

## Question 1

**c**

For these hypotheses, I would use the function  $f(\theta, \sigma) = \theta - \theta_0$ , which is zero under the null, and non-zero under the alternative. All three test statistics (Wald, Score and Likelihood-ratio) converge to the same distribution under the null, that is a chi-square with the degrees of freedom being the dimension of the  $f(\theta, \sigma)$  (in this case, it is  $k+1$ ). Thus, the test would be carried by comparing the test statistic to the critical value to the chi-square distribution with  $k+1$  degrees of freedom such that the probability of being below this value is significance level desired. If the test statistic is above this value, then we would reject the null (and not reject otherwise).

Below I compute the three test statistics and compare them.

### Score Test

The score test statistic is:

$$T_n = nS' \left\{ D_{\theta, \sigma} f(\tilde{\theta}_n, \tilde{\sigma}_n) \tilde{\Omega}_n D_{\theta, \sigma} f(\tilde{\theta}_n, \tilde{\sigma}_n)' \right\}^{-1} S$$

$$S = D_{\theta, \sigma} f(\tilde{\theta}_n, \tilde{\sigma}_n) (\tilde{B}_n)^{-1} D_{\theta, \sigma} L_n(\tilde{\theta}_n, \tilde{\sigma}_n)$$

Where  $\tilde{\theta}_n$  and  $\tilde{\sigma}_n$  are the estimators obtained by maximizing the likelihood function subject to the constraint:  $f(\theta, \sigma) = \theta - \theta_0 = 0$ ; and  $\tilde{\Omega}_n$  is a consistent estimator of  $\Omega = B^{-1}AB^{-1} = (-B)^{-1}$  (under the regularity conditions for the information equality to hold).

B can be written as follows (considering  $\theta$  and  $\sigma$  as the true parameters):

$$B = \mathbb{E} \begin{pmatrix} D_{\theta, \theta'}^2 \log p_{(\theta, \sigma)}(y_i | x_i) & D_{\theta, \sigma}^2 \log p_{(\theta, \sigma)}(y_i | x_i) \\ D_{\sigma, \theta'}^2 \log p_{(\theta, \sigma)}(y_i | x_i) & D_{\sigma, \sigma}^2 \log p_{(\theta, \sigma)}(y_i | x_i) \end{pmatrix}$$

$$= \begin{pmatrix} \mathbb{E} \left[ D_{\theta'} \left( \frac{x_i y_i - x_i x_i' \theta}{\sigma^2} \right) \right] & D_{\sigma} \{ D_{\theta} \mathbb{E} [\log p_{(\theta, \sigma)}(y_i | x_i)] \} \\ D_{\theta'} \{ D_{\sigma} \mathbb{E} [\log p_{(\theta, \sigma)}(y_i | x_i)] \} & \mathbb{E} \left[ D_{\sigma} \left( -\frac{1}{\sigma} + \frac{(y_i - x_i' \theta)^2}{\sigma^3} \right) \right] \end{pmatrix}$$

$$= \begin{pmatrix} -\frac{1}{\sigma^2} \mathbb{E} [x_i x_i'] & 0 \\ 0 & \frac{1}{\sigma^2} - \frac{3}{\sigma^4} \mathbb{E} [(y_i - x_i' \theta)^2] \end{pmatrix}$$

where the off-diagonal terms are zero because the regularity conditions imply that, using the true  $\theta$  and  $\sigma$ :  $D_{\theta} \mathbb{E} [\log p_{(\theta, \sigma)}(y_i | x_i)] = 0 = D_{\sigma} \mathbb{E} [\log p_{(\theta, \sigma)}(y_i | x_i)]$  (and also that we move derivatives in and out of the expectation).

