ECOM40006/90013 Econometrics 3

Week 7 Extras: Solutions

Question 1: Intuition for the likelihood function

(a) Via independence, the likelihood function can be split up into the product of marginal densities:

$$\begin{aligned} f(x_1, x_2, \ldots, x_n) &= f_1(x_1) f_2(x_2) \ldots f_n(x_n) \\ &= \prod_{i=1}^{n} f_i(x_i) \\ &= \prod_{i=1}^{n} f(x_i) \end{aligned}$$

where the last line comes from the assumption of identical distributions. One thing worth noting is that if we don't make this assumption, it is possible for each individual data point to have its own distinct marginal density, as given by the subscript i on $f_i(x_i)$. The identical distributions assumption essentially says that $f_1(x) = f_2(x) = \cdots = f_n(x) = f(x)$ for all $i = 1, \ldots, n$.

(b) The log of the likelihood function is

$$\log L(\theta; x_i) = \log \left(\prod_{i=1}^{n} f(x_i) \right) = \sum_{i=1}^{n} \log f(x_i).$$

This is simply an extension of the rule that $\log(mn) = \log(m) + \log(n)$.

(c) The short answer is that x_i is stochastic. In other words: even if we held θ fixed, there are many different possible draws of x_i we can make.

In different terms, θ governs the data, but we don't actually observe θ in practice; only the observed data. Using maximum likelihood, it's possible to go back the other way: from observed x_i, we can make inference on what the underlying x_i happens to be. That is: in theory,

$$\theta \overset{\text{generates}}{\Longrightarrow} x_i$$

but maximum likelihood aims to go the other way:

$$x_i \overset{\text{infers}}{\Longrightarrow} \theta$$

Be aware that this method is subjective. You can probably guess that this is contingent on specifying the correct distribution from which x_i is generated. There are certain conditions under we can still get consistent estimates even if we mess this up, but we leave the discussion of this for another day. You may look up *pseudo maximum likelihood* if you are interested.

Question 2: Maximum likelihood, the univariate case

In this section, we abbreviate $\log L(\theta; x_i)$ *to just* $\log L(\theta)$.

(a) The log-likelihood, with expansion, is

$$\begin{aligned}\log L(\theta) &= \sum_{i=1}^{n} \log f(x_i) \\ &= \sum_{i=1}^{n} \log\left(\frac{1}{\theta}\exp\left[-\frac{x_i}{\theta}\right]\right) \\ &= \sum_{i=1}^{n}\left(-\log\theta - \frac{x_i}{\theta}\right) \\ &= -n\log\theta - \frac{1}{\theta}\sum_{i=1}^{n} x_i.\end{aligned}$$

(b) The MLE solves

$$\begin{aligned}&\frac{\partial \log L(\theta)}{\partial\theta} = -\frac{n}{\theta} + \frac{1}{\theta^2}\sum_{i=1}^{n} x_i = 0 \\ \implies &\frac{1}{\theta^2}\sum_{i=1}^{n} x_i = \frac{n}{\theta} \\ \implies &\frac{1}{n}\sum_{i=1}^{n} x_i = \frac{\theta^2}{\theta} \\ &\implies \hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i\end{aligned}$$

In words: the sample mean is the maximum likelihood estimator for θ when the data are drawn from this type of exponential distribution.

(c) The second derivative is

$$\frac{\partial^2 \log L(\theta)}{\partial\theta^2} = \frac{n}{\theta^2} - \frac{2}{\theta^3}\sum_{i=1}^{n} x_i.$$

If we evaluate this as $\hat{\theta}$ as required for a maximum, then observe that by our previous calculations,

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i \implies \sum_{i=1}^{n} x_i = n\hat{\theta}$$

so that

$$\begin{aligned}\left.\frac{\partial^2 \log L(\theta)}{\partial \theta^2}\right|_{\theta=\hat{\theta}} &= \frac{n}{\hat{\theta}^2} - \frac{2n\hat{\theta}}{\hat{\theta}^3} \\ &= \frac{n}{\hat{\theta}^2} - \frac{2n}{\hat{\theta}^2} \\ &= -\frac{n}{\hat{\theta}^2} \\ &< 0\end{aligned}$$

since n must be positive and $\hat{\theta} > 0$ as every $x_i > 0$. Therefore, $\hat{\theta}$ represents a maximum.

