Worked Examples for Extremum Estimation

Jörg Stoye

April 7, 2025

1 Estimating a Poisson Distribution

This is a somewhat easy example that we will use to work through application of the ML theorems in detail. (It also reproduces an old exam question.) Let Y be distributed i.i.d. Poisson with true parameter value $\lambda_0 > 0$. Recall the Poisson distribution has probability mass function $\Pr(Y = k) = \lambda^k e^{-\lambda}/k!$, and mean and variance λ.

It's intuitively obvious that we want to estimate λ by $\bar{y}$, and the Lindeberg-Lévy CLT immediately yields $\sqrt{n}(\bar{Y} - \lambda_0) \overset{d}{\to} N(0, \lambda_0)$. To analyze this as an example of Maximum Likelihood, write

$$\begin{aligned}
Q_n(\lambda) &= \frac{1}{n}\sum_i \left(y_i \log\lambda - \lambda - \log\left(y_i!\right)\right) \\
\frac{\partial Q_n(\lambda)}{\partial\lambda} &= \frac{1}{n}\sum_i \left(\frac{y_i}{\lambda} - 1\right) \\
\frac{\partial^2 Q_n(\lambda)}{\partial\lambda^2} &= -\frac{1}{n}\sum_i \frac{y_i}{\lambda^2} \\
Q(\lambda) &= \mathbb{E}\left(Y\log\lambda - \lambda - \log\left(Y!\right)\right) \\
\frac{\partial Q(\lambda)}{\partial\lambda} &= \mathbb{E}\left(\frac{Y}{\lambda} - 1\right) \\
\frac{\partial^2 Q(\lambda)}{\partial\lambda^2} &= -\mathbb{E}\left(\frac{Y}{\lambda^2}\right).
\end{aligned}$$

Inspection of $\frac{\partial^2 Q_n(\lambda)}{\partial\lambda^2}$ reveals strict concavity, so that $\hat{\lambda}_{ML}$ is characterized by a FOC. In particular,

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i}{\lambda} - 1\right) \overset{!}{=} 0$$

can be solved for $\hat{\lambda}_{ML} = \bar{y}$. Next, we can apply a consistency theorem for extremum estimators. Pointwise convergence of Q_n to Q is clear. The parameter space is not compact, but Q_n is strictly concave; because λ is a scalar, we can therefore invoke the relatively simple convergence theorem for concave objectives from a homework. (The version using compactness of Θ would work only after bounding λ from above but also from below by a strictly positive number below; else, uniform convergence is simply not true.)

We finally verify conditions for asymptotic normality. 1. was established above, 2. holds if $\lambda > 0$, 3. is evident from the above displays. 4. holds because Y is i.i.d. with expectation and variance λ_0, hence

$$\sqrt{n}\frac{\partial Q_n(\lambda_0)}{\partial \lambda} = \sqrt{n}\frac{1}{n}\sum_i \left(\frac{Y_i}{\lambda_0} - 1\right) \overset{d}{\to} N(0, \lambda_0^{-1}).$$

5. follows by inspection of the display. To verify 6., note that $\frac{\partial^2 Q(\lambda_0)}{\partial \lambda^2} = -\frac{\lambda_0}{\lambda_0^2} = -\lambda_0^{-1}$. 7. follows because $\lambda_0 > 0$. The theorem applies, and we can substitute into its conclusion to find

$$\sqrt{n}\big(\hat{\lambda} - \lambda_0\big) \overset{d}{\to} N\big(0, (-\lambda_0^{-1})^{-1}\lambda_0^{-1}(-\lambda_0^{-1})^{-1}\big) = N(0, \lambda_0)$$

as expected.

Our results do not apply if $\lambda_0 = 0$ because interiority is then violated. Note that in this special case, we have not only a failure of proof but a failure of result: The distribution of both Y and $\bar{Y}$ will be degenerate and completely concentrated at 0. As one might expect from this observation, the above result is also not uniformly true as $\lambda_0 \to 0$; it fails along drifting parameters of the form $\lambda_n = \gamma/\sqrt{n}$. For a red flag that points at this non-uniformity, note that the Hessian, which is really just a second derivative here, approaches degeneracy ("it becomes infinity") as $\lambda_0 \to 0$.

2 Maximum Likelihood Analysis of Linear Regression

2.1 Single Equation

We next reconsider the linear regression model:

$$\begin{aligned} Y &= X'\beta + \varepsilon \\ \varepsilon \mid X &\sim N(0, \sigma^2). \end{aligned}$$

We already analyzed identification. We will maximize the conditional log likelihood

$$\log f(Y|X,\theta) = -\frac{1}{2}\log 2\pi - \frac{1}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\left(Y - X'\beta\right)^2,$$

hence the ML estimator $\hat{\theta} = (\hat{\beta}, \hat{\sigma}^2)$ solves

$$\begin{aligned} \hat{\theta} &= \arg\max_{\theta\in\Theta} Q_n(\beta, \sigma^2), \\ Q_n(\beta, \sigma^2) &= \sum_{i=1}^n \Big(-\frac{1}{2}\log 2\pi - \frac{1}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(y_i - x_i'\beta)^2 \Big). \end{aligned}$$

Likelihood functions frequently have separability properties that make it convenient to solve this by *concentrating out*, that is, by solving for some parameters first and then extremizing the value function over the other parameters. The optimized-out objective function is often called *concentrated* with respect to the optimized-out parameter. In the specific example, let's first find $\hat{\beta}$, which must solve

$$\hat{\beta} = \arg\max_{\beta} \Big\{ -\sum_{i=1}^n (y_i - X_i'\beta)^2 \Big\} = \arg\min_{\beta} \sum_{i=1}^n (y_i - X_i'\beta)^2,$$

so it is the least squares estimator. In a second step, we optimize the concentrated objective function to find that

$$\begin{aligned} \hat{\sigma}^2 &= \arg\max_{\sigma^2} \sum_{i=1}^n \Big(-\frac{1}{2}\log 2\pi - \frac{1}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\big(y_i - x_i'\hat{\beta}\big)^2 \Big) \\ &= \arg\min_{\sigma^2} \Big\{ n\log\sigma^2 + \frac{1}{\sigma^2}\sum_{i=1}^n \big(y_i - x_i'\hat{\beta}\big)^2 \Big\}. \end{aligned}$$

This problem has FOC

$$\begin{aligned} &\frac{n}{\sigma^2} - \frac{1}{(\sigma^2)^2}\sum_{i=1}^n \big(y_i - x_i'\hat{\beta}\big)^2 = 0 \\ \Rightarrow \quad & \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n \big(y_i - x_i'\hat{\beta}\big)^2. \end{aligned}$$

The slight surprise here, if you haven't seen it before, is that the ML estimator $\hat{\sigma}^2$ does not feature the degrees-of-freedom-adjustment, i.e. it is not $s^2 \equiv \frac{1}{n-K}\sum_{i=1}^n \big(y_i - x_i'\hat{\beta}\big)^2$. On the one hand, this shows that s^2 is not as obviously optimal as you might have thought. On the other hand, it is a reminder that ML estimators are, in general, biased (because it remains true that under assumptions maintained here, s^2 is unbiased).