Thus, using WLLN and CMT, a consistent estimate of  $\Omega$  is (since  $\tilde{\theta} = \theta_0$  and  $\tilde{\sigma}_n^2 = \frac{1}{n} \sum (y_i - x_i' \tilde{\theta})^2$  are consistent):

$$\tilde{\Omega}_n = (-\tilde{B}_n)^{-1} = \begin{pmatrix} \frac{1}{\tilde{\sigma}_n^2} \left( \frac{1}{n} \sum x_i x_i' \right) & 0 \\ 0 & \frac{1}{\tilde{\sigma}_n^2} - \frac{3}{\tilde{\sigma}_n^4} \frac{1}{n} \sum (y_i - x_i' \tilde{\theta})^2 \end{pmatrix}^{-1} = \begin{pmatrix} \tilde{\sigma}_n^2 \left( \frac{1}{n} \sum x_i x_i' \right)^{-1} & 0 \\ 0 & \frac{1}{2} \tilde{\sigma}_n^2 \end{pmatrix}$$

And, since  $D_{\theta, \sigma} f(\tilde{\theta}_n, \tilde{\sigma}_n) = (\mathbb{I}_{k+1} \ 0)$ , the term inside brackets in  $T_n$  is:

$$\left\{ (\mathbb{I}_{k+1} \ 0) \begin{pmatrix} \tilde{\sigma}_n^2 \left( \frac{1}{n} \sum x_i x_i' \right)^{-1} & 0 \\ 0 & \frac{1}{2} \tilde{\sigma}_n^2 \end{pmatrix} \begin{pmatrix} \mathbb{I}_{k+1} \\ 0 \end{pmatrix} \right\}^{-1} = \frac{1}{\tilde{\sigma}_n^2} \left( \frac{1}{n} \sum x_i x_i' \right)$$

Also, we have that:

$$\begin{aligned}
D_{\theta,\sigma}L_n(\tilde{\theta}_n, \tilde{\sigma}_n) &= \begin{pmatrix} \frac{1}{\tilde{\sigma}_n^2} \frac{1}{n} \sum x_i(y_i - x_i'\tilde{\theta}_n) \\ -\frac{1}{\tilde{\sigma}_n} + \frac{1}{\tilde{\sigma}_n^3} \frac{1}{n} \sum (y_i - x_i'\tilde{\theta}_n)^2 \end{pmatrix} \\
D_{\theta,\sigma}f(\tilde{\theta}_n, \tilde{\sigma}_n)(\tilde{B}_n)^{-1} &= \begin{pmatrix} -\tilde{\sigma}_n^2 \left(\frac{1}{n} \sum x_i x_i'\right)^{-1} & 0 \end{pmatrix} \\
S &= \begin{pmatrix} -\tilde{\sigma}_n^2 \left(\frac{1}{n} \sum x_i x_i'\right)^{-1} & 0 \end{pmatrix} \begin{pmatrix} \frac{1}{\tilde{\sigma}_n^2} \frac{1}{n} \sum x_i(y_i - x_i'\tilde{\theta}_n) \\ -\frac{1}{\tilde{\sigma}_n} + \frac{1}{\tilde{\sigma}_n^3} \frac{1}{n} \sum (y_i - x_i'\tilde{\theta}_n)^2 \end{pmatrix} = \\
&= -\tilde{\sigma}_n^2 \left(\frac{1}{n} \sum x_i x_i'\right)^{-1} \frac{1}{\tilde{\sigma}_n^2} \frac{1}{n} \sum x_i(y_i - x_i'\tilde{\theta}_n)
\end{aligned}$$

Thus, the test statistic is:

$$\begin{aligned}
T_n &= n \left\{ - \left( \frac{1}{n} \sum x_i x_i' \right)^{-1} \frac{1}{n} \sum x_i(y_i - x_i'\tilde{\theta}_n) \right\}' * \\
&\quad * \left\{ \frac{1}{\tilde{\sigma}_n^2} \left( \frac{1}{n} \sum x_i x_i' \right) \right\} * \\
&\quad * \left\{ - \left( \frac{1}{n} \sum x_i x_i' \right)^{-1} \frac{1}{n} \sum x_i(y_i - x_i'\tilde{\theta}_n) \right\} \\
&= n \frac{1}{\tilde{\sigma}_n^2} \frac{1}{n} \sum (y_i - x_i'\tilde{\theta}_n) x_i' \left( \frac{1}{n} \sum x_i x_i' \right)^{-1} \frac{1}{n} \sum x_i(y_i - x_i'\tilde{\theta}_n)
\end{aligned}$$

