Econ 6200: Econometrics II
Prelim, May 14^{th}, 2022

This exam consists of 14 questions, not of equal length or difficulty, grouped into two exercises. Each question is worth 10 points. Remember to always explain your answer. Throughout, you may invoke theorems from class without proof. It is acceptable to invoke "existence of moments as needed."

Good luck!

1 ML Estimation of the Exponential Distribution. The exponential distribution is supported on $[0, \infty)$ and has p.d.f.

$$f(y) = \lambda \exp(-\lambda y) \times \mathbf{1}\{y \geq 0\}.$$

This is the distribution of times between observed events of constant hazard rate, i.e. "time between events in a Poisson process." $\lambda \in (0, \infty)$ is the parameter of the corresponding Poisson process. It can be verified that $\mathbb{E}Y = 1/\lambda$ and $\text{var}(Y) = 1/\lambda^2$. Assume that n independent observations $Y_1, \ldots, Y_n$ are available.

1.1 Show that the ML estimator is $\hat{\lambda} = 1/\bar{Y}$.

1.2 Use a theorem about extremum estimators to argue that $\hat{\lambda}$ is consistent. (Do not prove the theorem.)

1.3 Provide the asymptotic distribution by means of the theorem reproduced below. Verify assumptions; provide expressions for score, $\mathbf{\Sigma}$, and $\boldsymbol{H}$; use them to derive the asymptotic variance. (As a sanity check, you should know the asymptotic variance right away from information already given.)

Theorem:

Assume that

1. $\hat{\theta} \xrightarrow{p} \theta_0$,
2. $\theta_0 \in \text{int}(\Theta)$,
3. there exists a neighborhood $\mathcal{N}$ of θ_0 s.t. for any $\theta \in \mathcal{N}$, $Q_n(\theta)$ is a.s. twice continuously differentiable in θ,
4. $\sqrt{n}\frac{\partial Q_n(\theta_0)}{\partial \theta} \xrightarrow{d} N(0, \Sigma)$, Σ positive definite,
5. $\mathbf{H}(\theta) \equiv \frac{\partial^2 Q(\theta)}{\partial \theta \partial \theta'}$ is continuous at θ_0.

6. $\sup_{\theta\in\mathcal{N}}\left\|\frac{\partial^2 Q_n(\theta)}{\partial\theta\partial\theta'}-\frac{\partial^2 Q(\theta)}{\partial\theta\partial\theta'}\right\|\xrightarrow{p}0$.

7. $\mathbf{H}_0\equiv\mathbf{H}(\theta_0)$ is nonsingular.

Then

$$\sqrt{n}\big(\hat{\theta}-\theta_0\big)\xrightarrow{d}N\big(0,\mathbf{H}_0^{-1}\Sigma\mathbf{H}_0^{-1}\big).$$

1.4 Can you decide whether $\hat{\lambda}$ is unbiased?

1.5 You are interested in testing $H_0:\lambda=1$. Provide the constrained estimator $\widetilde{\lambda}$ corresponding to this null.

Next, explain how to test the null using one of the "trinity" of test statistics, reproduced here for your convenience in the lecture's notation:

$$\begin{aligned}
W &\equiv \sqrt{n}r\big(\hat{\theta}\big)'\left(\mathbf{R}\big(\hat{\theta}\big)\hat{\Sigma}^{-1}\mathbf{R}'\big(\hat{\theta}\big)\right)^{-1}\sqrt{n}r\big(\hat{\theta}\big)\\
LR &\equiv 2n\left(Q_n\big(\hat{\theta}\big)-Q_n\big(\tilde{\theta}\big)\right)\\
LM &\equiv n\frac{\partial Q_n\big(\tilde{\theta}\big)}{\partial\theta'}\tilde{\Sigma}^{-1}\frac{\partial Q_n\big(\tilde{\theta}\big)}{\partial\theta},
\end{aligned}$$

For the statistic of your choice, state its asymptotic distribution and express the actual test statistic as explicitly as possible as function of data.

1.6 For the test statistic of your choice, would you expect bootstrap computation of the critical value to be valid? Is there reason to believe that bootstrap computation will be better, about as good, or worse compared to asymptotic approximation?

1.7 In this particular setting, there is a way to compute critical values for $H_0:\lambda=1$ that beats both asymptotic approximation and the bootstrap. Explain.

