Likelihood Estimation for Censored Random Vectors

Wendelin Schnedler*
Department of Economics
Heidelberg University

April 11, 2005

Abstract

This article shows how to construct a likelihood for a general class of censoring problems. This likelihood is proven to be valid, i.e. its maximiser is consistent and the respective root-n estimator is asymptotically efficient and normally distributed under regularity conditions. The method generalises ordinary maximum likelihood estimation as well as several standard estimators for censoring problems (e.g. tobit type I - tobit type V).

JEL classification: C130; C240

Keywords: Censored variables; Limited dependent variables; Multivariate methods; Random censoring; Likelihood

*Alfred-Weber-Institut, Universität Heidelberg, Grabengasse 14, 69117 Heidelberg (wendelin.schnedler@awi.uni-heidelberg.de)

I am indebted to Joseph Lafranchi, who initiated my research on this topic by confronting me with a censoring problem; as well as to Simon Burgess, Ingolf Dittmann, Bernd Fitzenberger, Cheng Hsiao, Offer Lieberman, Winfried Pohlmeier, Jean-Marc Robin, Uwe Sunde, Harald Uhlig and seminar participants at Bristol and Humboldt University. I also like to thank two anonymous referees for comments which considerably improved the article. All errors remain my own.

1 Introduction

The value of a variable of economic interest can often only be observed under particular circumstances – the variable is censored. Ignoring censoring will generally lead to inconsistent estimators. The seminal example is from Tobin (1958): because household expenditure is only observed when it is positive, ordinary least squares estimators for the relationship between household expenditure and income are downwardly biased. Since Tobin's contribution, a plethora of censoring problems has been examined where the observation of a particular random variable depends on whether it is above or below a fixed threshold or the value of another random variable. The classical approach to obtain estimates under these circumstances is to derive a likelihood function and use its maximiser – Amemiya (1984) surveys and classifies respective articles. But writing down an objective function alone does not guarantee that the maximiser has the properties of a maximium likelihood estimator (for a recent example see Attanasio, 2000). In a seminal article Amemiya (1973) provides an involved proof why the tobit type I estimator has such properties. Fortunately, it is now considerably easier to ensure these properties for maximisers of a given objective function by using the results on M-estimation (for an overview see Newey and McFadden, 1996). This, however, does not solve the problem how to obtain the objective function in the first place. There is no rule which explains how to derive a valid likelihood in the presence of censoring. Without such a rule it is not only difficult to find a likelihood, it is –in principle– also necessary to ensure asymptotic properties afresh from first principles for each censoring problem.

Here, we provide an explicit, unified framework for maximum likelihood estimation with multi-dimensional censored variables.[1] Starting from the specifics of the censoring problem and the distribution of the latent variables, we explain how to find an objective function such that its maximiser has all the asymptotic properties of a maximum likelihood estimator. The framework is general, covers several standard problems, and incorporates the respective estimators such as tobit types I - V or Nelson's estimator (1977).

The next section introduces the necessary notation to describe the censoring problem. In section 3, this description is used to derive a valid likelihood. In section 4, the method is applied to some classical and new censoring problems. Finally, section 5 concludes.

[1] The seminal classification of Amemiya (1985) exclusively deals with one-dimensional censored variables; the important work of Gourieroux et al. (1987) is limited to exponential families, while their formula for the likelihood (2.3) is not particularly accessible.

2 Describing the censoring problem

In this section, we start out with a simple representation of the censoring mechanism (visibility sets) in order to derive another representation which can be used for estimation (state set). Finally, we introduce the notation needed to describe the data available for estimation.

The following example illustrates a typical censoring problem. An employer pays the moving costs of workers, who have to relocate. To keep costs low, the employer uses the following rule: the worker has to obtain two quotes and the cheaper one is paid for. The employer records the name of the selected moving company and its price. Is it possible to estimate the mean price suggested by a moving company using the records of the employer? Clearly, the average price observed for this moving company underestimates its mean price offer. The price is simply more likely to be observed when it is lower. Is there nevertheless a way to consistently estimate this mean?

This example is a special case of the more general problem, how to estimate a p-dimensional parameter $\theta \in \mathbb{R}^p$, which governs a random vector $Y = (Y_1, \ldots, Y_q)$ with realisation $y = (y_1, \ldots, y_q) \in \mathbb{R}^q$ when some components of y cannot be observed sometimes. We assume to know the joint density of Y with respect to some measure μ and denote it by $f(\cdot, \theta)$.[2] In the example of the moving companies, the parameters of interest are the mean prices. For simplicity, we want to assume that there are only two moving companies, so $\theta = (\mu_1, \mu_2)$. Then, Y describes how prices are generated and $y = (y_1, y_2)$ are the actual prices.

Before we can advance with the estimation, we need a formal description of how and when the censoring occurs. Suppose the vector Y consists of all censored variables as well as the variables which determine observability (which may or may not be censored themselves). This is not a very restrictive assumption because any missing relevant variable can simply be added as a component to Y. Since we have all relevant variables at hand, we can say for which realisations y of the vector Y, we can observe the j-th component y_j. We collect the respective realisations in the *visibility set* for the j-th component: $V_j := \{y | y_j \text{ visible}\}$. Conversely, $\bar{V}_j$ denotes the realisations for which y_j is not observable.

[2] If $F(\cdot, \theta)$ is the cumulative distribution function of Y, then the density and measure are formally defined as: $d\mu := dy$ if $F(y, \theta)$ absolutely continuous in y and one else; while $f(y, \theta)$ is the derivative of $F(y, \theta)$ if it is absolutely continuous and $P(Y = y)$ otherwise.

In the example, the price by the first moving company y_1 is observed whenever it is smaller or equal than y_2 (for simplicity, we suppose that the first offer is chosen when both prices are equal). Conversely, y_2 is observed when it is smaller than y_1. So, the visibility set for y_1 is $V_1 = \{y_1, y_2 | y_1 \leq y_2\}$ and that for y_2 is $V_2 = \{y_1, y_2 | y_2 < y_1\}$.

In order to be able to compute the probability that a component is visible, we need the following assumption.

Assumption 1. *The visibility set V_j is (μ-)measurable for all j.*

This assumption restricts the type of censoring problem to which we can apply the method proposed later. However, it is the only restriction and it is hard to imagine any other form of likelihood estimation once it is violated.

While the visibility sets already embody all relevant information about the censoring problem, they represent it in a form which cannot be used for estimation because the estimation procedure has to deal with the visibility of several components. For an individual component, visibility can be characterised by two outcomes: either it is observable or not. Describing the visibility of all q components together, there are 2^q different outcomes. We call these outcomes *(visibility) states*. These states can be numbered $s = 0, \ldots, 2^q - 1$, where we reserve the label $s = 0$ for the state in which no component is visible.

In the moving example, there are four states: no price is visible ($s = 0$), only the price of the first company is visible ($s = 1$), only the price of the second company is visible ($s = 2$), and both prices are visible, ($s = 3$).

Since the visibility state describes the visibility of every component, it also imposes restrictions on the realisation of Y. The state s occurs if and only if y is in the following *(visibility) state set*:

$$W_s := \bigcap_{\{j|j \text{ visible in } s\}} V_j \quad \cap \quad \bigcap_{\{j|j \text{ not visible in } s\}} \bar{V}_j$$

With the states and the state sets we have derived a representation of the censoring mechanism that we can (and will) use for estimation later.

In the moving example, the state sets are $W^0 = \bar{V}_1 \cap \bar{V}_2 = \emptyset$ for state $s = 0$ because one price is always observable, $W^1 = V_1$ if the first price is observable, $W^2 = V_2$ if the second price is observable, and $W^3 = V_1 \cap V_2 = \emptyset$

because both prices are never observable at the same time.