We can also show that:

$$\begin{aligned}
\frac{1}{n} \sum (y_i - x_i'\tilde{\theta}_n) x_i' &= \frac{1}{n} \sum (y_i - x_i'\hat{\theta}_n + x_i'\hat{\theta}_n - x_i'\theta_0) x_i' \\
&= \frac{1}{n} \sum (y_i - x_i'\hat{\theta}_n) x_i' + \frac{1}{n} \sum x_i'(\hat{\theta}_n - \theta_0) x_i' \\
&= (\hat{\theta}_n - \theta_0)' \left( \frac{1}{n} \sum x_i x_i' \right)
\end{aligned}$$

where we used the first order condition for  $\hat{\theta}_n$  to eliminate the first term after the second equality and the fact that  $x_i'(\hat{\theta}_n - \theta_0)$  is a scalar, to transpose it and get the final expression.

This brings  $T_n$  closer to the Wald test (and thus the OLS case, since we see below that the Wald test is the same in both cases, and in the OLS case it was equal to the score test, as seen in a previous problem set), but it is still different, due to the estimator of the variance in the denominator being different in this case:

$$T_n = n \frac{1}{\tilde{\sigma}_n^2} (\hat{\theta}_n - \theta_0)' \left( \frac{1}{n} \sum x_i x_i' \right) (\hat{\theta}_n - \theta_0)$$

### Wald Test

The Wald test statistic is:

$$T_n = n f(\hat{\theta}_n, \hat{\sigma}_n) \left\{ D_{\theta,\sigma}f(\hat{\theta}_n, \hat{\sigma}_n) \hat{\Omega}_n D_{\theta,\sigma}f(\hat{\theta}_n, \hat{\sigma}_n)' \right\}^{-1} f(\hat{\theta}_n, \hat{\sigma}_n)'$$

Using the expression for  $B$  above, we can obtain consistent estimate of  $\Omega$  similarly, due to the consistency of  $\hat{\theta}_n, \hat{\sigma}_n$ :

$$\hat{\Omega}_n = \begin{pmatrix} \hat{\sigma}_n^2 \left( \frac{1}{n} \sum x_i x_i' \right)^{-1} & 0 \\ 0 & \frac{1}{2} \hat{\sigma}_n^2 \end{pmatrix}$$

Also, using  $D_{\theta, \sigma} f(\hat{\theta}_n, \hat{\sigma}_n) = (\mathbb{I}_{k+1} \ 0)$ , we get that the inside term of  $T_n$  is:

$$\left\{ (\mathbb{I}_{k+1} \ 0) \begin{pmatrix} \hat{\sigma}_n^2 \left( \frac{1}{n} \sum x_i x_i' \right)^{-1} & 0 \\ 0 & \frac{1}{2} \hat{\sigma}_n^2 \end{pmatrix} \begin{pmatrix} \mathbb{I}_{k+1} \\ 0 \end{pmatrix} \right\}^{-1} = \frac{1}{\hat{\sigma}_n^2} \left( \frac{1}{n} \sum x_i x_i' \right)$$

This gives us:

$$T_n = n(\hat{\theta} - \theta_0)' \frac{1}{\hat{\sigma}_n^2} \left( \frac{1}{n} \sum x_i x_i' \right) (\hat{\theta} - \theta_0)$$

Since  $\hat{\theta}$  was equal to the OLS estimator and  $\frac{1}{\hat{\sigma}_n^2} \left( \frac{1}{n} \sum x_i x_i' \right)$  is a consistent estimate of the asymptotic variance of the OLS estimator under homoskedasticity (which is satisfied here), the Wald test is the same in both cases.

**Likelihood Ratio Test** The test statistic for this one is:

$$T_n = 2 \left( L_n(\hat{\theta}_n, \hat{\sigma}_n) - L_n(\tilde{\theta}_n, \tilde{\sigma}_n) \right)$$

where again  $L_n(\tilde{\theta}_n, \tilde{\sigma}_n)$  is the log-likelihood evaluated at the estimator obtained by maximizing it under the restriction given by  $f(\theta, \sigma) = 0$ .