For completeness, we observe that the estimator's asymptotic distribution can also be derived from our development for Maximum Likelihood estimators. In particular, defining $\theta \equiv (\beta', \sigma^2)'$, we have

$$\begin{aligned}
Q(\theta) &= \mathbb{E}\Big(-\frac{1}{2}\log 2\pi - \frac{1}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(Y - X'\beta)^2\Big) \\
\frac{\partial Q(\theta)}{\partial\theta} &= \begin{bmatrix} \frac{1}{\sigma^2}\mathbb{E}\big(X(Y - X'\beta)\big) \\ -\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4}\mathbb{E}(Y - X'\beta)^2 \end{bmatrix} \\
\frac{\partial^2 Q(\theta)}{\partial\theta\partial\theta'} &= \left[\begin{array}{c:c} -\frac{1}{\sigma^2}\mathbb{E}(XX') & -\frac{1}{\sigma^4}\mathbb{E}\big(X(Y - X'\beta)\big) \\ \hdashline -\frac{1}{\sigma^4}\mathbb{E}\big((Y - X'\beta)X'\big) & \frac{1}{2\sigma^4} - \frac{1}{\sigma^6}\mathbb{E}(Y - X'\beta)^2 \end{array}\right].
\end{aligned}$$

As a small exercise to check my algebra, you should be able to convince yourself that: (i) Assuming identification, $\frac{\partial Q(\theta)}{\partial\theta}$ is the zero vector iff evaluated at the true values of β and σ^2; (ii) without identification, it is the zero vector on a linear subspace, i.e. the likelihood has a ridge.

Fortunately, to compute the asymptotic variance, we need to invert the last matrix only if evaluated at the true parameter values, where it much simplifies:

$$\begin{aligned}
\frac{\partial^2 Q(\theta_0)}{\partial\theta\partial\theta'} &= \left[\begin{array}{c:c} -\frac{1}{\sigma^2}\mathbb{E}(XX') & -\frac{1}{\sigma^4}\mathbb{E}\big(X(Y - X'\beta)\big) \\ \hdashline -\frac{1}{\sigma^4}\mathbb{E}\big((Y - X'\beta)X'\big) & \frac{1}{2\sigma^4} - \frac{1}{\sigma^6}\mathbb{E}(Y - X'\beta)^2 \end{array}\right] \\
&= \left[\begin{array}{c:c} -\frac{1}{\sigma^2}\mathbb{E}(XX') & 0 \\ \hdashline 0 & -\frac{1}{2\sigma^4} \end{array}\right] \\
\Longrightarrow -\left(\frac{\partial^2 Q(\theta_0)}{\partial\theta\partial\theta'}\right)^{-1} &= \left[\begin{array}{c:c} \sigma^2\big(\mathbb{E}XX'\big)^{-1} & 0 \\ \hdashline 0 & 2\sigma^4 \end{array}\right].
\end{aligned}$$

We see that $\hat{\beta}$ and $\hat{\sigma}^2$ are asymptotically independent.[1] While we know that the asymptotic variance expression for $\hat{\beta}$ holds under much weaker conditions than we here imposed, the one for $\hat{\sigma}^2$ does not (and in general exists only if ε has a finite fourth moment).

We conclude by verifying the information matrix equality. The top left submatrix equals

$$\sigma^{-4}\mathbb{E}\big(X(Y - X'\beta)^2 X'\big) = \sigma^{-2}\mathbb{E}\big(XX'\big)$$

using the Law of Iterated Expectations and $\mathbb{E}\big((Y - X'\beta)^2|X\big) = \sigma^2$. The bottom right entry equals

$$\frac{1}{4\sigma^4} + \frac{1}{4\sigma^8}\mathbb{E}(Y - X'\beta)^4 - \frac{1}{2\sigma^6}\mathbb{E}(Y - X'\beta)^2 \underset{\theta=\theta_0}{=} \frac{1}{4\sigma^4} + \frac{3}{4\sigma^4} + \frac{1}{\sigma^4} = \frac{1}{2\sigma^4},$$

using that the fourth central moment of the Normal equals $3\sigma^4$. For the off-diagonal elements, we have

$$-\frac{1}{2\sigma^4}\mathbb{E}\big(X(Y - X'\beta)\big) + \frac{1}{2\sigma^6}\mathbb{E}\big(X(Y - X'\beta)^3\big) = 0.$$

[1] In fact, under the normality assumption, they are finite sample independent, which is essential in establishing the exact distributions of t- and F-statistic.

2.2 Multiple Equations

We next derive the estimator with many equations. Write

$$\boldsymbol{Y} = \boldsymbol{\Pi}' X + \boldsymbol{\nu},$$

where

$$\boldsymbol{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_M \end{bmatrix}, \boldsymbol{\Pi}' = \begin{bmatrix} \boldsymbol{\pi}_1' \\ \boldsymbol{\pi}_2' \\ \vdots \\ \boldsymbol{\pi}_M' \end{bmatrix}, \boldsymbol{\nu} = \begin{bmatrix} \nu_1 \\ \nu 2 \\ \vdots \\ \nu_M \end{bmatrix}.$$

To recall, this stacks M equations, each of which regresses a different outcome y_{im} on the same regressors. Recall our assumptions:

- $\{\boldsymbol{Y}, X\}$ is i.i.d.,
- $\mathbb{E}(\boldsymbol{\nu} \otimes X) = 0$,
- $\mathbb{E}(XX')$ is nonsingular,
- $\mathbb{E}(\boldsymbol{\nu}\boldsymbol{\nu}'|X) \equiv \boldsymbol{\Omega}$ positive definite.

These assumptions suffice to estimate the model by GMM, leading to $\hat{\boldsymbol{\Pi}}_{OLS} = \left(\sum_{i=1}^n X_i X_i'\right)^{-1} \sum_{i=1}^n X_i \boldsymbol{Y}_i'$. To estimate the model by ML, we also assume that $\boldsymbol{\nu} \mid X \sim N(0, \boldsymbol{\Omega})$.