Before we can describe the available data, we need a final piece of notation. Let v^s be an operator which extracts the observable components of y for a given state s. In addition, we define the operator $\bar{v}_s$ to extract the components of y which are not observable in state s.[3] In order to see how the operators work, reconsider the moving company example: $v_1 y = y_1$ because the observable component in state $s = 1$ is y_1, while $v_2 y = y_2$ in state $s = 2$ when the second provider submitted the lower bid ($y_2 < y_1$). Likewise the unobservable component in state $s = 1$ can be obtained by the invisibility operator $\bar{v}_1 y = y_2$, while the invisibility operator yields $\bar{v}_2 y = y_1$ in state $s = 2$ where only y_2 is observable.

With this operator in place, we can now formally express the data available for estimation. Let $i = 1, \ldots, n$ be the index of n observational units which are an independent random sample of Y. A particular realisation of this random sample is denoted by $\mathbf{y}_i = (y_{i1}, \ldots, y_{iq})$ and leads to a state s_i. Then, the visibility operator allows the following succinct representation of the data which is available for estimation:

$$(s_i, v_{s_i}\mathbf{y}_i)_{i=1,\ldots,n}.$$

In words: the econometrician knows which variables are observable and the values for those variables.

3 Estimation

How can we use the data to consistently estimate the parameters? In ordinary maximum likelihood estimation, we would take the density evaluated at the observed realisation and maximise it in the unknown parameter. Here, this is not possible as some components of y are not observed in some states. However, we can deduce that the values of the unobserved components lie in the state set of the respective state. Also, we know their distribution on this set. Using this information, we can eliminate the unobserved components

[3] Formally, the operator v is a function of both, the state s and the realisation y:

$$\begin{aligned} v : \{0, \ldots, 2^{q-1}\} \times I\!R^q &\longrightarrow I\!R^{l(s)} \subseteq I\!R^q \\ (s, y) &\longmapsto (y_{j_1}, y_{j_2}, \ldots, y_{j_{l(s)}}), \end{aligned}$$

where $j_1, \ldots, j_{l(s)} \in \{j | j \text{ visible in state } s\}$ and $l(s)$ is is the number of observable components in state s. The operator $\bar{v}_s$ is defined completely analogously.

in state s by integrating them on the state set W_s. The contribution of a realisation y in state s then only depends on observable components:

$$\tilde{f}_s(v_s y, \theta) := \int_{W_s} f(y, \theta) d\mu(\bar{v}_s y),$$

where the integration is ignored if all variables are observable. Note, that while looking complicated, the term $d\mu(\bar{v}_s y)$ simply indicates that integration should be carried out for the components which are not observed.

The contribution of realisation y can also be motivated differently. For discrete variables, the maximum likelihood estimator is obtained as the parameter value $\hat{\theta}$ under which an observed event is most likely. Take the event that the observable components $v_s y$ fall into some set A and that the state is s. This event has the probability:

$$P(v_s y \in A \wedge y \in W_s | \theta) = \int_A \int_{W_s} f(y, \theta) d\mu(\bar{v}_s y) d\mu(v_s y) = \int_A \tilde{f}_s(v_s y, \theta) d\mu(v_s y).$$

The parameter which maximises this probability also maximises a multiple of this probability. As the integral is proportional to the integrand for small changes, we can eliminate the outer integral and take the inner integral as the contribution of y. This leads to the same contribution as suggested before: $\tilde{f}_s(v_s y, \theta)$.

There is also a third motivation for the form of the contributions. In state s, the variables $v_s y$ are observable. We could directly use the density of $v_s y$ for a given state s: $f_s(v_s y, \theta)$.[4] This density accurately describes the observed values, however it does not take into account the likelihood of state s itself. In order to incorporate this likelihood, we weigh the density with the probability of state s. This results in $f_s(v_s y, \theta) \cdot P(y \in W_s)$, which again is identical to $\tilde{f}_s(v_s y, \theta)$.

If we want to compute the contribution for the i-th observational unit, we simply replace s and $v_s y$ by the visibility state and observed values of this unit: s_i and $v_{s_i} \mathbf{y}_i$. The joint objective function for n independently drawn units is then obtained by multiplying the contributions. This leads to the *likelihood estimator*:

$$\hat{\theta}_n := \text{argmax}_\theta \prod_{i=1}^{n} \tilde{f}_{s_i}(v_{s_i} \mathbf{y}_i, \theta) \tag{1}$$

[4]The density is computed as $f_s(v_s y, \theta) := \int_{W_s} f(y, \theta) d\mu(\bar{v}_s y) / \int_{W_s} f(y, \theta) d\mu(y)$.

Of course, this is only one of many possible ways to define a likelihood estimator. The chosen definition is based on the idea of visibility states and there is a-priori no reason to classify contributions by these states. Even when contributions are linked to visibility states, they could be calculated differently, for example by not restricting the unobserved components (see formula 5.1 in Rubin, 1976 or formula 5.11 in the textbook of Little and Rubin, 1987) or by limiting them to some other set instead of the state set W_s. Also, contributions could be weighed differently. A specific example for a different weighing is the *state conditional likelihood estimator* which uses $f_s(v_s y, \theta)$ in place of $\tilde{f}_s(v_s y, \theta)$.

Nevertheless, the name likelihood estimator for $\hat{\theta}_n$ is justified because it has various desirable properties known from ordinary maximum likelihood estimators:

Theorem 1. *Under regularity conditions, the maximiser $\hat{\theta}_n$ is consistent. $\sqrt{n}\hat{\theta}_n$ is asymptotically normally distributed and asymptotically efficient.*

In order to prove this theorem, we can directly apply the results for M-estimators to the general form of the objective function in (1). Hence, the novelty of the theorem does not lie in its proofs (the interesting reader finds them and the regularity conditions in the appendix) but in the theorem itself. We extended the idea of maximum likelihood estimation to the censored context by providing a rule how to derive a likelihood (definition of visibility sets, state sets and calculation of contributions). The theorem guarantees that the respective estimator also has the properties of a maximum likelihood estimator. Due to this result, our approach puts an end to the problem of how to find an objective function when there is censoring and to the ensuing difficulty of ensuring that the maximiser of this objective function is "good".

The method proposed here does not only resemble the ordinary maximum likelihood method in its relative ease of application and the properties of the estimator $\hat{\theta}_n$, this estimator is even identical to the ordinary maximum likelihood estimator if no censoring is present. So, the approach can truly be regarded as an extension of the maximum likelihood method to censored variables.

The concept of state sets is instrumental in the proofs of the theorem because it ensures that each realisation y only contributes to the likelihood in one way:[5] the state sets are a disjoint decomposition of all possible realisations

[5]This is a rule of thumb sometimes used to "check" likelihoods under censoring.

(see Appendix A). This is not necessarily the case if state sets are not used for integration. The state conditional likelihood estimator is based on the same state sets as the likelihood estimator and similar proofs can be used to show that it is also consistent. On the other hand, it does not employ the information which is embodied in the probability of observing a particular state and is thus not root-n asymptotically efficient (see Appendix B.3). Also, its contributions are more difficult to calculate as $P(y \in W_s)$ needs to be computed and thus additional integration is required.

4 Applications

The developed framework covers a large range of censoring problems. This section reconsiders some classical censoring problems, derives the likelihood, and compares it with likelihood functions that were used for the respective problem by other authors. In the end, we construct the likelihood for different variations of the moving company problem.