2 Consider the following model ("Tobit Type V"):

$$
\begin{aligned}
Y_1^* &= \alpha_1 + \beta_1 X_1 + \varepsilon_1 \\
Y_2^* &= \alpha_2 + \beta_2 X_2 + \varepsilon_2 \\
Y_3^* &= \alpha_3 + \beta_3 X_3 + \varepsilon_3 \\
Y_1 &= \mathbf{1}\{Y_1^* \geq 0\} \\
Y_2 &= Y_2^* \cdot Y_1 \\
Y_3 &= Y_3^* \cdot (1 - Y_1) \\
\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{bmatrix} &\sim N\left(0, \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 \end{bmatrix}\right),
\end{aligned}
$$

where all error terms are independent of the covariates. The researcher observes i.i.d. realizations of the scalars $(Y_1, Y_2, Y_3, X_1, X_2, X_3)$.

2.1 In this model there is *one* selection equation and *two* outcome variables. Explain and verbally describe the selective observation process described here. Which parameters do you expect to be of primary (causal) interest?

2.2 Argue that the model is identified after one normalization. You may invoke results shown in class.

2.3 Write down (or at least verbally describe) the objective function for an estimator of β_3 that makes efficient use of likelihood information.

2.4 Verbally explain an alternative estimator.

Suppose now that you also observe Z_3 s.t. $\text{cov}(Z_3, X_3) \neq 0$ and where Z_3 is independent of error terms.

The next questions are independent of the preceding ones. Also, the questions are not about the wisdom of making certain assumptions, but about what would be true if assumptions were known to hold.

2.5 If (contrary to earlier parts of this question) you observed Y_3^*, could you estimate β_3 by IV regression of Y_3^* on X_3 using Z_3 as instrument? Would this be an efficient thing to do?

2.6 If (contrary to earlier parts of this question) you observed Y_3^*, could you estimate β_3 by TSLS regression of Y_3^* on X_3 using (X_3, Z_3) as instruments? Would this be an efficient thing to do?

2.7 Going back to the original setting (i.e., we only observe Y_3), could you estimate β_3 by IV regression of Y_3 on X_3 using Z_3 as instrument? Would this be an efficient thing to do?

Answers

1.1 The log likelihood (rescaled by $1/n$ for convenience) is

$$Q_n(\lambda) = \frac{1}{n}\sum_{i=1}^{n} \log\left[\lambda \exp(-\lambda y_i)\right] = \frac{1}{n}\sum_{i=1}^{n} \left[\log \lambda - \lambda y_i\right]$$

with first derivative

$$\frac{1}{n}\sum_{i=1}^{n} \left[1/\lambda - y_i\right]$$

and second derivative

$$-\frac{1}{n}\sum_{i=1}^{n} \lambda^{-2} = -\lambda^{-2} < 0,$$

thus the objective is strictly concave (that's also useful later), and the estimator is characterized by FOC

$$n/\hat{\lambda} = \sum_{i=1}^{n} y_i \Longrightarrow \hat{\lambda} = 1/\overline{y}$$

1.2 We don't have compact Θ but we do have strict concavity of Q_n, so use the "consistency through concavity" theorem. To verify further assumptions, write

$$\begin{aligned} Q(\lambda) &= \operatorname{plim} Q_n = \mathbb{E}\log\left[\lambda \exp(-\lambda Y)\right] = \mathbb{E}\left(\log \lambda - \lambda Y\right) \\ Q'(\lambda) &= \mathbb{E}\left(1/\lambda - Y\right) \\ Q''(\lambda) &= -\lambda^{-2}. \end{aligned}$$

$Q_n \overset{a.s./p}{\to} Q$ follows by SLLN/WLLN. Pointwise convergence suffices because of concavity. Identification obtains because knowledge of the true distribution would imply knowledge of $\lambda = 1/\mathbb{E}Y$. Thus, the true parameter value uniquely maximizes the expected likelihood by the KL inequality. (Alternatively, directly verify that the true value λ_0 uniquely maximizes $Q(\lambda) = \mathbb{E}\left(\log \lambda - \lambda Y\right)$, seen as function of λ but with the expectation evaluated under λ_0.)

Note that other than compactness, we also don't have uniform convergence of $Q_n = \frac{1}{n}\sum_{i=1}^{n}\left[\log \lambda - \lambda Y_i\right]$ to $Q = \mathbb{E}\left[\log \lambda - \lambda Y\right]$. For any data set, the two are far apart for large choice of λ. So using the non-concave consistency theorem here is problematic. (If you wonder why a uniform LLN does not apply, consider the behavior of Q_n at both $\lambda \to 0$ and $\lambda \to \infty$.)

1.3

1. $\hat{\theta} \overset{p}{\to} \theta_0$ was established.