Using the closed-form expression for multivariate normal densities, we find that the conditional log likelihood function is

$$\log f(\boldsymbol{Y}|X; \boldsymbol{\Pi}, \boldsymbol{\Omega}) = -\frac{M}{2}\log 2\pi + \frac{1}{2}\log\left|\boldsymbol{\Omega}^{-1}\right| - \frac{1}{2}\left(\boldsymbol{Y} - \boldsymbol{\Pi}'X\right)' \boldsymbol{\Omega}^{-1} \left(\boldsymbol{Y} - \boldsymbol{\Pi}'X\right),$$

hence our objective function is

$$\begin{aligned} Q_n(\boldsymbol{\Pi}, \boldsymbol{\Omega}) &= -\frac{M}{2}\log 2\pi + \frac{1}{2}\log\left|\boldsymbol{\Omega}^{-1}\right| - \frac{1}{2n}\sum_{i=1}^n \left(\boldsymbol{Y}_i - \boldsymbol{\Pi}'X_i\right)' \boldsymbol{\Omega}^{-1} \left(\boldsymbol{Y}_i - \boldsymbol{\Pi}'X_i\right) \\ &= -\frac{M}{2}\log 2\pi + \frac{1}{2}\log\left|\boldsymbol{\Omega}^{-1}\right| - \frac{1}{2n}\sum_{i=1}^n \operatorname{tr}\big(\left(\boldsymbol{Y}_i - \boldsymbol{\Pi}'X_i\right)' \boldsymbol{\Omega}^{-1} \left(\boldsymbol{Y}_i - \boldsymbol{\Pi}'X_i\right)\big) \\ &= -\frac{M}{2}\log 2\pi + \frac{1}{2}\log\left|\boldsymbol{\Omega}^{-1}\right| - \frac{1}{2n}\sum_{i=1}^n \operatorname{tr}\big(\boldsymbol{\Omega}^{-1}(\boldsymbol{Y}_i - \boldsymbol{\Pi}'X_i)(\boldsymbol{Y}_i - \boldsymbol{\Pi}'X_i)'\big) \\ &= -\frac{M}{2}\log 2\pi + \frac{1}{2}\log\left|\boldsymbol{\Omega}^{-1}\right| - \frac{1}{2}\operatorname{tr}\bigg(\frac{\boldsymbol{\Omega}^{-1}}{n}\sum_{i=1}^n (\boldsymbol{Y}_i - \boldsymbol{\Pi}'X_i)(\boldsymbol{Y}_i - \boldsymbol{\Pi}'X_i)'\bigg) \\ &\equiv -\frac{M}{2}\log 2\pi + \frac{1}{2}\log\left|\boldsymbol{\Omega}^{-1}\right| - \frac{1}{2}\operatorname{tr}\big(\boldsymbol{\Omega}^{-1}\hat{\boldsymbol{\Omega}}(\boldsymbol{\Pi})\big), \end{aligned}$$

where the last line defines $\hat{\boldsymbol{\Omega}}(\boldsymbol{\Pi})$ in analogy to $\hat{\sigma}^2$ in our warm-up example. (The notation here anticipates that this is going to be our estimator, which we strictly speaking don't know yet.)

We first optimize with respect to $\boldsymbol{\Omega}$. A fact that you need not memorize is that if $\boldsymbol{A}$ and $\boldsymbol{B}$ are symmetric, positive definite, and conformable, then

$$\arg\max_{\boldsymbol{A}} \{\log|\boldsymbol{A}| - \operatorname{tr}(\boldsymbol{AB})\} = \boldsymbol{B}^{-1}.$$

Thus, we find that $\boldsymbol{\Omega}^* = \hat{\boldsymbol{\Omega}}(\boldsymbol{\Pi})$.

We can now concentrate the objective function by plugging in the estimator. Then

$$\begin{aligned}
\hat{\boldsymbol{\Pi}}_{ML} &= \arg\max_{\boldsymbol{\Pi}} \Big\{ -\frac{M}{2}\log 2\pi + \frac{1}{2}\log\big|\hat{\boldsymbol{\Omega}}(\boldsymbol{\Pi})^{-1}\big| - \frac{1}{2}\operatorname{tr}\left(\boldsymbol{I}_M\right)\Big\} \\
&= \arg\max_{\boldsymbol{\Pi}} \Big\{ -\frac{M}{2}\log 2\pi - \frac{1}{2}\log\big|\hat{\boldsymbol{\Omega}}(\boldsymbol{\Pi})\big| - \frac{M}{2}\Big\} \\
&= \arg\min_{\boldsymbol{\Pi}} \ \big|\hat{\boldsymbol{\Omega}}(\boldsymbol{\Pi})\big| \\
&= \arg\min_{\boldsymbol{\Pi}} \ \left|\frac{1}{n}\sum_{i=1}^n \left(\boldsymbol{Y}_i - \boldsymbol{\Pi}' X_i\right)\left(\boldsymbol{Y}_i - \boldsymbol{\Pi}' X_i\right)'\right|.
\end{aligned}$$

To see that this is solved by the OLS estimator, write

$$\begin{aligned}
&\sum_{i=1}^n \left(\boldsymbol{Y}_i - \boldsymbol{\Pi}' X_i\right)\left(\boldsymbol{Y}_i - \boldsymbol{\Pi}' X_i\right)' \\
= &\sum_{i=1}^n \big(\boldsymbol{Y}_i - \hat{\boldsymbol{\Pi}}'_{OLS} X_i + \hat{\boldsymbol{\Pi}}'_{OLS} X_i - \boldsymbol{\Pi}' X_i\big)\big(\boldsymbol{Y}_i - \hat{\boldsymbol{\Pi}}'_{OLS} X_i + \hat{\boldsymbol{\Pi}}'_{OLS} X_i - \boldsymbol{\Pi}' X_i\big)' \\
= &\sum_{i=1}^n \big(\hat{\boldsymbol{\nu}}_i + (\hat{\boldsymbol{\Pi}}'_{OLS} - \boldsymbol{\Pi}')X_i\big)\big(\hat{\boldsymbol{\nu}}_i + (\hat{\boldsymbol{\Pi}}'_{OLS} - \boldsymbol{\Pi}')X_i\big)' \\
= &\sum_{i=1}^n \big(\hat{\boldsymbol{\nu}}_i \hat{\boldsymbol{\nu}}_i' + \hat{\boldsymbol{\nu}}_i X_i'(\hat{\boldsymbol{\Pi}}_{OLS} - \boldsymbol{\Pi}) + (\hat{\boldsymbol{\Pi}}'_{OLS} - \boldsymbol{\Pi}')X_i \hat{\boldsymbol{\nu}}_i' + (\hat{\boldsymbol{\Pi}}'_{OLS} - \boldsymbol{\Pi}')X_i X_i'(\hat{\boldsymbol{\Pi}}_{OLS} - \boldsymbol{\Pi})\big) \\
= &\sum_{i=1}^n \hat{\boldsymbol{\nu}}_i \hat{\boldsymbol{\nu}}_i' + \sum_{i=1}^n \big(\hat{\boldsymbol{\Pi}}'_{OLS} - \boldsymbol{\Pi}'\big)X_i X_i'\big(\hat{\boldsymbol{\Pi}}_{OLS} - \boldsymbol{\Pi}\big),
\end{aligned}$$

where the last step uses that from the geometry of OLS, $\sum_{i=1}^n X_i \hat{\boldsymbol{\nu}}_i' = 0$. Now,

$$\sum_{i=1}^n \big(\hat{\boldsymbol{\Pi}}_{OLS} - \boldsymbol{\Pi}\big)' X_i X_i' \big(\hat{\boldsymbol{\Pi}}_{OLS} - \boldsymbol{\Pi}\big)$$

is a quadratic form, and it is true that if $\boldsymbol{A}$ and $\boldsymbol{B}$ are positive semidefinite, then $|\boldsymbol{A} + \boldsymbol{B}| \geq |\boldsymbol{A}|$. The objective is therefore minimized by $\hat{\boldsymbol{\Pi}}_{OLS}$, which achieves value $\sum_{i=1}^n \hat{\boldsymbol{\nu}}_i \hat{\boldsymbol{\nu}}_i'$. Substituting into $\boldsymbol{\Omega}^* = \hat{\boldsymbol{\Omega}}(\boldsymbol{\Pi})$, it now follows that $\hat{\boldsymbol{\Omega}}_{ML} = \hat{\boldsymbol{\Omega}}(\hat{\boldsymbol{\Pi}}_{OLS})$.