(d) Repeating the same steps as before, the log-likelihood is

$$\begin{aligned}\log L(\theta) &= \sum_{i=1}^{n} \log(\theta(1-\theta)^{x_i}) \\ &= \sum_{i=1}^{n} [\log\theta + x_i \log(1-\theta)] \\ &= n\log\theta + \log(1-\theta)\sum_{i=1}^{n} x_i\end{aligned}$$

with the FOC with respect to θ implying the MLE:

$$\begin{aligned}\frac{\partial \log L(\theta)}{\partial\theta} = \frac{n}{\theta} - \frac{1}{1-\theta}\sum_{i=1}^{n} x_i = 0 \\ \implies \frac{1}{1-\theta}\sum_{i=1}^{n} = \frac{n}{\theta} \\ \implies \underbrace{\frac{1}{n}\sum_{i=1}^{n} x_i}_{:=\bar{x}} = \frac{n}{\theta} \\ \implies \bar{x} = \frac{1}{\theta} - 1 \\ \implies \bar{x} + 1 = \frac{1}{\theta} \\ \implies \hat{\theta} = \frac{1}{1+\bar{x}}.\end{aligned}$$

The second derivative is

$$\frac{\partial^2 \log L(\theta)}{\partial\theta^2} = -\frac{n}{\theta^2} - \frac{1}{(1-\theta)^2}\sum_{i=1}^{n} x_i$$

where we observe that $\hat{\theta} > 0$ by construction and $\sum_{i=1}^{n} x_i \geq 0$. Hence the second-order derivative is negative when evaluated at $\hat{\theta}$, confirming a maximum.

Question 3: Bivariate maximum likelihood

(a) The log-likelihood function is

$$\begin{aligned}\log L(\theta) &= \sum_{i=1}^{n} \log f(y_i;\theta) \\ &= \sum_{i=1}^{n} \log\left[\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)\right] \\ &= -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\end{aligned}$$

(b) The gradient vector is given by

$$G(\theta) = \frac{\partial \log L(\theta)}{\partial \theta} = \begin{bmatrix} \dfrac{\partial \log L(\theta)}{\partial \mu} \\ \dfrac{\partial \log L(\theta)}{\partial \sigma^2} \end{bmatrix} = \begin{bmatrix} \dfrac{1}{\sigma^2}\displaystyle\sum_{i=1}^{n}(y_i-\mu) \\ -\dfrac{n}{2\sigma^2} + \dfrac{1}{2\sigma^4}\displaystyle\sum_{i=1}^{n}(y_i-\mu)^2 \end{bmatrix}$$

Keep in mind that it is a common algebraic error to take a derivative with respect to σ instead of σ^2 in the second expression.

(c) The first-order conditions are as follows:

$$\begin{aligned}\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i-\mu) &= 0 \\ -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(y_i-\mu)^2 &= 0.\end{aligned}$$

Solving the first one yields

$$\sum_{i=1}^{n} y_i - \sum_{i=1}^{n}\mu = 0 \implies \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

Solving the second expression directly also yields

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{\mu})^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y}).$$

These expressions are the sample mean and variance of our data, respectively.

(d) The Hessian is a 2×2 matrix with the following components:

$$H(\theta) = \frac{\partial^2 \log L(\theta)}{\partial \theta^2} = \begin{bmatrix} \dfrac{\partial^2 \log L(\theta)}{\partial \mu^2} & \dfrac{\partial^2 \log L(\theta)}{\partial \mu \partial \sigma^2} \\ \dfrac{\partial^2 \log L(\theta)}{\partial \sigma^2 \partial \mu} & \dfrac{\partial^2 \log L(\theta)}{\partial (\sigma^2)^2} \end{bmatrix}$$