2. $\theta_0 \in \operatorname{int}(\Theta)$ is for free here because Θ is open.

3. $Q_n'' = -\lambda^{-2}$ is trivially continuous.

4. $$\frac{\partial Q_n(\lambda_0)}{\partial \lambda} = \frac{1}{n}\sum_i [1/\lambda_0 - Y_i]$$

   is a sample average of $(1/\lambda_0 - Y_i)$, which has expected value $\mathbb{E}[1/\lambda_0 - Y] = 0$ and variance equal to the variance of Y, i.e. $1/\lambda_0^2$. A Lindeberg-Levy CLT yields $\sqrt{n}\frac{\partial Q_n(\lambda_0)}{\partial \lambda} \xrightarrow{d} N(0, 1/\lambda_0^2)$. We conclude that $\Sigma = \lambda_0^{-2}$.

5. $\mathbf{H} = -\lambda_0^{-2}$ is trivially continuous.

6. $\sup_{\theta\in\mathcal{N}} \left\| \frac{\partial^2 Q_n(\theta)}{\partial\theta\partial\theta'} - \frac{\partial^2 Q(\theta)}{\partial\theta\partial\theta'} \right\| \xrightarrow{p} 0$ is for free here because $Q_n'' = Q'' = -\lambda^{-2}$, so the expression is identically equal to zero.

7. $\mathbf{H} = -\lambda_0^{-2}$ is nonzero and therefore nonsingular.

The asymptotic variance then is $\mathbf{H}^{-1}\Sigma\mathbf{H}^{-1} = \left(-\lambda_0^{-2}\right)^{-1}\lambda_0^{-2}\left(-\lambda_0^{-2}\right)^{-1} = \lambda_0^2$.

Bonus Content: Reminder of Delta Method. We should be able to recover the same result through the Delta Method. Indeed, a simple application of this method yields

$$\sqrt{n}(\hat\theta - \theta_0) \xrightarrow{d} N(0, \sigma^2) \implies \sqrt{n}(1/\hat\theta - 1/\theta_0) \xrightarrow{d} N(0, \sigma^2/\theta_0^4).$$

In the example, we have $\sigma^2 = 1/\lambda_0^2$ but also $\theta_0 = 1/\lambda_0$, so we get the expected limit distribution of $N(0, \lambda_0^2)$.

Warning: In this particular example, it turns out that "the variance of one over the estimator is one over the variance of the estimator." This is coincidental!

1.4 The estimator is upward biased. Note that $\lambda = 1/\mathbb{E}Y$. The function $f(x) = x^{-1}$ is convex, hence $\mathbb{E}(1/\overline{Y}) > 1/\mathbb{E}\overline{Y}$ by Jensen's inequality (equality cannot obtain because no parameter value that we allow makes $\overline{Y}$ a degenerate r.v.).

I gave 7 points for just noting that expectation does not pass through nonlinear functions. The following answers are really just wrong: "$E(1/\overline{Y}) = 1/\mathbb{E}\overline{Y}$" and "ML estimators are unbiased."

1.5 The constrained estimator is just $\widetilde{\lambda} = 1$ and the above expressions simplify at that point, so from a simplicity point of view, my pick would have been

$$LM \equiv n\frac{\partial Q_n(\tilde\theta)}{\partial \theta'}\tilde\Sigma^{-1}\frac{\partial Q_n(\tilde\theta)}{\partial \theta}.$$

Write

$$\begin{aligned} \left.\frac{\partial Q_n(\lambda)}{\partial \lambda}\right\|_{\lambda=1} &= \frac{1}{n}\sum_i [1 - y_i] = 1 - \overline{y} \\ \widetilde{\Sigma} &= 1^{-2} = 1, \end{aligned}$$

so the test statistic is

$$LM = n\left(1-\overline{y}\right)^2$$

We can also derive

$$\begin{aligned} LR &= 2n\big(Q_n(\hat{\lambda})-Q_n(\tilde{\lambda})\big) = 2n\left(\frac{1}{n}\sum_i\left[\log\hat{\lambda}-\hat{\lambda}y_i\right]-\frac{1}{n}\sum_i\left[\log 1-y_i\right]\right) \\ &= 2n\left(\log\left(1/\overline{y}\right)-\frac{1}{n}\sum_i y_i/\overline{y}-0+\frac{1}{n}\sum_i y_i\right) = 2n\left(\overline{y}-\log\overline{y}-1\right) \end{aligned}$$

and

$$\begin{aligned} W &\equiv \sqrt{n}r(\hat{\lambda})'\big(r'(\hat{\lambda})\hat{\Sigma}^{-1}r'(\hat{\lambda})\big)^{-1}\sqrt{n}r(\hat{\lambda}) \\ &= n(\hat{\lambda}-1)^2/(\hat{\lambda}^{-2}) = n\left(\frac{1}{\overline{y}}-1\right)^2\overline{y}^2 = n\left(1-\overline{y}\right)^2, \end{aligned}$$

using that $r(\lambda)=\lambda-1$ with constant derivative $r'(\lambda)=1$.