We will not go through consistency and asymptotic normality of this. Our analysis of GMM implies that the assumptions used to construct this estimator, i.e. independent draws and normality, are not really needed for its consistency and asymptotic normality. In cases where these assumptions do not hold, we could therefore think of the OLS estimator through its ML justification as a pseudo-ML estimator, illustrating that pseudo-ML can be "successful."

3 Binary Response

Binary response models can generally be expressed in the form

$$Y = \mathbf{1}\left\{\phi(X, \varepsilon; \theta_0) \geq 0\right\}$$

which is frequently specialized to

$$Y = \mathbf{1}\left\{X'\beta - \varepsilon \geq 0\right\}$$

$$\iff \Pr(Y = 1|X) = \Pr(\varepsilon \leq X'\beta) = F_\varepsilon\left(X'\beta\right).$$

Assume that $\mathbb{E}XX'$ is nonsingular and that F_ε is strictly increasing, then the model is identified up to a scale normalization and (if X has a constant component) a location normalization. Other than that, different assumptions about F_ε lead to different models. Thus, assume that ε is logistically distributed, then we have the *logit model*

$$\Pr(Y = 1|X_i) = \frac{\exp(X'\beta)}{1 + \exp(X'\beta)}.$$

We then have log likelihood

$$\log f(Y|X;\theta) = Y \log \frac{\exp(X'\beta)}{1 + \exp(X'\beta)} + (1 - Y) \log \frac{1}{1 + \exp(X'\beta)},$$

and the maximum likelihood estimator is characterized as maximizer of

$$\begin{aligned} Q_n(\beta) &= \frac{1}{n} \sum_i \left(Y_i \log \frac{\exp(X_i'\beta)}{1 + \exp(X_i'\beta)} + (1 - y_i) \log \frac{1}{1 + \exp(X_i'\beta)} \right) \\ &= \frac{1}{n} \sum_i \big(Y_i X_i'\beta - Y_i \log\left(1 + \exp(X_i'\beta)\right) - (1 - Y_i) \log\left(1 + \exp(X_i'\beta)\right) \big) \\ &= \frac{1}{n} \sum_i \big(Y_i X_i'\beta - \log\left(1 + \exp(X_i'\beta)\right) \big). \end{aligned}$$

We can thus write

$$\frac{\partial Q_n(\beta)}{\partial \beta} = \frac{1}{n} \sum_i \left(Y_i X_i - \frac{\exp(X_i'\beta)}{1 + \exp(X_i'\beta)} X_i \right) = \frac{1}{n} \sum_i \big(Y_i - F_\varepsilon\left(X_i'\beta\right) \big) X_i$$

and

$$\frac{\partial^2 Q_n(\beta)}{\partial \beta \partial \beta'} = -\frac{1}{n} \sum_i F_\varepsilon\left(X_i'\beta\right) \big(1 - F_\varepsilon\left(X_i'\beta\right)\big) X_i X_i',$$

applying the formula $F(t) = e^t/(1 + e^t) \Rightarrow F'(t) = F(t)(1 - F(t))$ to $F_\varepsilon\left(X\beta\right) = \frac{\exp(X'\beta)}{1+\exp(X'\beta)}$.

We can now establish consistency and asymptotic normality as in the previous example. Note in particular that the Hessian of Q_n is negative definite and the sample criterion function therefore concave, which together with likelihood identification and pointwise consistency of Q_n for Q implies consistency even without compactness of Θ. Also, the last display above immediately gives uniform boundedness of the (sample and population) Hessian.

We close with two asides:

- The above algebra, with expectations replacing sample averages, yields

$$
\begin{aligned}
\frac{\partial Q(\beta)}{\partial \beta} &= \mathbb{E}\big(\big(Y - F_\varepsilon\left(X'\beta\right)\big)X\big) \\
\frac{\partial^2 Q(\beta)}{\partial \beta \partial \beta'} &= -\mathbb{E}\big(F_\varepsilon\left(X'\beta\right)\big(1 - F_\varepsilon\left(X'\beta\right)\big)XX'\big).
\end{aligned}
$$

  At first glance, the information matrix equality (specifically, that the line marked (*) below equals minus the Hessian) may not appear obvious. However, write

$$
\begin{aligned}
& \mathbb{E}\big[\big(\big(Y - F_\varepsilon\left(X'\beta\right)\big)X\big)\big(\big(Y - F_\varepsilon\left(X'\beta\right)\big)X\big)'\big] \\
= \; & \mathbb{E}\big[\big(Y - F_\varepsilon\left(X'\beta\right)\big)^2 XX'\big] \quad (*) \\
= \; & \mathbb{E}\big[\mathbb{E}\big[\big(Y - F_\varepsilon\left(X_i'\beta\right)\big)^2 | X\big]XX'\big] \\
= \; & \mathbb{E}\big[\mathrm{var}(Y|X)XX'\big] \\
= \; & \mathbb{E}\big[F_\varepsilon\left(X'\beta\right)\big(1 - F_\varepsilon\left(X'\beta\right)\big)XX'\big] \\
= \; & -\frac{\partial^2 Q(\beta)}{\partial \beta \partial \beta'},
\end{aligned}
$$

  where we used the Law of Iterated Expectations followed by our knowledge that, conditionally on X, Y is distributed Bernoulli with parameter $F_\varepsilon\left(X'\beta\right)$ and therefore has mean $F_\varepsilon\left(X'\beta\right)$ and variance $F_\varepsilon\left(X'\beta\right)\big(1 - F_\varepsilon\left(X'\beta\right)\big)$.

