsymbol{X_1^T X_1}\right)^{-1}\boldsymbol{X_0^T}\left[\boldsymbol{I_0} + \boldsymbol{X_0}\left(\boldsymbol{X_1^T X_1}\right)^{-1}\boldsymbol{X_0^T}\right]^{-1}\left(\boldsymbol{y_0} - \boldsymbol{X_0}\boldsymbol{\beta_1}\right) \Longrightarrow$$

$$\left(\boldsymbol{X_1^T X_1}\right)^{-1}\boldsymbol{X_0^T}\left[\boldsymbol{I_0} + \boldsymbol{X_0}\left(\boldsymbol{X_1^T X_1}\right)^{-1}\boldsymbol{X_0^T}\right]^{-1}\boldsymbol{y_0} =$$

$$\left(\boldsymbol{X_1^T X_1}\right)^{-1}\boldsymbol{X_0^T}\left[\boldsymbol{I_0} + \boldsymbol{X_0}\left(\boldsymbol{X_1^T X_1}\right)^{-1}\boldsymbol{X_0^T}\right]^{-1}\boldsymbol{X_0}\boldsymbol{\beta_1} - \sigma\left(\boldsymbol{X_1^T X_1}\right)^{-1}\boldsymbol{X_0^T}\boldsymbol{\gamma_0}$$

premultiplying by $\left(\mathbf{X_1^T X_1}\right)$ gives:

$$X_0^T\left[I_0 + X_0\left(X_1^T X_1\right)^{-1} X_0^T\right]^{-1} y_0 = X_0^T\left[I_0 + X_0\left(X_1^T X_1\right)^{-1} X_0^T\right]^{-1} X_0\beta_1 - \sigma X_0^T\gamma_0 \quad (7)$$

What Eq. (7) represents is a system of kn_0 equations with n_0 unknowns, where k is the number of explanatory variables. In this case an estimate of $\boldsymbol{y_0}$ is obtained by premultiplying Eq. (7) firstly by the generalized inverse of X_0^T and secondly by $\left[I_0 + X_0\left(X_1^T X_1\right)^{-1} X_0^T\right]$, which results in:

$$\begin{aligned}\hat{y}_0 &= X_0\beta_1 - \sigma\left[I_0 + X_0\left(X_1^T X_1\right)^{-1} X_0^T\right]\gamma_0 = X_0\beta_1 - \sigma\gamma_0 - \sigma X_0\left(X_1^T X_1\right)^{-1} X_0^T\gamma_0 \\ &= X_0\beta_1 - \sigma X_0\left(X_1^T X_1\right)^{-1} X_0^T\gamma_0 - \sigma\gamma_0 = X_0\left[\beta_1 - \sigma\left(X_1^T X_1\right)^{-1} X_0^T\gamma_0\right] - \sigma\gamma_0 \\ &= X_0\beta_{ML} - \sigma\gamma_0\end{aligned} \quad (8)$$

A straightforward comparison of Eqs (5) and (8) gives us:

$$\delta = -\sigma\gamma_0 \quad (9)$$

which, in turn, implies that:

$$E\left(y_0^*\right) = X_0\beta_{ML} - \sigma\gamma_0 = X_0\beta_{ML} - \sigma\frac{f_0}{(1 - F_0)} \quad (10)$$

Thus, formula Eq. (10) produces desired predictions for the censored values of the dependent variable.

Experimental investigation

To examine the effectiveness of the limit-prediction method, the following research experiment was designed.

Step 1: Without loss of generality, a simple model with a single explanatory variable was used.

Step 2: Through the random number generator, the variables y^* and X were constructed for 1,000 observations. For X, a univariate distribution between 0 and 10 was assumed, while for y^*, a univariate distribution between –20 and 100 was assumed. Although the choice of these two univariate distributions is arbitrary, it does not affect the findings in any way.

Step 3: From y^*, 13 y variables were constructed. These 13 variables differ in the number of limit observations, namely 10%, 20%, 25%, 30%, 33%, 40%, 50%, 60%, 67%, 70%, 75%, 80%, and 90%.

Step 4: Each of the 13 models was estimated using the Tobit method.

Step 5: In each of the 13 sets of limit observations, two predictions were made: one based on Eq. (3), and the other on Eq. (10).

For each method and each set of observations, Table 1 shows the number of limit observations that were incorrectly predicted as non-limit. For completeness, the corresponding results for the inappropriate OLS method are also included.

In cases where non-limit values were incorrectly predicted, when in fact they were limit values, it is interesting to examine the range of predicted values. In all cases without exception, the incorrectly predicted values are the closest to the limit, 0 in this case. This is reasonable and expected because as we move away from the limit, $\sigma\gamma_0$ tends to zero, with the consequence that Eqs (3) and (10) tend to be identical.

In all experiments, and in order to compare the prediction errors of Eqs (3) and (10), Wilcoxon signed-rank tests were first applied as appropriate non-parametric tests that do not require normality of the tested values, followed by paired t-tests, which require normality of the tested values. The said tests were conducted across all censored observations.

In all cases, the Wilcoxon signed-rank tests yielded $W^+ = 0$ with corresponding p-values ≈0.000, which under the conditions is a very interesting and reasonable finding. This means that all differences are negative, that is, in all censored observations across all simulation settings, Eq. (10) had a smaller absolute error than Eq. (3).

Table 1. Limit observations that were incorrectly predicted as non-limit.

% of limit observations	Number of limit observations (n_0)	$X_0\beta_{ML}$ Eq. (3)	$X_0\beta_{ML}$-$\sigma\gamma_0$ Eq. (10)	OLS
10%	100	10	0	31
20%	200	14	0	68
25%	250	13	0	94
30%	300	9	0	111
33%	333	11	0	128
40%	400	17	0	173
50%	500	3	0	250
60%	600	11	0	327
67%	666	7	0	377
70%	700	6	0	406
75%	750	0	0	446
80%	800	0	0	492
90%	900	0	0	578

This result is logical because formula Eq. (10) is clearly more accurate, and the number of observations is large (so it is not a coincidence).