Recall the simple tobit model in which a (one-dimensional) realisation y_1 cannot be observed when it is below zero. Thus the visibility sets are $V_1 = \{y_1 | y_1 \geq 0\}$ and $\bar{V}_1 = \{y_1 | y_1 < 0\}$. The model has only two states: state $s = 0$ with state set $W_0 = \bar{V}_1$ and state $s = 1$ with state set $W_1 = V_1$. The invisibility and visibility operators yield $\bar{v}_0 y = y_1$ and $v_1 y = y_1$. Given a normally distributed Y_1 with mean μ, variance σ^2, density $\phi(\cdot|\mu, \sigma^2)$ and cumulative distribution function $\Phi(\cdot|\mu, \sigma^2)$, the calculated contribution for state $s = 0$ is:

$$\tilde{f}_0(\theta) = \int_{W_0} f(y, \theta) d\bar{v}_0 y = \int_{\bar{V}_1} f(y_1, \theta) dy_1 = P(y_1 < 0) = \Phi(0|\mu, \sigma^2),$$

where $\theta = (\mu, \sigma^2)$. If y_1 is observable ($s = 1$), the formula for the contribution yields:

$$\tilde{f}_1(y_1, \theta) = f(y_1, \theta) = \phi(y_1|\mu, \sigma^2).$$

Using the data, the objective function from equation (1) becomes:

$$\prod_{\{i|s_i=0\}} \Phi(0|\mu, \sigma^2) \prod_{\{i|s_i=1\}} \phi(y_{i1}|\mu, \sigma^2)$$

But this is exactly the likelihood which is usually used to obtain the tobit estimator. Consequently, this estimator is a special case of estimator $\hat{\theta}_n$ and inherits its properties. In the case of the tobit estimator, this might not be

very exciting since Amemiya (1973) has already derived its properties. However, for other estimators, the properties of which have not been proven, the method is more useful.

As an example take the tobit type II model as introduced by Amemiya (1984). In this model, there are two components $y = (y_1, y_2)$ which are normally distributed around the means μ_1 and μ_2 with variance-covariance matrix

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix} \text{ so that } \theta = (\mu_1, \mu_2, \sigma_{11}, \sigma_{12}, \sigma_{22}).$$

The realisation y_1 is never observable but y_2 is observable whenever $y_1 > 0$. The visibility set for y_2 is $V_1 = \{(y_1, y_2) | y_1 > 0\}$. There are two relevant states $s = 0$ and $s = 1$ and the corresponding state sets are $W_0 = \bar{V}_1$ and $W_1 = V_1$. The respective contribution for state $s = 0$ is:

$$\tilde{f}_0(\theta) = \int_{W_0} f(y, \theta) d\bar{v}_0 y = \iint_{W_0} f(y_1, y_2, \theta) dy_1 dy_2 = P(y_1 \leq 0) = \Phi(0 | \mu_1, \sigma_{11})$$

while state $s = 1$ contributes:

$$\tilde{f}_1(y_2, \theta) = \int_{W_1} f(y, \theta) d\bar{v}_1 y = \phi(y_2 | y_1 > 0, \mu_1, \mu_2, \Sigma) \cdot P(y_1 > 0).$$

Plugging in the data, we get the following likelihood, which coincides with the objective function given by Amemiya (1984) for the tobit type II estimator:

$$\prod_{\{i | s_i = 0\}} \Phi(0 | \mu_1, \sigma_{11}) \prod_{\{i | s_i = 1\}} \phi(y_{i2} | y_1 > 0, \mu_1, \mu_2, \Sigma) \cdot (1 - \Phi(0 | \mu_1, \sigma_{11})).$$

Thus, the tobit type II estimator is also a special case of the estimator $\hat{\theta}_n$. Unlike for the tobit type I estimator, Amemiya (1984, 1985) gives no proof, why the type II estimator is consistent and root-n asymptotically normal distributed. Since the estimator is a special case of $\hat{\theta}_n$, these properties are now formally ensured by Theorem 1. It can be shown that the likelihood function for the tobit models of type III to V according to Amemiya' s classification (1984) are also special cases of the objective function in (1). Hence, the respective maximisers all have desirable properties (under regularity conditions).

Nelson (1977) examined a censoring problem, which does not fall into the five categories of Amemiya (1984). In this model, there are again two realisations y_1 and y_2 from normally distributed random variables and the same

parameters as in the tobit type II model. This time the second component operates as an unobservable censoring threshold. That means y_1 is observable whenever it is above y_2, the visibility set for y_1 is $V_1 = \{y|y_1 > y_2\}$. Nelson (1977) proposes the following likelihood function:

$$\prod_{\{i|s_i=0\}} \Phi\left(\frac{\mu_2 - \mu_1}{\sigma_{11} + \sigma_{22} - 2\sigma_{12}}\right) \prod_{\{i|s_i=1\}} \int_{-\infty}^{y_{i1}} \phi(y_{i1}, y_{i2}|\mu_1, \mu_2, \Sigma) dy_2.$$

Nelson does not prove why this likelihood is valid but he conducts a small simulation study which suggests that its maximiser has the standard asymptotic properties. Again, the properties can be formally affirmed if the proposed likelihood function coincides with the likelihood function in (1). To check this, compute the contribution for the state $s = 0$, in which y_1 is not observable:

$$\tilde{f}_0(\theta) = \int_{V_0=\bar{V}_1} f(y, \theta) d\bar{v}_0 y = \int_{-\infty}^{\infty} \int_{y_1}^{\infty} \phi(y_1, y_2) dy_2 dy_1 = \Phi\left(\frac{\mu_2 - \mu_1}{\sigma_{11} + \sigma_{22} - 2\sigma_{12}}\right).$$

The contribution for the state where y_1 can be observed ($s = 1$) is:

$$\tilde{f}_1(y_1, \theta) = \int_{V_1} f(y, \theta) d\bar{v}_1 y = \int_{-\infty}^{y_1} \phi(y_1, y_2|\mu_1, \mu_2, \Sigma) dy_2.$$

This implies that Nelsons objective function is indeed a likelihood function and that the estimator has the standard asymptotic properties.

Recent examples for likelihood functions, which are embedded in the current framework, are the multi-variate analysis of milk product purchases by Cornick et al. (1994) and the estimation of a productivity distribution from bonus data by Ferrall & Shearer (1999).

Not only does the framework accommodate various classical approaches, it also allows to generate estimators for new censoring problems. Let us return to the two moving companies. We already know that the mean of the observed prices is inconsistent because high prices are not observed. Often, robust methods can be successfully applied to censoring problems (for an overview of the use of quantile methods in censoring models see Fitzenberger, 1997). Here, however, taking the median of the observed prices is problematic because it is only known for a company when it submitted the lowest price in more than half of the cases. On the other hand, if we are willing to

make distributional assumptions, we can employ the proposed methodology. If the price of company j was below that of company k, the contribution is:

$$\tilde{f}_j(y_j, \theta) := \int_{y_j}^{\infty} f(y_j, y_k, \theta) dy_k,$$

where θ is a vector summarising the parameters of interest (e.g. the mean prices). If companies are not colluding, the price distributions are independent for fixed characteristics, the density can be rewritten as $f(y_1, y_2, \theta) = f_1(y_1, \theta) f_2(y_2, \theta)$, and the contribution simplifies to:

$$\tilde{f}_j(y_j, \theta) = (1 - F_k(y_j, \theta))\, f_j(y_j, \theta).$$

What if there are q bidding moving companies and we can only observe the identity j of the company with the lowest price and its price y_j? Then, the contribution is:

$$\tilde{f}_j(y_j, \theta) := \int_{y_j}^{\infty} \cdots \int_{y_j}^{\infty} f(y, \theta) dy_1 \cdots dy_{j-1} dy_{j+1} \cdots dy_q.$$

Given indepently distributed price bids, the contribution becomes:

$$\tilde{f}_j(y_j, \theta) = \prod_{k \neq j} (1 - F_k(y_j, \theta))\, f_j(y_j, \theta).$$

What if the reimbursing firm runs a second-price auction between two firms, where the moving company with the lower bid wins the contract but gets paid the higher bid? If only the reimbursed price is recorded, the visibility set for company j is $V_j := \{y | y_j > y_k\}$. Accordingly, the contribution of an observation where k wins the bid and j's price is observable is: $\int_{-\infty}^{y_j} f(y, \theta) dy_k$. All these examples should convince the reader that the estimation approach is rather versatile and that it can be applied to a large range of problems if the visibility sets are known and measurable.

When we discussed the independence of prices, we already hinted at the possibility that observable explanatory characteristics may be incorporated in the estimation problem.[6] As usually, the parameter of the i-th observation θ_i may be regarded as a function of an underlying common parameter and these explanatory characteristics – for more details see Appendix C.

[6]The explanatory characteristics may be censored random variables themselves. For this case, Donald , Paarsch (1993) examine how their distribution can be retrieved.