- This model can also be estimated by GMM. (It is just identified, so really we're just doing plain method of moments.) To do so, we need a function $\boldsymbol{g}$ of the dimensionality of X s.t. $\mathbb{E}\boldsymbol{g}(Y, X; \beta) = 0$. As a general rule, if conditional expectations can be written out, they immediately give rise to such functions. In the present example, the Law of Iterated Expectations yields

$$
\mathbb{E}\left(X\left(Y - F_\varepsilon\left(X'\beta\right)\right)\right) = \mathbb{E}\left(X\mathbb{E}\left(Y - F_\varepsilon\left(X'\beta\right)|X\right)\right) = 0.
$$

  so our method of moments estimator is defined by the sample analog of this,

$$
\frac{1}{n}\sum_{i=1}^{n} X_i\big(y_i - F_\varepsilon(X_i'\hat{\beta})\big) = 0.
$$

  We see that the estimators algebraically coincide. That is, the natural GMM estimator uses the score equations as moment conditions and therefore is exactly the ML estimator. Hence, this estimator is efficient in the strong sense of replicating the asymptotic variance of ML.

4 Tobit Type II and Heckman Two-Step

We next analyze the Type II Tobit and the Heckman Two-Step ("Heckit") strategy of adjusting for nonignorable censoring. This development can be generalized in many ways and stands at the beginning of a huge literature in applied econometrics.

4.1 The Model

The following model is known as Type II Tobit. (To economize on subscripts, I drop the "0" subscript for true parameter value.)

$$\begin{aligned} Y_1^* &= X_1'\beta_1 + \varepsilon_1 \\ Y_2^* &= X_2'\beta_2 + \varepsilon_2 \\ Y_1 &= \mathbf{1}\{Y_1^* \geq 0\} \\ Y_2 &= Y_2^* \times \mathbf{1}\{Y_1^* \geq 0\} \\ \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix} &\sim N\left(0, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\right). \end{aligned}$$

The original motivating examples for the model are that the decision to enter the labor force may depend on some covariates (like having children) that do not determine the wage conditional on entering the labor force, and that a government's decision to extend development aid to some country may be driven by some considerations (like human rights) that do not influence the amount of aid conditional on there being any. (The causes need not be separate – the model generalizes Type I Tobit, in which selection and outcome equation coincide.) The first equation is also called *selection equation*, the second one is the *outcome equation.* Typically, the parameter of substantive interest is β_2. Note that, under assumptions maintained here, β_2 could be estimated by regressing Y_2^* on X_2 if the former were observable. However, unless $\rho = 0$, we cannot estimate β_2 by regressing Y_2 on X_2 in the subsample where $Y_1 = 1$. This is because by conditioning on $Y_1 = 1$, we select for high ε_1 and therefore for high [low] ε_2 if ρ is positive [negative]. That is, the subsample on which Y_2^* is observed is selective.

The selection model embodied in the first and third equations gives rise to likelihood

$$\Pr(Y_1 = 0|X_1, X_2, \theta) = \Phi\left(\frac{-X_1'\beta_1}{\sigma_1}\right).$$

The likelihood of Y_2 conditionally on X_1, X_2, and $Y_1 = 1$ (i.e. Y_2 being observed) equals

$$\begin{aligned} f(Y_2|X_1X_2, Y_1 = 1; \theta) &= \Pr(Y_1 = 1|Y_2, X_1, X_2; \theta) \times f(Y_2|X_1, X_2; \theta)/\Phi\left(\frac{X_1'\beta_1}{\sigma_1}\right) \\ &= \Pr(\varepsilon_1 > -X_1\beta_1|Y_2, X_1, X_2; \theta) \times \frac{1}{\sigma_2}\phi\left(\frac{Y_2 - X_2'\beta_2}{\sigma_2}\right)/\Phi\left(\frac{X_1'\beta_1}{\sigma_1}\right) \\ &= \frac{1}{\sigma_2}\left(1 - \Phi\left(\frac{-X_1'\beta_1 - \rho\frac{\sigma_1}{\sigma_2}(Y_2 - X_2'\beta_2)}{\sigma_1\sqrt{1-\rho^2}}\right)\right)\phi\left(\frac{Y_2 - X_2'\beta_2}{\sigma_2}\right)/\Phi\left(\frac{X_1'\beta_1}{\sigma_1}\right) \\ &= \frac{1}{\sigma_2}\Phi\left(\frac{X_1'\beta_1 + \rho\frac{\sigma_1}{\sigma_2}(Y_2 - X_2'\beta_2)}{\sigma_1\sqrt{1-\rho^2}}\right)\phi\left(\frac{Y_2 - X_2'\beta_2}{\sigma_2}\right)/\Phi\left(\frac{X_1'\beta_1}{\sigma_1}\right), \end{aligned}$$

where we used that the distribution of $\varepsilon_1|\varepsilon_2$ is $N\big(\rho\frac{\sigma_1}{\sigma_2}\varepsilon_2, \sigma_1^2(1-\rho^2)\big)$. Note that all appearances of Y_2 on the r.h.s. really are Y_2^*; I anticipate that the two are the same on the event we condition on.

4.1.1 Identification and ML Estimation

We will establish identification by conducting the following thought experiment: If we had perfect knowledge of the distribution of the data and therefore of the above likelihoods as "black-box functions" of (X_1, X_2, Y_1, Y_2), could we back out the true parameter values? Indeed, it is immediately clear that the probit part of the model identifies β_1/σ_1 (as long as $\mathbb{E}X_iX_i'$ is invertible) but not either of these parameters in isolation. Observe also that $f(Y_2|X_1, X_2, Y_2 = 1; \theta)$ can be rewritten as

$$f(Y_2|X_1, X_2, Y_2 = 1; \theta) = \frac{1}{\sigma_2}\Phi\left(\frac{X_1'\frac{\beta_1}{\sigma_1} + \frac{\rho}{\sigma_2}(Y_2 - X_2'\beta_2)}{\sqrt{1-\rho^2}}\right)\phi\left(\frac{Y_2 - X_2\beta_2}{\sigma_2}\right)/\Phi\left(\frac{X_1'\beta_1}{\sigma_1}\right).$$

Since this expression also depends on β_1 and σ_1 only through β_1/σ_1, we can really identify only this ratio. The natural way to deal with this is to normalize $\sigma_1 = 1$ and define $\theta \equiv (\beta_1, \beta_2, \sigma_2, \rho)$. This leads to the following simplification:

$$\begin{aligned} \Pr(Y_1 = 0|X_1, X_2, \theta) &= \Phi\left(-X_1'\beta_1\right) \\ f(Y_2|X_1, X_2, Y_2 = 1; \theta) &= \frac{1}{\sigma_2}\Phi\left(\frac{X_1'\beta_1 + \frac{\rho}{\sigma_2}(Y_2 - X_2'\beta_2)}{\sqrt{1-\rho^2}}\right)\phi\left(\frac{Y_2 - X_2'\beta_2}{\sigma_2}\right)/\Phi\left(X_1'\beta_1\right). \end{aligned}$$