The individual components may be calculated as follows:

$$\frac{\partial^2 \log L(\theta)}{\partial \mu^2} = -\frac{n}{\sigma^2}$$
$$\frac{\partial^2 \log L(\theta)}{\partial (\sigma^2)^2} = \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{i=1}^{n}(y_i - \mu)^2$$
$$\frac{\partial^2 \log L(\theta)}{\partial \mu \partial \sigma^2} = \frac{\partial^2 \log L(\theta)}{\partial \sigma^2 \partial \mu} = -\frac{1}{\sigma^4}\sum_{i=1}^{n}(y_i - \mu)$$

(e) Let's first consider the cross-partials when evaluated at the MLE. From our earlier conditions for $\hat{\mu}$, we know that

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i \implies n\hat{\mu} = \sum_{i=1}^{n} y_i \implies \sum_{i=1}^{n}(y_i - \hat{\mu}) = 0.$$

When plugged into the Hessian evaluated at $\hat{\theta}$, this results in the cross-partials being equal to zero. If we consider the element in the lower right corner of the Hessian, let's use the fact that from the MLEs, we can derive that

$$\hat{\sigma}^2 = n\sum_{i=1}^{n}(y_i - \hat{\mu})^2$$

so that

$$\begin{aligned}\left.\frac{\partial^2 \log L(\theta)}{\partial (\sigma^2)^2}\right|_{\theta=\hat{\theta}} &= \frac{n}{2\hat{\sigma}^4} - \frac{1}{\hat{\sigma}^6}\sum_{i=1}^{n}(y_i - \hat{\mu})^2 \\ &= \frac{n}{2\hat{\sigma}^4} - \frac{1}{\hat{\sigma}^6}(n\hat{\sigma}^2) \\ &= \frac{n}{2\hat{\sigma}^4} - \frac{n}{\hat{\sigma}^4} \\ &= -\frac{n}{2\hat{\sigma}^4}\end{aligned}$$

leading us to the final expression for the Hessian evaluated at the MLE:

$$H(\hat{\theta}) = \begin{bmatrix} -\frac{n}{\hat{\sigma}^2} & 0 \\ 0 & -\frac{n}{2\hat{\sigma}^4} \end{bmatrix}$$

The method of leading principal minors can be used to calculate negative definiteness. In this case we have

- The first leading principal minor is negative by default since $\hat{\sigma}^2 > 0$ and $n > 0$.
- The second leading principal minor is

$$-\frac{n}{\hat{\sigma}^2} \times -\frac{n}{2\hat{\sigma}^4} - 0 \times 0 = \frac{n^2}{2\hat{\sigma}^6} > 0.$$

Putting these two together imply that the Hessian is negative definite at $\hat{\theta}$, and therefore a maximum is attained.

Question 4: More univariate maximum likelihood

(a) Begin by abbreviating $p(y_i|\theta)$ to $p(\theta)$ and $L(\theta|y_i)$ to $L(\theta)$ in an abuse of notation . The likelihood function is simply

$$\begin{aligned} L(\theta) &= p(y_1, y_2, \ldots, y_n) \\ &= p(y_1)p(y_2)\ldots p(y_n) && (y_i \text{ i.i.d.}) \\ &= \prod_{i=1}^{n} p(y_i) \\ &= \prod_{i=1}^{n} \theta^{y_i}(1-\theta)^{1-y_i} \end{aligned}$$

(b) The log-likelihood function can be obtained by taking the log of the likelihood function derived in (a):

$$\begin{aligned} \log L(\theta) &= \log\left(\prod_{i=1}^{n} p(y_i)\right) \\ &= \sum_{i=1}^{n} \log p(y_i) \\ &= \sum_{i=1}^{n} \log(\theta^{y_i}(1-\theta)^{1-y_i}) \\ &= \sum_{i=1}^{n} [y_i \log\theta + (1-y_i)\log(1-\theta)] \\ &= \log\theta \sum_{i=1}^{n} y_i + \log(1-\theta)\sum_{i=1}^{n}(1-y_i) \end{aligned}$$