5 Conclusion

This article proposes a method to obtain likelihood estimators when there is censoring. The chosen approach has several virtues.

First, it is general. The method can be applied to an almost arbitrary censoring problem. For many standard censoring problems, the resulting estimator then coincides with known estimators. Thus, it can be regarded as a generalisation of these estimators. In addition, it embedds maximum likelihood estimation when there is no censoring as a special case. Since the estimator has the typical asymptotical properties of maximum likelihood estimators, the method can be seen as a natural extension of the maximum likelihood method to the problem of censored variables.

Second, likelihood estimation of censored variables becomes more accessible for applied econometricians. At the moment, econometricians have to rely on experience, conceived wisdom and folk theorems to find a valid likelihood when being confronted with a censoring problem. The present approach provides an explicit rule. In order to increase the accessibility further, it would be helpful to implement the method in a computer programme; the generality of the approach could be maintained by calculating contributions on the basis of Monte-Carlo simulations.

Third, the approach renders likelihood estimation with censoring more transparent and clarifies why asymptotic properties hold. As Davidson and MacKinnon point out (1993, p. 539), likelihoods for censoring problems sometimes appear "fishy" to the uninitiated, because they are neither (Lebesgue-)density nor probability functions but seemingly odd mixtures. While Amemiya (1973) has proven that the mixture employed in the tobit type I estimator yields the usual asymptotic properties, the approach here identifies a whole class of such mixtures which are valid likelihood functions.

Fourth, it provides a structured way to describe censoring problems. In order to write down the likelihood, the censoring problem needs to be put in a particular structure. This imposes a discipline on the presentation of the model which helps the reader (and possibly also the modeller) to better understand the nature of the censoring problem.

References

Amemiya, T. (1973). Regression analysis when the dependent variable is truncated normal. *Econometrica*, *41*(6), 997–1016.

Amemiya, T. (1984). Tobit models: A survey. *Journal of Econometrics*, *24*, 3–61.

Amemiya, T. (1985). *Advanced Econometrics*. Oxford: Basil Blackwell.

Attanasio, O. P. (2000). Consumer durables and inertial behaviour: Estimation and aggregation of (s,s) rules for automobile purchases. *Review of Economic Studies*, *67*(4), 667–696.

Cornick, J., Cox, T., Gould, B. W. (1994). Fluid milk purchases: A multivariate tobit analysis. *American Journal of Agricultural Economics*, *76*(1), 74–82.

Davidson, R., MacKinnon, J. G. (1993). *Estimation and Inference in Econometrics*. Oxford: Oxford University Press.

Donald, S. G., Paarsch, H. J. (1993). Piecewise pseudo-maximum likelihood estimation in empirical models of auctions. *International Economic Review*, *34*(1), 121–148.

Ferrall, C., Shearer, B. (1999). Incentives and transaction costs within the firm: Estimating an agency model using payroll records. *Review of Economic Studies*, *66*, 309–338.

Fitzenberger, B. (1997). A guide to censored quantile regression. In G. Maddala , C. Rao (Eds.), *Handbook of Statistics, Vol 15: Robust Inference* (pp. 405–435). Amsterdam: Elsevier. Reprint.

Gourieroux, C., Monfort, A., Renault, E., Trognon, A. (1987). Generalised residuals. *Journal of Econometrics*, *34*, 5–32.

Little, R. J. A., Rubin, D. B. (1987). *Statistical Analysis with Missing Data*. New York: John Wiley.

Nelson, F. D. (1977). Censored regression models with unobserved stochastic censoring thresholds. *Journal of Econometrics*, *6*(3), 309–328.

Newey, W. K., McFadden, D. (1996). Large sample estimation and hypothesis testing. In Z. Griliches & M. D. Intriligator (Eds.), *Handbook of Econometrics*, volume 4 (pp. 2111–2241). North-Holland.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*, 581–592.

Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica*, *26*, 24–36.

A Collection of states decomposes $\mathbb{R}^q$

Lemma 1. *$\{W_s\}_{s\geq 0}$ is a disjoint decomposition of $\mathbb{R}^q$.*

Proof. Part 1: $\bigcup_s W_s = \mathbb{R}^q$: Take any $y \in \mathbb{R}^q$. Call the respective visibility state s'. For this visibility state s', it must hold that $y \in V_j$ if y_j observable and $y \in \bar{V}_j$ if y_j not observable. Thus, $y \in \bigcap_{\{j|j \text{ visible in } s'\}} V_j \cap \bigcap_{\{j|j \text{ not visible in } s'\}} \bar{V}_j$, which by definition is equivalent to $y \in W_{s'}$ and hence $y \in \bigcup_s W_{s'}$. Overall, $y \in \mathbb{R}^q \Rightarrow y \in \bigcup_s W_{s'}$ and hence $\bigcup_s W_s \supseteq \mathbb{R}^q$. On the other hand, $W_{s'} \subseteq \mathbb{R}^q$ for all s' and thus $\bigcup_s W_s \subseteq \mathbb{R}^q$. Together we get: $\bigcup_s W_s = \mathbb{R}^q$.

Part 2: $W_s \cap W_{s'} = \emptyset$ for $s \neq s'$. If states differ ($s \neq s'$), it follows that there exists a component k which is visible in one state but not in the other. Without loss of generality, let s be the state where it is visible, then $W_s \subseteq V_k$ and $W_{s'} \subseteq \bar{V}_k$ by the definition of the visibility set. Hence, $(W_s \cap W_{s'}) \subseteq (V_k \cap \bar{V}_k)$. But the latter is by construction the empty set: $V_k \cap \bar{V}_k = \emptyset$. Thus, the visibility sets must be disjoint: $(W_s \cap W_{s'}) = \emptyset$. □

B Properties of the maximiser (Theorem 1)

In this section, we analyse the properties of the estimator defined by (1). The first part deals with consistency, the second part with asymptotic normality, and the third part with efficiency.

B.1 Consistency

In this section, consistency is proven by using a standard result on the consistency of M-estimators. As $\hat{\theta}_n$ is the maximiser of any monotone transformation of the objective function in (1), one can alternatively work with the following objective function:

$$Q_n(\theta) := \frac{1}{n} \sum_{i=1}^{n} \log \tilde{f}_{s_i}(v_{s_i}\mathbf{y}_i, \theta) \tag{2}$$

and use the machinery of M-estimation to determine the properties of the maximiser and in particular to check whether it is consistent. To do so we employ the following standard result (see e.g. Amemiya, 1985 or Newey and McFadden, 1996):

Theorem 2 (Consistency of M-estimators). *If there are measurable functions $Q_n(\theta)$ and a non-stochastic function $Q_0(\theta)$ such that (i) $Q_n(\theta)$ converges uniformly in probability to $Q_0(\theta)$, (ii) $Q_0(\theta)$ is continuous, (iii) $Q_0(\theta)$ is uniquely maximised at θ_0, and (iv) the parameter space is compact, then $\hat{\theta}_n$ is consistent for θ_0: $\hat{\theta}_n \xrightarrow{p} \theta_0$.*

The rest of this section will be devoted to find primitive conditions for (i) to (iii) to hold.