We will now show that the abridged θ *is* identified. To begin, recall that the selection equation identifies β_1. We can therefore treat β_1, hence $\Phi\left(X_1'\beta_1\right)$, as known. Use this to define

$$\begin{aligned} \tilde{f}(Y_2|X_1, X_2; \theta) &= f(Y_2|X_1, X_2, Y_2 = 1; \theta)\Pr(Y_1 = 1|X_1, X_2, \theta) \\ &= \frac{1}{\sigma_2}\Phi\left(\frac{X_1'\beta_1 + \frac{\rho}{\sigma_2}(Y_2 - X_2'\beta_2)}{\sqrt{1-\rho^2}}\right)\phi\left(\frac{Y_2 - X_2'\beta_2}{\sigma_2}\right). \end{aligned}$$

Observe that

$$\begin{aligned}
\frac{\partial \tilde{f}(\cdot)}{\partial Y_2} &= \frac{1}{\sigma_2}\phi\left(\frac{X_1'\beta_1 + \frac{\rho}{\sigma_2}\left(Y_2 - X_2'\beta_2\right)}{\sqrt{1-\rho^2}}\right)\frac{\rho}{\sigma_2\sqrt{1-\rho^2}}\phi\left(\frac{Y_2 - X_2'\beta_2}{\sigma_2}\right) \\
&\quad + \frac{1}{\sigma_2^2}\Phi\left(\frac{X_1'\beta_1 + \frac{\rho}{\sigma_2}\left(Y_2 - X_2'\beta_2\right)}{\sqrt{1-\rho^2}}\right)\phi'\left(\frac{Y_2 - X_2'\beta_2}{\sigma_2}\right) \\
\frac{\partial \tilde{f}(\cdot)}{\partial X_2} &= \frac{1}{\sigma_2}\phi\left(\frac{X_1'\beta_1 + \frac{\rho}{\sigma_2}\left(Y_2 - X_2'\beta_2\right)}{\sqrt{1-\rho^2}}\right)\left(-\frac{\rho\beta_2}{\sigma_2\sqrt{1-\rho^2}}\right)\phi\left(\frac{Y_2 - X_2'\beta_2}{\sigma_2}\right) \\
&\quad - \frac{1}{\sigma_2^2}\Phi\left(\frac{X_1'\beta_1 + \frac{\rho}{\sigma_2}\left(Y_2 - X_2'\beta_2\right)}{\sqrt{1-\rho^2}}\right)\phi'\left(\frac{Y_2 - X_2'\beta_2}{\sigma_2}\right)\beta_2 \\
\frac{\partial \tilde{f}(\cdot)}{\partial X_1} &= \frac{1}{\sigma_2}\phi\left(\frac{X_1'\beta_1 + \frac{\rho}{\sigma_2}\left(Y_2 - X_2'\beta_2\right)}{\sqrt{1-\rho^2}}\right)\frac{\beta_1}{\sqrt{1-\rho^2}}\phi\left(\frac{Y_2 - X_2'\beta_2}{\sigma_2}\right).
\end{aligned}$$

A first observation is that

$$\beta_2 = -\frac{\partial \tilde{f}(\cdot)/\partial X_2}{\partial \tilde{f}(\cdot)/\partial Y_2}.$$

(Without algebra, this is really clear from the fact that Y_2 and X_2 enter the likelihood only through $(Y_2 - X_2'\beta_2)$.)

Having identified β_1 and β_2, we can choose to evaluate $\tilde{f}$ at arguments where $X_1'\beta_1 = Y_2 - X_2'\beta_2 = 0$. At any such value of the argument, we have

$$\begin{aligned}
\left.\frac{\partial \tilde{f}(\cdot)}{\partial X_1}\right\|_{X_1'\beta_1 = Y_2 - X_2'\beta_{2i} = 0} &= \beta_1 \times \frac{1}{\sigma_2\sqrt{1-\rho^2}}\big(\phi(0)\big)^2 \\
\left.\frac{\partial \tilde{f}(\cdot)}{\partial Y_2}\right\|_{X_1'\beta_1 = Y_2 - X_2'\beta_{2i} = 0} &= \frac{\rho}{\sigma_2^2\sqrt{1-\rho^2}}\big(\phi(0)\big)^2.
\end{aligned}$$

With β_1 known, these are basically two equations in the two remaining unknowns (σ_2, ρ).

We close with two remarks:

- This is an example of a non-constructive identification proof: We established in a thought experiment that perfect knowledge of the likelihood would allow us to back out parameter values, but our estimation strategy should not be to solve sample analogs of these equations. After all, they involve derivatives of likelihoods evaluated at specific (themselves in practice estimated) parameter values.

- The argument made use of some *support conditions*. For the purpose of our thought experiment, we can freely take derivatives of likelihoods to be known, but we may evaluate these derivatives only at values of (X_1, X_2, Y_1, Y_2) on the support of the true distribution. Because of the normality assumption on ε_{2i}, we actually know that $Y_2 - X_2'\beta_2 = 0$ occurs on the support. We do not really know that $X_1'\beta_1 = 0$ can occur but our use of that assumption was not tight, i.e. we could work around it as long as there is some variation in X_1. Support assumptions on covariates are often made to enable arguments like the above.

Now, how should we estimate this model? An obvious approach is Maximum Likelihood. Substituting in for the standard normal p.d.f. and dropping constants, the objective function is

$$Q_n(\theta) =$$
$$\frac{1}{n}\sum_{i=1}^{n}\left[(1-Y_1)\log\Phi\left(-X_1'\beta_1\right)+Y_1\left(\log\Phi\left(\frac{X_1'\beta_1+\frac{\rho}{\sigma_2}\left(Y_2-X_2'\beta_2\right)}{\sqrt{1-\rho^2}}\right)-\frac{1}{2}(Y_2-X_2'\beta_2)^2-\log\sigma_2\right)\right],$$

where $\theta \equiv (\beta_1, \beta_2, \sigma_2, \rho)$.

4.2 "Heckit"

ML is the statistically efficient way to estimate such models; many implementations exist, and there are also some shortcuts like pre-estimating β_1 by probit to get a good initial point. However, with larger parameter vectors, the problem remains involved even by modern standards because the likelihood is multimodal. Heckman proposed a two-step method for this type of model. This method builds on the observation that

$$\begin{aligned}\mathbb{E}(Y_2|X_1, X_2, Y_1 = 1) &= \mathbb{E}(X_2'\beta_2 + \varepsilon_2|X_1, X_2, Y_1 = 1)\\ &= X_2'\beta_2 + \mathbb{E}(\varepsilon_2|\varepsilon_1 \geq -X_1'\beta_1).\end{aligned}$$