All have asymptotic distribution χ^2_1. This is especially intuitive with LM, which is the square of $\sqrt{n}\left(\overline{y}-1\right)$, but the latter is easily seen to go to standard normal under H_0. (This also illustrates that the Wald statistic is just a two-sided t-test here.)

Bonus content: Asymptotic equivalence The test statistics are not numerically equivalent, but I claimed in lecture that they are asymptotically equivalent. In the example, this can be verified. Write

$$\begin{aligned} \Delta_n(\overline{y}) &\equiv (LM-LR)/n = (1-\overline{y})^2-2(\overline{y}-\log\overline{y}-1) \\ &= 3+\overline{y}^2-4\overline{y}+2\log\overline{y} \\ \Longrightarrow \Delta_n'(\overline{y}) &= 2\overline{y}-4+2/\overline{y} \\ \Delta_n''(\overline{y}) &= 2-2/\overline{y}^2. \end{aligned}$$

We can therefore write the second order expansion of $\Delta_n(\cdot)$ about 1 as

$$\Delta_n(\overline{y}) = (3+1-4+0)+(2-4+2)(\overline{y}-1)+\tfrac{1}{2}(2-2)(\overline{y}-1)^2+O((\overline{y}-1)^3) = O((\overline{y}-1)^3),$$

and so for $\overline{y}-1=O_P(n^{-1/2})$ we get

$$\Delta_n(\overline{y}) = O_P(n^{-3/2}) \implies LM-LR = n\Delta_n = O_P(n^{-1/2}).$$

Since we also know that both test statistics converge to χ^2_1 (in particular, they cannot concentrate at the critical value), it follows that the probability of them disagreeing vanishes.

1.6 Bootstrap should outperform asymptotic approximation because the test statistic is an asymptotic pivot (and we can think of this test as equal-tailed two-sided).

1.7 We know the exact population distribution of Y under the null and can therefore formulate an exact finite-sample hypothesis test.

I didn't ask the following, but to see the idea, suppose $n = 1$. Under the null, the single observation that we see has p.d.f. $f(y) = \exp(-y)$ and c.d.f. $F(y) = \int_{t=0}^{y} \exp(-t)dt = [-\exp(-t)]_0^y = 1 - \exp(-y)$. At 5% significance, an exact equal-tailed two-sided test of the null has nonrejection region $[.025, 3.7]$. Note that an equal-tailed test based on an asymmetric distribution is typically not symmetric about the null parameter value.

2.1 The first equation is selection and the next two are outcome. Parameters of substantive interest are $(\alpha_2, \beta_2, \alpha_3, \beta_3)$ or in most cases really (β_2, β_3) (both answers accepted). Intuitively, there are two outcome equations of interest and each data point selects into exactly one of them. For example, each subject could choose the (for them!) more attractive of two jobs and we see outcomes for the chosen option ("Roy model").

2.2 After setting $\sigma_1 = 1$, the model is identified by the exact arguments from lecture notes. To argue this, it suffices to point out that really we are looking at two Tobit Type 2 models whose selection equations are the logical complements of each other (but we can analyze their identification separately).

Note: The question contained a slight mistake. I was not interested in, and consequently overlooked, σ_{23}. Since we have no joint observations of Y_2, Y_3 (even disregarding censoring), σ_{23} is not identified beyond the implications of the overall variance being positive definite. This did not affect anybody's grade because nobody spotted it (and I obviously did not deduct points).

2.3 The ML objective function for Tobit Type 2, see lecture notes.

2.4 Heckit; see lecture notes.

2.5 Yes, but it would be very inefficient. For example, under the assumptions we could just do OLS and that is necessarily more efficient (by comparison of variance expressions). But also...

2.6 That is actually efficient because Z_3 can be used as overidentifying instrument (very similar to SURE). Two-stage GMM is not necessary because homoskedasticity was assumed.

2.7 No! Z_3 will not be independent of, nor uncorrelated with, the composite error term in the observed outcome equations. Indeed, if it were, we would never have had a problem to begin with since X_3 fulfils the assumptions we make about Z_3.