The objective function needs to converge to the non-stochastic function $Q_0(\theta)$. Before we are able to find the maximiser of the limit of the objective function, we must ensure that this function exists and is finite. Thus, we assume that the expected value of the logarithm of the density exists:

$$\mathrm{E}|\log f(Y,\theta)| < \infty. \tag{FIN}$$

From this condition on the general density, we can conclude the finiteness of the expectation of the contributions of state s.[7]

Lemma 2. *From* (FIN) *follows:*

$$E_{Y|S}\left|\log(\tilde{f}_s(v_s Y),\theta)\right| < \infty.$$

Proof. By the mean-value theorem for integrals, we can rewrite the contribution of state s:

$$\tilde{f}_s(v_s y) = \int_{W_s} f(h(v_s y, \bar{v}_s \tilde{y}),\theta) d\mu(\bar{v}_s y) = f(h(v_s y,\zeta)) \text{ for some } \zeta \in W_s, \tag{3}$$

where h is the appropriate permutation of the values such that y_1 is the first argument of $f(\cdot)$ and y_n is the last. Now, take condition (FIN) and rewrite it.

$$\begin{aligned}
\infty &> \mathrm{E}\left|\log\left\{f(Y,\theta)\right\}\right| \\
&= \mathrm{E}_S\left[\mathrm{E}_{Y|S}\left|\log\left\{f(Y,\theta)\right\}\right|\right] \\
&= \mathrm{E}_S\left[\mathrm{E}_{\bar{v}_s Y, v_s Y|S}\left|\log\left\{f(Y,\theta)\right\}\right|\right] \\
&= \mathrm{E}_S\left[\mathrm{E}_{\bar{v}_s Y|S}\left[\mathrm{E}_{v_s Y|\bar{v}_s Y,S}\left|\log\left\{f(Y,\theta)\right\}\right|\right]\right].
\end{aligned}$$

[7] Note that the states s are generated by a well-defined random variable S because W_s is $\mu-$measurable.

This implies:

$$\begin{aligned} & \forall(\bar{v}_S y):\ \mathrm{E}_{(v_S Y)|(\bar{v}_S Y),S}\left|\log\left\{f(h(v_S Y,(\bar{v}_S Y)),\theta)\right\}\right|<\infty \\ \Rightarrow\ & \mathrm{E}_{(v_s Y)|\zeta,S}\left|\log\left\{f(h(v_s Y,\zeta),\theta)\right\}\right|<\infty. \end{aligned}$$

Together with (3), we get:

$$\infty > \mathrm{E}_{(v_s Y)|\zeta,S}\left|\log\left\{f(h(v_s Y,\zeta),\theta)\right\}\right| = \mathrm{E}_{(v_s Y)|S}\left|\log\left\{\tilde{f}_s(v_s Y,\theta)\right\}\right|.$$

□

Using the finiteness and the law of large numbers, we can determine the probability limit of the objective function when the number of observations tends to infinity.

Proposition 1 (Convergence). *Given condition* (FIN)*, $Q_n(\theta)$ converges uniformly in probability to*

$$Q_0(\theta) = \sum_s \int_{W_s} \log\left\{\tilde{f}_s(v_s y,\theta)\right\}\ \tilde{f}_s(v_s y,\theta_0)\, d\mu(v_s y). \tag{4}$$

Proof. Since observational units are drawn independently, $Q_n(\theta)$ is the mean of independent random variables. The law of the large numbers applies, and the mean converges in probability to its expected value

$$\begin{aligned} & \mathrm{E}_{Y|\theta_0}\left[\log\left\{\tilde{f}_S(v_S Y,\theta)\right\}\right] \\ = \ & \mathrm{E}_{S|\theta_0}\left[\mathrm{E}_{Y|S,\theta_0}\left[\log\left\{\tilde{f}_S(v_S Y,\theta)\right\}\right]\right],\ \text{which is finite by (FIN)} \\ = \ & \sum_{s=0}^{2^q-1} P(S=s,\theta_0)\int_{W_s}\log\left\{\tilde{f}_s(v_s y,\theta)\right\} f_s(v_s y,\theta_0) d\mu(v_s y) \\ = \ & \sum_{s=0}^{2^q-1}\int_{W_s}\log\left\{\tilde{f}_s(v_s y,\theta)\right\}\tilde{f}_s(v_s y,\theta_0) d\mu(v_s y). \end{aligned}$$

□

Next, we have to ensure continuity of the limiting objective function $Q_0(\theta)$ so that stochastic convergence of the argument leads to stochastic convergence of the values of the function.

Proposition 2 (Continuity). *If*

$$f(y,\theta)\ \textit{is continuous in}\ \theta \quad \text{(CON)},$$

$Q_0(\theta)$ is continuous in θ.

Proof. If $f(y,\theta)$ is continuous in θ, $\tilde{f}_s(v_s y,\theta) = \int_{W_s} f(y,\theta)d\mu(\bar{v}_s y)$ is continuous in θ, and so is $Q_0(\theta)$. □

Like in the case of maximum likelihood estimation, it must be possible to extract the desired information about parameters from the observations. Two *different* parameter values which generate the *same* observations cannot be distinguished. In other words it must be possible to identify the parameter from the observations. We define the statistical model to be *identified* if and only if

$$\forall\theta \neq \theta' \,\exists s\, P(S=s) > 0 :\; f_s(v_s y,\theta) \neq f_s(v_s y,\theta'). \tag{ID}$$

In other words, there must exist at least one state under which differences in the parameter translate into differences in the conditional density. Similarly, to the identification condition in maximum likelihood estimation, this condition may be difficult to verify. The next result proves that the parameter is indeed uniquely determined when the condition can be verified. The proof is very similar to the proof of the uniqueness of the maximiser of the limiting objective function when working with ordinary likelihoods.

Proposition 3 (Unique maximiser). *Under* (ID) *and* (FIN), $Q_0(\cdot)$ *is uniquely maximised at the true parameter* θ_0.

Proof. Consider the difference between the limiting objective function evaluated at the true parameter, $Q_0(\theta_0)$, and at a different parameter $Q_0(\theta)$:

$$\begin{aligned} Q_0(\theta_0) - Q_0(\theta) &= \mathrm{E}_{Y|\theta_0}\left[\log\left\{\tilde{f}_s(v_s y,\theta_0)\right\} - \log\left\{\tilde{f}_s(v_s y,\theta)\right\}\right] \\ &= \mathrm{E}_{Y|\theta_0}\left[-\log\left\{\frac{\tilde{f}_s(v_s y,\theta)}{\tilde{f}_s(v_s y,\theta_0)}\right\}\right] \\ &> \mathrm{E}_{S|\theta_0}\left[-\log\left\{\mathrm{E}_{Y|S,\theta_0}\left[\frac{\tilde{f}_s(v_s y,\theta)}{\tilde{f}_s(v_s y,\theta_0)}\right]\right\}\right], \end{aligned} \tag{5}$$

where the last inequality follows from the strict version of Jensen's inequality for non-constant random variables. By (ID), the expected value is indeed taken over a non-constant random variable. As $\mathrm{E}_{Y|S}\left[\frac{\tilde{f}_s(v_s y,\theta)}{\tilde{f}_s(v_s y,\theta_0)}\right] = 1$, we get $Q_0(\theta_0) - Q_0(\theta) > 0$, and θ_0 is the unique maximum. □

The consistency of the estimator (1) can now be deduced from the conditions (FIN), (CON), and (ID) which together with the propositions 1 to 3 imply that the requirements of the standard consistency theorem are valid.

Theorem 3 (Consistency of the estimator). *If there are measurable functions $Q_n(\theta)$,* (FIN)*,* (CON)*,* (ID) *hold, and the parameter space is compact, then $\hat{\theta}_n \xrightarrow{p} \theta_0$.*

It is simple to construct an alternative to estimator (1) by replacing $\tilde{f}_s(\cdot)$ by the conditional density contributions $f_s(\cdot)$ where the "density" for state $s = 0$ is defined as $f_0(y, \theta) \equiv P(S = 0, \theta)$. Multiplying these conditional density contributions, we get the *state conditional likelihood function* Q_n^{SCL}; the maximiser will be referred to as *state conditional likelihood estimator* or as *SCL-estimator*. All propositions and proofs in this section can be adapted to the state conditional likelihood estimator, so that it is also consistent under (FIN), (CON), (ID). In addition to the evaluation of an integral, which is also necessary to obtain the generalised likelihood, the state conditional likelihood requires the computation of the probability of all states. If there is no closed form for the respective probabilities, this will increase the computational effort substantially. In order to save space, the likelihood estimator $\hat{\theta}_n$ defined in (1), is abbreviated to *L-estimator.*