(c) Before taking the gradient, first note via the Chain Rule

$$\frac{\partial \log(1-\theta)}{\partial\theta} = \frac{1}{1-\theta} \times (-1) = -\frac{1}{1-\theta}.$$

Then, the gradient, or score function, is given as

$$\begin{aligned} S(\theta) &= \frac{\partial \log L(\theta)}{\partial\theta} \\ &= \frac{\partial}{\partial\theta}\left[\log\theta \sum_{i=1}^{n} y_i + \log(1-\theta)\sum_{i=1}^{n}(1-y_i)\right] \\ &= \frac{1}{\theta}\sum_{i=1}^{n} y_i - \frac{1}{1-\theta}\sum_{i=1}^{n}(1-y_i) \end{aligned}$$

(d) Before moving to the Hessian, or second derivative, first note that

$$\frac{\partial}{\partial\theta}(1-\theta)^{-1} = -(1-\theta)^{-2} \times (-1) = \frac{1}{(1-\theta)^2}$$

Now let's examine the Hessian, or second derivative:

$$\begin{aligned} H(\theta) &= \frac{\partial^2 \log L(\theta)}{\partial \theta^2} \\ &= \frac{\partial}{\partial \theta}\left[\frac{1}{\theta}\sum_{i=1}^{n} y_i - \frac{1}{1-\theta}\sum_{i=1}^{n}(1-y_i)\right] \\ &= -\frac{1}{\theta^2}\sum_{i=1}^{n} y_i - \frac{1}{(1-\theta)^2}\sum_{i=1}^{n}(1-y_i) \end{aligned}$$

(e) To solve for the maximum likelihood estimator, let's return to the score function. Setting it to zero allows us to solve for the first-order conditions. Doing so gives

$$\begin{aligned} &\frac{1}{\theta}\sum_{i=1}^{n} y_i - \frac{1}{1-\theta}\sum_{i=1}^{n}(1-y_i) = 0 \\ &\qquad \implies \frac{1}{\theta}\sum_{i=1}^{n} y_i = \frac{1}{1-\theta}\sum_{i=1}^{n}(1-y_i) \\ &\qquad\qquad = \frac{n}{1-\theta} - \frac{1}{1-\theta}\sum_{i=1}^{n} y_i \\ &\implies \underbrace{\left(\frac{1}{\theta} + \frac{1}{1-\theta}\right)}_{=\frac{1}{\theta(1-\theta)}}\sum_{i=1}^{n} y_i = \frac{n}{1-\theta} \\ &\implies \frac{1}{\theta(1-\theta)}\sum_{i=1}^{n} y_i = \frac{n}{1-\theta} \\ &\qquad \implies \frac{1}{\theta}\sum_{i=1}^{n} y_i = n \\ &\qquad\qquad \implies \hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} y_i \end{aligned}$$

So here, the maximum likelihood estimator (MLE), θ, is the sample average.

(f) Now that we're taking expectations, first note that one of the properties of the Bernoulli distribution is $\mathbb{E}(y_i) = \theta_0$, as mentioned in the question. Note that this also implies that $\mathbb{E}(1-y_i) = 1 - \mathbb{E}(y_i) = 1 - \theta_0$. First, evaluate the score at the true value θ_0 and call it

$S(\theta_0)$. Then,

$$\begin{aligned}
\mathbb{E}(S(\theta_0)) &= \mathbb{E}\left[\frac{1}{\theta_0}\sum_{i=1}^{n} y_i - \frac{1}{1-\theta_0}\sum_{i=1}^{n}(1-y_i)\right] \\
&= \frac{1}{\theta_0}\sum_{i=1}^{n}\mathbb{E}(y_i) - \frac{1}{1-\theta_0}\sum_{i=1}^{n}\mathbb{E}(1-y_i) \\
&= \frac{1}{\theta_0}\sum_{i=1}^{n}\theta_0 - \frac{1}{1-\theta_0}\sum_{i=1}^{n}(1-\theta_0) \\
&= \frac{n\theta_0}{\theta_0} - \frac{n(1-\theta_0)}{1-\theta_0} \\
&= n - n \\
&= 0
\end{aligned}$$

So the expected score is zero... as expected. ;)

(g) In order to calculate the variance of the score, it can get a bit tricky to do it directly. So what we're going to do is to take a step back and first calculate the variance of the score *for a single observation*.