To get a grip on the last expression, let z_i be distributed standard normal and recall the following, standard algebra:

$$\begin{aligned}\mathbb{E}(z_i|z_i \geq t) &= \frac{\int_{z=t}^{\infty} z\phi(z)dz}{\int_{z=t}^{\infty}\phi(z)dz} = \frac{(2\pi)^{-1/2}\int_{z=t}^{\infty} ze^{-z^2/2}dz}{\Phi(-t)}\\ &= \frac{(2\pi)^{-1/2}\left[-e^{-z^2/2}\right]_t^{\infty}}{\Phi(-t)} = \frac{(2\pi)^{-1/2}e^{-t^2/2}}{\Phi(-t)} = \frac{\phi(t)}{\Phi(-t)} \equiv \lambda(-t),\end{aligned}$$

where the last equality defines the *Inverse Mills Ratio.*

Now,

$$\begin{aligned}\mathbb{E}(\varepsilon_2|\varepsilon_1 \geq -X_1'\beta_1) &= \frac{\int_{-X_1'\beta_1}^{\infty}\mathbb{E}(\varepsilon_2|\varepsilon_1)\phi\left(\varepsilon_1\right)d\varepsilon_1}{\int_{-X_1'\beta_1}^{\infty}\phi\left(\varepsilon_1\right)d\varepsilon_1}\\ &= \frac{\int_{-X_1'\beta_1}^{\infty}\rho\sigma_2\varepsilon_1\phi\left(\varepsilon_1\right)d\varepsilon_1}{\int_{-X_1'\beta_1}^{\infty}\phi\left(\varepsilon_1\right)d\varepsilon_1} = \rho\sigma_2\frac{\int_{-X_1'\beta_1}^{\infty}\varepsilon_1\phi\left(\varepsilon_1\right)d\varepsilon_1}{\int_{-X_1'\beta_1}^{\infty}\phi\left(\varepsilon_1\right)d\varepsilon_1} = \rho\sigma_2\lambda(X_1'\beta_1),\end{aligned}$$

where we used that due to our normalization, ε_{1i} is standard normal.

Thus, for those data points where $Y_1 = 1$, we can write

$$Y_2 = X_2'\beta_2 + \rho\sigma_2\lambda(X_1'\beta_1) + \eta_i$$

where $\mathbb{E}(\eta_i|X_1, X_2, Y_1 = 1) = 0$. If we knew β_1, we could therefore just estimate β_2 by running a OLS regression of Y_2 on $(X_2, \lambda\left(X_1'\beta_1\right))$. (The last component of the estimator would estimate $\rho\sigma_2$. The population variance of $(Y_2|X_1, X_2i, Y_1 = 1)$ equals $\sigma_2^2(1-\rho^2)$, so that separate estimates of ρ and σ_2 can be backed out.) In reality, things are a bit more complicated because β_2 is unknown. Heckman established that the following two-step procedure ("Heckit") works:

Step 1. Use probit to estimate β_1. Call this estimator $\hat{\beta}_1$.

Step 2. Restrict attention to observations with $Y_1 = 1$. Use OLS to estimate the equation

$$Y_2 = X_2'\beta_2 + \rho\sigma_2\lambda(X_1'\hat{\beta}_1) + \eta.$$

To reiterate, that this works is not completely obvious because of the estimated regressor $\lambda(X_1'\hat{\beta}_1)$ on the r.h.s., but it is nonetheless true (and arguments of this sort have since been much generalized). Unsurprisingly, inference theory changes and OLS standard errors would not be valid.

The Heckman two-step method generalizes easily to variations on the above model. It can be seen as the first appearance of a control function approach: correcting for selectivity by introducing a function into the regression that compensates selection bias. The "Heckit" estimator is hardcoded in most canned packages. Notice, though, that it is inefficient: It exploits normality assumptions for identification (and is sensitive to their failure!) but then fails to fully use them for estimation, and in this case there is no reason to believe that the moment conditions coincide with the score equations. In particular, an ML estimator would use second-stage information also in the estimation of β_1. The choice between ML and Heckit estimation of this model depends on how complex the likelihood is in a given application. Both estimation methods are implemented in Stata and similar packages.

5 (Smoothed) Maximum Score

Consider the binary choice model

$$Y = \mathbf{1}\{X'\beta + \varepsilon \geq 0\},$$

where the researcher observes (Y, X). We do not assume an exact distribution for ε, but we do assume that ε is continuous and that $\text{med}(\varepsilon) = 0$. For simplicity, we also assume continuous X with full support other than having a constant.

While the median assumption amounts to a location normalization, the assumptions are still weaker than for probit, say, and so β can only be identified up to scale. Unlike in earlier examples, it is convenient to normalize $|\beta| = 1$ and impose the same restriction on estimators. Then

$$\beta \in \arg \max_{\beta : |\beta|=1} \mathbb{E}\left((2Y-1)\, \mathbf{1}\{X'\beta \geq 0\}\right)$$

because

$$\mathbb{E}\left((2Y-1)\, \mathbf{1}\{X'\beta \geq 0\}\right) = \mathbb{E}\big(\mathbb{E}(2Y-1|X)\mathbf{1}\{X'\beta \geq 0\}\big)$$

and the r.h.s outer integrand is maximized pointwise by β because

$$\mathbb{E}\left(2Y-1|X\right) \geq 0 \Leftrightarrow X'\beta \geq 0.$$

If X has full support, then for any $\tilde{\beta} \neq \beta$, the set $\{x : x'\tilde{\beta} \times x'\beta < 0\}$ has positive probability and so the above arg max is unique. Without this assumption, β may be *partially identified*: Knowledge of the population distribution of observables restricts β to a nontrivial but also nonsingleton subset of the unit sphere.

Note that the empirical distribution of X_i has at most n mass points and so cannot have full support. Thus, the estimator

$$\hat{\beta} = \arg \max_{\beta|\beta|=1} \frac{1}{n} \sum_{i=1}^{n} (2Y_i - 1)\, \mathbf{1}\{X_i'\beta \geq 0\}$$

is not well-defined as written because the arg max is not unique. In the following, let $\hat{\beta}$ be a measurable selection from the arg max, e.g. the element that minimizes the first, then the second, etc. component.

This is the *Maximum Score* estimator (Manski, 1975). It has historic importance as one of the first *nonparametric* (actually semiparametric in modern terminology because of the linear index structure) estimators. It is also supremely ill-behaved. Under reasonable conditions, consistency can be established along the lines of the theorems provided in this lecture. As an aside, this shows the theorems' power because the resulting proof is much shorter than the original one. But with regard to the asymptotic distribution, red flags abound. To begin, the Hessian at the sample arg max is 0. This suggests, but does of course not prove, a slower than $\sqrt{n}$ rate of convergence. This conjecture is true:

The true rate is $n^{1/3}$ (Kim and Pollard, 1990); an estimator with $\sqrt{n}$-convergence does not exist under the assumptions (Chamberlain, 1986, who formalizes the intuition just alluded to), and the asymptotic distribution is intractable.[2]

Many of these issues are due to the extreme non-smoothness of the objective function. This raises the possibility that artifically smoothing the objective function might make for better behaved estimators. There is even a vague intuition that it might help with efficiency because it brings to bear whether a given $X_i'\beta$ is very close to zero or not. All these intuitions are correct. In particular, write

$$\hat{\beta} = \arg \max_{\beta:|\beta|=1} \frac{1}{n} \sum_{i=1}^{n} (2Y_i - 1)\, g_n(X_i'\beta),$$

where the function $g_n(t)$ is smooth, has $g(0) = 1/2$, asymptotes 0 [1] as $t \to -\infty$ [∞], and converges to $\mathbf{1}\{t \geq 0\}$ at a certain rate as $n \to \infty$. If you have seen kernel density estimation or kernel regression before, you'll realize that this is very similar.