B.2 Asymptotic normality

Another property of maximum likelihood estimators is their asymptotic normality and efficiency. In this section, conditions are derived under which (1) has these properties. More precisely, we assume that the objective function in (1) has an interior maximum and examine the solution to the first-order condition which results from maximising the objective function. Again, the conditions resemble the respective conditions for maximum likelihood estimators. This is no coincidence, since the proof is based on a standard result for M-estimators (Theorem 4.1.3 in Amemiya, 1985 where assumption B is replaced using Theorem 4.1.5):

Proposition 4 (Asymptotic normality of M-estimators). *If (i) $\hat{\theta}_n$, the maximiser of $Q_n(\cdot)$, is consistent for θ_0, (ii) θ_0 lies in the interior of the parameter space Θ, (iii) Q_n is twice continuously differentiable in an open and convex neighbourhood $\mathcal{N}$ of θ_0, (iv) $\sqrt{n}\,\nabla_\theta Q_n(\theta)|_{\theta=\theta_0} \xrightarrow{d} N(0, J)$, (v) $\nabla_{\theta\theta} Q_n(\theta)|_{\theta=\hat{\theta}_n} \xrightarrow{P} H(\theta_0)$ with $H(\theta)$ finite, non-singular, and continuous at θ_0, then $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, H^{-1}JH^{-1})$.*

Under the assumptions of theorem 3, condition (i) is valid. Condition (ii) ensures that the maximum is not a corner solution and hence that the first derivative of $Q_0(\cdot)$ disappears at θ_0. Subsequently, conditions (iii) to (v) should be replaced by primitive conditions on the density $f(y, \theta)$.

We begin by assuming:

$$f(y,\theta) \text{ is twice continuously differentiable at } \theta_0. \tag{DIFF}$$

Denote the operator which yields the first derivative of a vector-valued function by ∇_θ and use the convention that this operator turns a real-valued component of the function into a $(1,p)$-vector, where the first value is the derivative with respect to θ_1, the second with respect to θ_2 and so forth. Likewise $\nabla_{\theta\theta}$ is the operator which gives the second derivative of a real-valued function, the Jacobian matrix. We want the differentiation operators to be exchangeable with integration which is for example fulfilled if the area over which is integrated does not depend on θ:

$$\begin{aligned} \nabla_\theta \int f(y,\theta) d\mu(y) &= \int \nabla_\theta f(y,\theta) d\mu(y) \\ \nabla_{\theta\theta} \int f(y,\theta) d\mu(y) &= \int \nabla_{\theta\theta} f(y,\theta) d\mu(y), \end{aligned} \tag{EID}$$

where we require the equalities to hold only evaluated in the neighbourhood of $\theta = \theta_0$. We can use the exchangebility to compute the first and second derivative of $\tilde{f}^s(v_s y, \theta)$ with respect to θ at θ_0:

$$\begin{aligned} \nabla_\theta \tilde{f}_s(v_s y,\theta)\Big|_{\theta=\theta_0} &= \int_{W_s} \nabla_\theta f(y,\theta) d\mu(\bar{v}_s y)\Big|_{\theta=\theta_0} \\ \nabla_{\theta\theta} \tilde{f}_s(v_s y,\theta) d\mu(\bar{v}_s y)\Big|_{\theta=\theta_0} &= \int_{W_s} \nabla_{\theta\theta} f(y,\theta) d\mu(\bar{v}_s y)\Big|_{\theta=\theta_0} \end{aligned} \tag{6}$$

Next, define:

$$J(\theta) := \mathrm{E}_{Y|\theta_0}\left[\frac{\nabla_\theta \tilde{f}_S(v_S Y,\theta)' \nabla_\theta \tilde{f}_S(v_S Y,\theta)}{\tilde{f}_S(v_S Y,\theta)\tilde{f}_S(v_S Y,\theta)}\right], \tag{7}$$

where expectations are taken with respect to the true parameter θ_0 and where the prime denotes the transpose of a vector or matrix. Later, it will be proven that $J(\theta)$ is the second derivative of the limit function $Q_0(\theta)$ and thus deserves the letter "J" indicating that it is a Jacobian matrix. Since, we want this second derivative to exist and to be finite at its maximiser θ_0, we suppose that

$$J := J(\theta_0) \text{ exists and is finite.} \tag{EX}$$

Under the defined conditions, it is now possible to calculate the distribution of the first derivative of the objective function:

Proposition 5 (Asymptotic normality of the first derivative). *Under* (DIFF), (FIN), (EX), (EID), *and if* J *defined in* (EX) *is non-singular, then*

$$\sqrt{n}\,\nabla_\theta Q_n(\theta)|_{\theta=\theta_0} \xrightarrow{d} N(0, J).$$

Proof. $\sqrt{n}\,\nabla_\theta Q_n(\theta)|_{\theta=\theta_0}$ can be rewritten as the sum of i.i.d. random variables:

$$\sqrt{n}\,\nabla_\theta Q_n(\theta)|_{\theta=\theta_0} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \underbrace{\frac{\nabla_\theta \tilde{f}_s(v_{s_i}\mathbf{y}_i, \theta)|_{\theta=\theta_0}}{\tilde{f}_s(v_{s_i}\mathbf{y}_i, \theta)|_{\theta=\theta_0}}}_{=:Q^i_\nabla}, \tag{8}$$

and by the central-limit theorem its distribution converges to a normal distribution with mean $\mathrm{E}(\mathrm{Q}^{\mathrm{i}}_\nabla)$ and variance-covariance matrix $\mathrm{COV}\left[Q^i_\nabla\right]$. The existence of $\mathrm{E}(\mathrm{Q}^{\mathrm{i}}_\nabla)$ is assured by (EX) and Jensen's inequality, its value is:

$$\begin{aligned} \mathrm{E}(Q^i_\nabla) = &\quad \sum_s P(S=s,\theta_0) \int \frac{\nabla_\theta \tilde{f}_s(v_s y,\theta)|_{\theta=\theta_0}}{\tilde{f}_s(v_s y,\theta_0)} \cdot \frac{\tilde{f}_s(v_s y,\theta_0)}{P(S=s,\theta_0)} d\mu(v_s y) \\ \overset{(\mathrm{EID})}{=} &\quad \sum_s P(S=s,\theta) \cdot \nabla_\theta \underbrace{\int \frac{\tilde{f}_s(v_s y,\theta)}{P(S=s,\theta)} d\mu(v_s y)}_{=1}\Bigg|_{\theta=\theta_0} = 0. \end{aligned} \tag{9}$$

As the expected value is the zero vector, $\mathrm{COV}\left[Q^i_\nabla\right] = \mathrm{E}\left[(Q^i_\nabla)' Q^i_\nabla\right]$. By plugging in Q^i_∇ one immediately gets $\mathrm{COV}\left[Q^i_\nabla\right] = J$, the existence of which is ensured by (EX). □

Assumptions (DIFF), (EX), and (EID) do not only enable us to compute the first but also the limit of the second derivative when it is evaluated at the maximiser:

Proposition 6 (Convergence of the second derivative). *Given* (DIFF), (EX), (EID), *and* $\hat{\theta}_n \to \theta_0$, *it follows that* $\nabla_{\theta\theta} Q_n(\theta)|_{\theta=\hat{\theta}_n} \xrightarrow{P} -J$, *where* J *is positive definite.*

Proof. The second derivative of the objective function with respect to θ is:

$$\nabla_{\theta\theta} Q_n(\theta) = \quad \frac{1}{n} \sum_{i=1}^{n} \underbrace{\frac{\tilde{f}_s(v_s y,\theta) \nabla_{\theta\theta} \tilde{f}_s(v_s y,\theta) - \nabla_\theta \tilde{f}_s(v_s y,\theta)' \nabla_\theta \tilde{f}_s(v_s y,\theta)}{\tilde{f}_s(v_s y,\theta)^2}}_{=:Q^i_{\nabla\nabla}}. \tag{10}$$