The paired t-tests comparing absolute prediction errors from Eqs (3) and (10) across all censoring levels showed statistically significant improvements in favor of Eq. (10), with p-values consistently ≈ 0.000. The only exception was at the 90% censoring level, where the difference was not statistically significant (p = 0.165), highlighting the expected limitations of any predictive correction when only 10% of the data is uncensored. A result that is consistent with the idea that prediction accuracy deteriorates rapidly when the effective sample size (i.e., the uncensored portion in this case) becomes too small to support reliable estimation and/or testing.

Since, in this experiment, y^* is known, two of the most popular and widely used error statistics were applied, namely mean absolute deviation (MAD) and root mean squared error (RMSE), to compare the accuracy of predictions. Table 2 presents the prediction performances, in terms of MAD and RMSE, for all the limit cases of all the datasets (13 in total) used. It was observed that Eq. (10) had a very slight footprint over Eq. (3).

To validate that the observed prediction performance was not due to random variation, 6,498 similar experiments were conducted and the corresponding MAD and RMSE were computed for both Eqs (3) and (10). Again, paired t-tests and Wilcoxon signed-rank tests were applied to the distribution of differences in prediction errors. These tests showed that the differences between the two equations are statistically significant and, therefore, are not the result of chance.

Furthermore, the robustness of the correction method was analyzed under different data distributions. The predictability of each method was compared under the assumption that errors follow either the standard normal distribution, the t distribution with 30 degrees of freedom, or the uniform distribution with range (–1.73, 1.73). The t and uniform distributions were chosen to have a mean as close to zero as possible and a variance as close to unity as possible. The first choice ($\mu \approx 0$) was made because it is one of the basic assumptions of regression models, while the second ($\sigma^2 \approx 1$) was made for simplicity reasons. The relevant results are reported in Table 3.

Table 2. Prediction performances (% of limit observations).

	$X_0\beta_{ML}$ Eq. (3)	$X_0\beta_{ML}$-$\sigma\gamma_0$ Eq. (10)	OLS
MAD	5.30	5.27	30.88
RMSE	6.71	6.70	36.53

Table 3. Prediction performances (different error distribution).

Error distribution	MAD			RMSE		
	$X_0\beta_{ML}$ Eq. (3)	$X_0\beta_{ML}$-$\sigma\gamma_0$ Eq. (10)	OLS	$X_0\beta_{ML}$ Eq. (3)	$X_0\beta_{ML}$-$\sigma\gamma_0$ Eq. (10)	OLS
Normal	0.83	0.66	1.85	1.03	0.83	2.04
t	0.82	0.68	1.86	1.04	0.84	2.05
Uniform	0.85	0.67	1.84	0.99	0.81	2.03

Even under the presence of heavier-tailed error distributions, Eq. (10) maintained a clear advantage in predictive accuracy over both Eq. (3) and the OLS-based approach. Consistent with earlier findings, statistical tests—including paired t-tests and Wilcoxon signed-rank tests—were conducted on the resulting differences in prediction errors. These tests confirmed that the improvements offered by Eq. (10) remained statistically significant across all evaluated cases. Nonetheless, it is important to note that under non-normal errors, the assumptions underlying the MLE are violated, and it no longer guarantees optimality in estimation.

Conclusions

This method revisited the classical problem of predicting censored outcomes in the Tobit model. While the final correction formula for censored observations Eq. (10) is mathematically equivalent to the classical IMR adjustment Eq. (3), the contribution of this work lies in offering an alternative derivation. By linking OLS and MLE, the proposed approach provides deeper insight into the structure of the correction and may offer practical advantages in implementation and interpretation.

Theoretical equivalence, however, does not guarantee empirical equivalence. Through extensive simulation experiments across varying levels of censoring, it was shown that Eq. (10) systematically reduces prediction errors compared to the classical approach. These improvements were confirmed using both parametric (paired t-tests) and non-parametric (Wilcoxon signed-rank) tests, which demonstrated statistically significant gains in all settings except at extreme censoring (90%), where the effective sample size is severely limited. Additional robustness checks with non-normal error distributions confirmed that the correction retains its performance under model misspecification.

Overall, this study not only reinforces the validity of the classical correction formula but also enriches understanding of it through a novel derivation and comprehensive empirical evaluation. The findings suggest that the proposed derivation may serve as a useful alternative pedagogical and computational route for researchers and practitioners dealing with censored data.

Author contributions

The author confirms sole responsibility for the following: study conception and design, data collection, analysis and interpretation of results, manuscript preparation, and approval of the final version of the manuscript.

Data availability

The dataset generated during the current study is available from the corresponding author on reasonable request.

Acknowledgments

I am indebted to the associate editor and to the two anonymous referees for encouragement, helpful hints, and constructive comments on an earlier draft of the manuscript.

Conflict of interest

The author declares that there is no conflict of interest.

Dates

Received 19 March 2025; Revised 11 June 2025; Accepted 8 July 2025; Published online 11 August 2025

References

1. Fair RC. 1977. A note on the computation of the tobit estimator. *Econometrica* 45(7):1723–27
2. Maddala GS. 1983. *Limited-dependent and qualitative variables in econometrics (Econometric Society Monographs)*. Cambridge, MA: Cambridge University Press
3. Green WH. 2017. *Econometric Analysis*. 8th Edition. New York: Pearson International
4. Phillips GDA. 1977. Recursions for the two-stage least-squares estimators. *Journal of Econometrics* 6:65–77