This works because we're looking at

$$\log L(\theta) = \sum_{i=1}^{n} \log p(y_i),$$

which is in fact the sum of log-likelihoods for single data points. So if we focused our attention on just one of those, say $\log p(y_i)$, then the derivative of that would be the score for a single observation. So in this case, we have

$$s_i(\theta_0) = \frac{y_i}{\theta_0} - \frac{1-y_i}{1-\theta_0}$$

and so the variance of the score for a single observation is

$$\begin{aligned}
\text{Var}(s_i(\theta_0)) &= \text{Var}\left(\frac{y_i}{\theta_0} - \frac{1-y_i}{1-\theta_0}\right) \\
&= \text{Var}\left(\frac{y_i}{\theta_0} + \frac{y_i}{1-\theta_0} - \frac{1}{1-\theta_0}\right) \\
&= \text{Var}\left(\frac{y_i}{\theta_0} + \frac{y_i}{1-\theta_0}\right) && \text{(Variance ignores constants)} \\
&= \text{Var}\left(\left[\frac{1}{\theta_0} + \frac{1}{1-\theta_0}\right] y_i\right) \\
&= \text{Var}\left(\frac{1}{\theta_0(1-\theta_0)} y_i\right) \\
&= \frac{1}{(\theta_0(1-\theta_0))^2}\text{Var}(y_i) && \because \text{Var}(aX) = a^2\text{Var}(X) \\
&= \frac{\theta_0(1-\theta_0)}{(\theta_0(1-\theta_0))^2} && \because \text{Var}(y_i) = \theta_0(1-\theta_0) \\
&= \frac{1}{\theta_0(1-\theta_0)}.
\end{aligned}$$

Then, the variance of the score can be obtained by taking the variance of the sum:

$$\begin{aligned}
\text{Var}(S(\theta_0)) &= \text{Var}\left(\sum_{i=1}^{n} s_i(\theta_0)\right) \\
&= \sum_{i=1}^{n} \text{Var}(s_i(\theta_0)) && (y_i \text{ is i.i.d.}) \\
&= n\text{Var}(s_i(\theta_0)) \\
&= \frac{n}{\theta_0(1-\theta_0)}.
\end{aligned}$$

(h) Since we already have an expression for $\text{Var}(S)$, all we need to find is the negative expected Hessian. Evaluating the Hessian at the true value $\theta = \theta_0$ and taking the negative expectation, we find:

$$\begin{aligned}
-\mathbb{E}(H(\theta_0)) &= \mathbb{E}\left[\frac{1}{\theta_0^2}\sum_{i=1}^{n} y_i + \frac{1}{(1-\theta_0)^2}\sum_{i=1}^{n}(1-y_i)\right] \\
&= \frac{1}{\theta_0^2}\sum_{i=1}^{n}\mathbb{E}(y_i) + \frac{1}{(1-\theta_0)^2}\sum_{i=1}^{n}\mathbb{E}(1-y_i) \\
&= \frac{1}{\theta_0^2}\sum_{i=1}^{n}\theta_0 + \frac{1}{(1-\theta_0)^2}\sum_{i=1}^{n}(1-\theta_0) \\
&= \frac{n\theta_0}{\theta_0^2} + \frac{n(1-\theta_0)}{(1-\theta_0)^2} \\
&= \frac{n}{\theta_0} + \frac{n}{1-\theta_0} \\
&= n\left(\frac{1}{\theta_0} + \frac{1}{1-\theta_0}\right) \\
&= \frac{n}{\theta_0(1-\theta_0)}
\end{aligned}$$

which matches our expression for $\text{Var}(S)$ from before, hence demonstrating the result $-\mathbb{E}(H) = \text{Var}(S)$ as required.