By the law of large numbers $Q^i_{\mathbb{W}}$ approaches its expected value:

$$\begin{aligned}
\mathrm{E}(Q^i_{\mathbb{W}}(\theta)) = & \quad \mathrm{E}\left(\frac{\nabla_{\theta\theta}\tilde{f}_s(v_s y,\theta)}{\tilde{f}_s(v_s y,\theta)}\right) - \mathrm{E}\left(\frac{\nabla_\theta \tilde{f}_s(v_s y,\theta)'\nabla_\theta \tilde{f}_s(v_s y,\theta)}{\tilde{f}_s(v_s y,\theta)^2}\right) \\
= & \quad \sum_s P(S=s,\theta_0)\int \nabla_{\theta\theta}\tilde{f}_s(v_s y,\theta)d\mu(v_s y)\cdot\frac{1}{P(S=s,\theta_0)} - J(\theta) \\
\overset{\text{(EID)}}{=} & \quad \sum_s P(S=s,\theta_0)\underbrace{\nabla_{\theta\theta}\int f(v_s y,\theta)d\mu(v_s y)}_{=0} - J(\theta) = -J(\theta).
\end{aligned} \tag{11}$$

As this expected value is continuous in θ around θ_0, we can use Theorem 4.1.5 in Amemiya (1985) to conclude that from $\hat{\theta}_n \xrightarrow{p} \theta_0$ it follows that $\mathrm{E}[Q^i_{\mathbb{W}}(\hat{\theta}_n)] \xrightarrow{p} \mathrm{E}\left[Q^i_{\mathbb{W}}(\theta_0)\right]$. So overall, we get $\nabla_{\theta\theta}Q_n(\theta)|_{\theta=\hat{\theta}_n} \xrightarrow{p} \mathrm{E}\left[Q^i_{\mathbb{W}}(\theta_0)\right] = -J$. As $-J$ is the second derivative of the objective function evaluated at a unique and interior maximum, it must be negative definite. So, J must be positive definite. □

Using theorem 3 and propositions 4 to 6, we can state:

Theorem 4 (Asymptotic normality of the L-estimator). *If* (ID), (EX), (FIN), (EID), *and* (DIFF) *hold, and* θ_0 *is in the interior of the compact parameter space* Θ, *then* $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, J^{-1})$.

Again, the results for the state conditional estimator $\hat{\theta}^{\mathrm{SCL}}_n$ can be derived by replacing the contributions $\tilde{f}_s(\cdot)$ by the conditional density contributions $f_s(\cdot)$ and the generalised likelihood Q_n by the state conditional likelihood Q^{SCL}_n in propositions 4 to 6. This yields the following result.

Corollary 1 (Asymptotic normality of the SCL-estimator). *If* (ID), (EX), (FIN), (EID), *and* (DIFF) *hold, and* θ_0 *is in the interior of the compact parameter space* Θ, *then* $\sqrt{n}(\hat{\theta}^{SCL}_n - \theta_0) \xrightarrow{d} N(0, (\Sigma^{SCL})^{-1})$, *where*

$$\Sigma^{SCL} := E_{Y|\theta_0}\left[\frac{\nabla_\theta f_s(v_s Y,\theta)'\nabla_\theta f_s(v_s Y,\theta)}{f_s(v_s Y,\theta)f_s(v_s Y,\theta)}\right]. \tag{12}$$

The asymptotic variance-covariance matrix of the derivative of the state dependent likelihood function Σ^{SCL} is identical to the variance-covariance matrix for the generalised likelihood J, if the probabilities of the various states do not depend on the unknown parameter: $\nabla_\theta P(S=s,\theta) = 0 \Rightarrow \Sigma^{\mathrm{SCL}} = J$. However, generally the two matrices will be different. Then, the respective root-n estimators have different asymptotic properties. Is one of the estimators preferable because it has a smaller asymptotic variance-covariance matrix?

B.3 Asymptotic Efficiency

To show that root-n times the L-estimator is asymptotically efficient, we proceed in two steps. First, we determine the Cramer-Rao lower bound for the class of censoring problems under consideration. This yields the "smallest" variance-covariance matrix which can be attained using the available (censored) information. Second, we observe that the asymptotic variance-covariance matrix of the root-n L-estimator coincides with this lower bound. Hence, the root-n L-estimator must be asymptotically efficient.

Proposition 7 (Cramer-Rao lower bound). *In the censoring problem described above and given that* (ID)*,* (EX)*,* (FIN)*,* (EID)*, and* (DIFF) *hold, the asymptotic variance-covariance matrix* $\lim_{n\to\infty} COV[\sqrt{n}T]$ *of any asymptotically unbiased estimator* $\sqrt{n}T$ *for* θ_0 *is larger or equal to* J *according to the Löwner ordering:*

$$\forall x: \lim_{n\to\infty} x' COV\left[\sqrt{n}T\right] x \geq x' J^{-1} x.$$

Proof. Denote the observable sample by $v_s y := (v_{s_1} y_{1.}, \ldots, v_{s_n} y_{n.})$ Let $\sqrt{n}T(\cdot)$ be an asymptotically unbiased estimator for the true parameter: $\mathrm{E}\left[T(v_S Y)\right] = \theta_0$, for $n \to \infty$. Then, write out the expected value using the independence of the observations:

$$\begin{aligned}
\theta_0 &= \lim_{n\to\infty} \mathrm{E}_{Y|\theta_0}\left[T(v_S Y)\right] \\
&= \lim_{n\to\infty} \sum_{s_1=0}^{2^q-1} P(S=s_1,\theta_0) \cdots \sum_{s_n=0}^{2^q-1} P(S=s_n,\theta_0) \cdot \\
&\qquad \cdot \int_{W_{s_1}} \cdots \int_{W_{s_n}} T(v_{s_1} y_1, \ldots, v_{s_n} y_n) \prod_{i=1}^{n} f_{s_i}(v_{s_i} y_i, \theta_0)\, d\mu(v_{s_1} y_1) \cdots d\mu(v_{s_n} y_n).
\end{aligned}$$

Using relationship $\tilde{f}_s(v_s y, \theta) = P(S=s) f_s(v_s y, \theta)$, this simplifies to:

$$\theta|_{\theta=\theta_0} = \lim_{n\to\infty} \sum_{s_1=0}^{2^q-1} \cdots \sum_{s_n=0}^{2^q-1} \int_{W_{s_1}} \cdots \int_{W_{s_n}} T(v_s y) \prod_{i=1}^{n} \tilde{f}_{s_i}(v_{s_i} y_i, \theta_0)\, d\mu(v_{s_1} y_1) \cdots d\mu(v_{s_n} y_n). \tag{13}$$

Now, take the derivative with respect to θ on both sides:

$$I = \lim_{n\to\infty} \sum_{s_1=0}^{2^q-1} \cdots \sum_{s_n=0}^{2^q-1} \int_{W_{s_1}} \cdots \int_{W_{s_n}} T(v_s y) \nabla_\theta \prod_{i=1}^{n} \tilde{f}_{s_i}(v_{s_i} y_i, \theta_0)\, d\mu(v_{s_1} y_1) \cdots d\mu(v_{s_n} y_n). \tag{14}$$

Next, consider the following sophisticated expression for a zero matrix:

$$\theta_0 \, \nabla_\theta I = \theta_0 \, \nabla_\theta \sum_{s_1=0}^{2^q-1} \cdots \sum_{s_n=0}^{2^q-1} \int_{W_{s_1}} \cdots \int_{W_{s_n}} \nabla_\theta \prod_{i=1}^{n} \tilde{f}_{s_i}(v_{s_i} y_i, \theta_0) \, d\mu(v_{s_1} y_1) \cdots d\mu(v_{s_n} y_n). \tag{15}$$