This estimator is called the *Smoothed Maximum Score* estimator. It was suggested by Horowitz (1992), who also showed that it is asymptotically normal. Its rate of convergence is not $n^{1/2}$, but can be arbitrarily close to $n^{1/2}$ if the distribution of X_i is sufficiently favorable and g_n is chosen smartly.[3]

[2] A naive reaction might be to "just bootstrap" the distribution, but the simple nonparametric bootstrap is demonstrably inconsistent here (Abrevaya and Huang, 2005).

[3] Moreover, the estimator cannot only be bootstrapped; Horowitz (2002) shows that for test statistics based on a studentized estimator, the bootstrap achieves asymptotic refinement.

6 Maximum of a Uniform Distribution

This is a less regular example and one in which ML and GMM may disagree. Let X be i.i.d. uniformly distributed on $[0, \alpha_0]$. The aim is to estimate α_0. We briefly note that $\mathbb{E}X = \alpha_0/2$ and so $\hat{\alpha} = 2\bar{X}$ is a GMM estimator based on moment condition $\mathbb{E}(2X - \alpha_0) = 0$. We understand the behavior of this estimator very well. But in this example, it is not the ML estimator and turns out to be rather inefficient indeed. (That said, it is BLUE. So the ML estimator must be either biased or nonlinear. We will see that it is both.)

Let's compute the ML estimator. The likelihood for a single observation is $f(x; \alpha) = 1/\alpha \times \mathbf{1}\{0 \le x \le \alpha\}$, and so the sample criterion function equals

$$Q_n(\alpha) = \frac{1}{n}\sum_{i=1}^{n} \log\left(1/\alpha \times \mathbf{1}\{X_i \in [0, \alpha]\}\right) = \begin{cases} -\log\alpha & \text{if } \max_i\{X_i\} \le \alpha \\ -\infty & \text{otherwise} \end{cases}$$

with population analog

$$Q(\alpha) = \mathbb{E}\log f(X; \alpha) = \begin{cases} -\log\alpha & \text{if } \alpha \le \alpha_0 \\ -\infty & \text{otherwise} \end{cases}$$

By inspection, these problems are solved by $\hat{\alpha} = \max_i\{X_i\}$ respectively by the true value α_0.

These objective functions are discontinuous, and therefore of course not differentiable, at their maxima. This means that none of our extremum estimator theorems are immediately applicable, and it is a massive red flag with regard to our asymptotic distribution theorem. Indeed, while consistency of the estimator obtains (it's intuitively obvious from the closed-form expression, follows from developments below, but could also be shown by recovering well-separatedness of the population maximum), neither $\sqrt{n}$-consistency nor asymptotic normality are true. Plotting the objective function can provide an intuition: The Hessian at the solution is not well-defined, but the vertical drop loosely suggests an "unbounded Hessian," which by the theorem's formula would suggest an asymptotic variance of 0 for $\sqrt{n}(\hat{\alpha} - \alpha_0)$. In the same way that an asymptotic variance of ∞ would suggest failure of $\sqrt{n}$-consistency, this observation suggests, but does of course not prove, so-called "superconsistency." Is the estimator faster than $\sqrt{n}$-consistent?

The answer is yes; in fact, the true rate of convergence is n. To see this, let's approximate the c.d.f. of $n(\hat{\alpha} - \alpha_0)$. This c.d.f. is obviously 1 for nonnegative arguments, another pointer that asymptotic normality will fail (and also implying that the estimator cannot be unbiased). For arguments $t \le 0$, we have

$$\begin{aligned}
& \Pr(n(\hat{\alpha}-\alpha_0)\leq t) \\
= & \Pr\big(\max_{i=1,...,n}\{X_i\}\leq \alpha_0+t/n\big) \\
= & \Pr\left(X_1\leq\alpha_0+t/n,...,X_n\leq\alpha_0+t/n\right) \\
= & \Pr\left(X_1/\alpha_0\leq 1+t/\left(n\alpha_0\right),...,X_n/\alpha_0\leq 1+t/\left(n\alpha_0\right)\right) \\
= & \left(1+t/\left(n\alpha_0\right)\right)^n \\
\to & e^{t/\alpha_0},
\end{aligned}$$

where the last step invokes a well-known nonstochastic limit formula. Note this is 1 at $t=0$ as expected, so we derived a coherent limit c.d.f. In fact, we showed that $-n(\hat{\alpha}-\alpha_0)$ converges to an exponential distribution whose dispersion increases with α_0. A corollary is that $n(\hat{\alpha}-\alpha_0)=O_p(1)$, i.e. the estimator is consistent at rate n.

We conclude the example with some additional observations.

- Our standard inference theory does not apply, but it is relatively easy to construct hypothesis tests for this setting. Because the model is fully parameterized, we can even do exact tests. In particular, suppose that $H_0:\alpha_0=\alpha$ holds for some fixed value α. We clearly reject this null if $\hat{\alpha}>\alpha$, so the testing problem is effectively one-sided. From the above algebra, we have (setting $\alpha=\alpha_0$ and flipping the sign on t) $\Pr(n(\alpha-\hat{\alpha})>t)=(1-t/\left(n\alpha\right))^n$, so that the 95% critical value is $n\alpha(1-.05^{1/n})$.

  We can construct a CI by inverting this test. Notice that the critical value depends on α (the test statistic is not asymptotically *pivotal*) and so in principle we need to recompute it at each α. In this example, the condition that the test statistic be smaller than the critical value can be solved in closed form, and we get a CI of $[\hat{\alpha},20^{1/n}\hat{\alpha}]$.

  While there is no need for asymptotic CI's if we have computationally cheap finite sample ones, in principle we can compute an asymptotic counterpart as well. From the last line of algebra above, it would be $[\hat{\alpha},\hat{\alpha}\frac{n}{n+\log .05}]$. Note that this is well-defined only if $n+\log .05>0$, which is the case for $n\geq 3$. This weird feature is an artefact of using an asymptotic approximation (namely the limit formula) at very small n. The intervals become very similar for moderate n.

- The ML estimator has negative bias, but standard formulae for order statistics imply that $\mathbb{E}\hat{\alpha}=\frac{n}{n+1}\alpha_0$, so we can define a bias-corrected (in fact unbiased) and still superconsistent estimator $\tilde{\alpha}=\frac{n+1}{n}\hat{\alpha}$. For this particular estimation problem, that is the preferred estimator. (Theorems about asymptotic efficiency of ML do not apply due to the problem's irregularity.)