Subtracting this sophisticated zero from the right-hand side in (14) yields:

$$I = \lim_{n\to\infty} \sum_{s_1=0}^{2^q-1} \cdots \sum_{s_n=0}^{2^q-1} \int_{W_{s_1}} \cdots \int_{W_{s_n}} (T(v_s y) - \theta_0) \, \nabla_\theta \tilde{f}_{s_\cdot}(v_s y, \theta_0) \, d\mu(v_{s_1} y_1) \cdots d\mu(v_{s_n} y_n),$$

where $\tilde{f}_{s_\cdot}(v_s y, \theta_0) := \prod_{i=1}^{n} \tilde{f}_{s_i}(v_{s_i} y_i, \theta_0)$. By multiplying with and dividing by

$$\sqrt{n} f_{s_\cdot}(v_s y, \theta_0) := \prod_{i=1}^{n} f_{s_i}(v_{s_i} y_i, \theta_0),$$

we get:

$$I = \lim_{n\to\infty} \sum_{s_1=0}^{2^q-1} \cdots \sum_{s_n=0}^{2^q-1} \int_{W_{s_1}} \cdots \int_{W_{s_n}} \sqrt{n} \, (T(v_s y) - \theta_0) \, \frac{\nabla_\theta \tilde{f}_{s_\cdot}(v_s y, \theta_0)}{\sqrt{n} f_{s_\cdot}(v_s y, \theta_0)} f_{s_\cdot}(v_s y, \theta_0) d\mu(v_s y)$$

$$= \lim_{n\to\infty} \mathrm{E}\left[\mathcal{T}\mathcal{W}\right], \tag{16}$$

where $\mathcal{T} = \sqrt{n} \, (T(v_s y) - \theta_0)$ and $\mathcal{W} = \frac{\left(\nabla_\theta \tilde{f}_{s_\cdot}(v_s y, \theta_0)\right)'}{\sqrt{n} f_{s_\cdot}(v_s y, \theta_0)} = \frac{\left(\nabla_\theta \log\left\{\tilde{f}_{s_\cdot}(v_s y, \theta_0)\right\}\right)'}{\sqrt{n}}$.
Next, write the complete asymptotic variance-covariance matrix of $(\mathcal{T}, \mathcal{W})$.

$$\lim_{n\to\infty} \left(\mathrm{E}\left[\begin{pmatrix} \mathcal{T}\mathcal{T}' & \mathcal{T}\mathcal{W}' \\ \mathcal{W}\mathcal{T}' & \mathcal{W}\mathcal{W}' \end{pmatrix} \right] - \begin{pmatrix} \mathrm{E}\left[\mathcal{T}\right] \mathrm{E}\left[\mathcal{T}\right]' & \mathrm{E}\left[\mathcal{T}\right] \mathrm{E}\left[\mathcal{W}\right]' \\ \mathrm{E}\left[\mathcal{W}\right] \mathrm{E}\left[\mathcal{T}\right]' & \mathrm{E}\left[\mathcal{W}\right] \mathrm{E}\left[\mathcal{W}\right]' \end{pmatrix} \right).$$

Recall that $\sqrt{n}T$ is an asymptotically unbiased estimator such that $\lim_{n\to\infty} \mathrm{E}\left[\mathcal{T}\right]$ is a zero vector of length q. By writing out $\mathrm{E}\left[\mathcal{W}\right]$ and exchanging the order of integration and differentiation, it can be shown that $\mathrm{E}\left[\mathcal{W}\right]$ is also a zero vector of length q. Thus, the subtracted matrix cancels and the asymptotic variance-covariance matrix is:

$$\lim_{n\to\infty} \mathrm{COV}\left[\mathcal{T}\mathcal{W}\right] = \lim_{n\to\infty} \mathrm{E}\left[\begin{pmatrix} \mathcal{T}\mathcal{T}' & \mathcal{T}\mathcal{W}' \\ \mathcal{W}\mathcal{T}' & \mathcal{W}\mathcal{W}' \end{pmatrix} \right].$$

Next, note that $\mathrm{E}\left[\mathcal{T}\mathcal{T}'\right] = \mathrm{COV}\left[\mathcal{T}\right] = \mathrm{COV}\left[\sqrt{n}(T - \theta_0)\right] = \mathrm{COV}\left[\sqrt{n}T\right]$, while

$$\mathrm{E}\left[\mathcal{W}\mathcal{W}'\right] = \frac{1}{n} \sum_{i=1}^{n} \mathrm{E}\left[\nabla_\theta \log\left\{\tilde{f}_{s_i}(v_s y, \theta_0)\right\}' \nabla_\theta \log\left\{\tilde{f}_{s_i}(v_s y, \theta_0)\right\} \right] = J.$$

and $\mathrm{E}[\mathcal{T}\mathcal{W}'] = I$. So overall, we get:

$$\lim_{n\to\infty} \mathrm{COV}[\mathcal{T}\mathcal{W}] = \lim_{n\to\infty} \begin{pmatrix} \mathrm{COV}[\sqrt{n}T] & I \\ I & J \end{pmatrix}. \tag{17}$$

Being a variance-covariance matrix, this expression must be positive semi-definite, so in particular

$$\forall a \quad (a', -a'J^{-1}) \lim_{n\to\infty} \begin{pmatrix} \mathrm{COV}[\sqrt{n}T] & I \\ I & J \end{pmatrix} \begin{pmatrix} a \\ -J^{-1}a \end{pmatrix} \geq 0.$$

If we multiply out this inequality, we get the result:

$$\lim_{n\to\infty} \forall a \quad a' \left(\mathrm{COV}\left[\sqrt{n}T\right] - J^{-1}\right) a \geq 0.$$

□

Proposition 7 gives us the lower bound on the variance-covariance matrix. Because this bound is asymptotically attained by the root-n estimator, we can conclude immediately:

Corollary 2 (Asymptotic efficiency). *In the censoring problem described above and given that* (ID), (EX), (FIN), (EID), *and* (DIFF) *hold, the root-n estimator* $\sqrt{n}\hat{\theta}_n$ *is asymptotically efficient.*

The L-estimator is thus superior to the SCL-estimator in the sense that its root-n estimator has a lower asymptotic variance-covariance matrix. In other words, the L-estimator makes better use of the available information.

C A remark on censored regression

In many applications, observational units will differ by observable characteristics X_i which have an effect on the distribution of Y. To allow for this in our modelling framework, we suppose that the formerly fixed θ is an individual parameter which results from the interplay of observable characteristics X_i with a fixed parameter β: $\theta_i = g(\beta, X_i)$.

Since observational units are drawn randomly, the observable characteristics X_i can be modelled by a random variable. The joint density of Y and X can be decomposed: $f_{Y,X}(y, g(\beta, x)) = f_{Y|X}(y|g(\beta, x)) \cdot f_X(x)$. Accordingly, the contribution of a particular state s becomes:

$$\hat{f}_s(v_s y, \beta, x) = \underbrace{\int_{W_s} f(y|\beta, x) d\mu(\bar{v}_s y)}_{=:\tilde{f}(v_s y|\beta, x)} \cdot f_X(x),$$

and thus the logarithmitised objective function is:

$$Q_n(\theta) = \sum_{i=1}^{n} \log\left(\tilde{f}_s(v_{s_i}\mathbf{y}_i|\beta, x_i)\right) + \sum_{i=1}^{n} \log\left(f_X(x_i)\right). \tag{18}$$

As the last term does not change in β, it can be ignored when maximising. So we are left with an objective function which closely resembles the objective function from formula (2) which we analysed in the preceding sections. All conditions, proofs, and theorems can be adapted to this new objective function by replacing $f(y, \theta)$ by $f(y|\beta, x)$ and $\tilde{f}_s(v_s y, \theta)$ by $\tilde{f}_s(v_s y|\beta, x)$, and requiring the respective statement to hold for all x. Additionally, one needs $\int \log(f_X(x)) f_X(x) dx < \infty$, to ensure finiteness of the limiting objective function. The identifiability condition becomes:

$$\forall \beta \neq \beta' \, \exists s, \mathcal{X} : \, P_X(x \in \mathcal{X}) > 0 : \, \tilde{f}_s(v_s y, \beta, x) \neq \tilde{f}_s(v_s y, \beta', x). \tag{ID'}$$

For the linear case $g(X, \beta) = X\beta$ and given (ID), this is fulfilled if X has full rank with positive probability.

On the asymptotic normality result the introduction of X has no effect: all proofs are based on derivatives of the objective function with respect to the parameter. Since the second sum in the objective function is independent of the parameter β, it cancels when taking derivatives.