## Question 2

**a**

Since the density function  $p_\theta(X_i)$  for each  $X_i$  is  $\frac{1}{\theta}$  if this  $X_i$  is between  $\theta$  and  $2\theta$ , and zero otherwise, the likelihood function can be written as:

$$l_n(\theta) = \frac{1}{\theta^n} \mathbb{1}_{\{\theta \leq X_i \leq 2\theta, \forall i\}}$$

For the range where the indicator function is one, this is a decreasing function of  $\theta$  (everywhere else it is zero). So we must pick the smallest  $\hat{\theta}_n$  such that  $\min X_i \geq \hat{\theta}_n$  and  $2\hat{\theta}_n \geq \max X_i$ . Putting these together we get that  $\hat{\theta}_n$  must be the smallest that satisfy:  $\min X_i \geq \hat{\theta}_n \geq \frac{1}{2} \max X_i$ . We know that  $\min X_i \geq \frac{1}{2} \max X_i$  is always true, because  $\min X_i \geq \theta \geq \frac{1}{2} \max X_i$  must hold for the true  $\theta$ . Thus the smallest  $\hat{\theta}_n$  satisfying the condition is just  $\hat{\theta}_n = \frac{1}{2} \max X_i$ .

**b**

No. Since expectation is a linear operator,  $\mathbb{E}[\hat{\theta}_n] = \mathbb{E}[\frac{1}{2} \max X_i] = \frac{1}{2} \mathbb{E}[\max X_i]$ . Since  $\max X_i$  also can take values between  $\theta$  and  $2\theta$ , unless the distribution of  $X_i$  is degenerate,  $\mathbb{E}[\max X_i]$  will be strictly between these two values and thus we have  $\mathbb{E}[\hat{\theta}_n] = \frac{1}{2} \mathbb{E}[\max X_i] < \frac{1}{2} 2\theta = \theta$ .

We can also see this calculating  $\mathbb{E}[\max X_i]$ .

First notice that:

$$\begin{aligned} G(x) &:= \Pr(\max X_i \leq x) &= [\Pr(X_i \leq x)]^n = \left[\frac{x-\theta}{\theta}\right]^n \\ g(x) &:= n \left[\frac{x-\theta}{\theta}\right]^{n-1} \frac{1}{\theta} \end{aligned}$$

where  $G()$  is the cdf and  $g()$  the pdf of  $\max X_i$ , and we used the fact that the cdf of the  $Unif(\theta, 2\theta)$  is  $\frac{x-\theta}{\theta}$ .

Thus, we can calculate  $\mathbb{E}[\max X_i]$ :

$$\begin{aligned} \mathbb{E}[\max X_i] &= \int_{\theta}^{2\theta} x n \left[\frac{x-\theta}{\theta}\right]^{n-1} \frac{1}{\theta} dx \\ &= \left[ x \left[\frac{x-\theta}{\theta}\right]^n \right]_{\theta}^{2\theta} - \int_{\theta}^{2\theta} \left[\frac{x-\theta}{\theta}\right]^{n-1} \left[\frac{x-\theta}{\theta}\right] dx \\ &= 2\theta - \left( \int_{\theta}^{2\theta} x \left[\frac{x-\theta}{\theta}\right]^{n-1} \frac{1}{\theta} dx + \int_{\theta}^{2\theta} \left[\frac{x-\theta}{\theta}\right]^{n-1} dx \right) \frac{n}{\theta} \\ &= 2\theta - \frac{1}{n} \int_{\theta}^{2\theta} x n \left[\frac{x-\theta}{\theta}\right]^{n-1} \frac{1}{\theta} dx + \frac{\theta}{n} \int_{\theta}^{2\theta} n \left[\frac{x-\theta}{\theta}\right]^{n-1} \frac{1}{\theta} dx \\ &= 2\theta - \frac{1}{n} \mathbb{E}[\max X_i] + \frac{\theta}{n} \\ &\iff \\ \mathbb{E}[\max X_i] &= \frac{n}{n+1} 2\theta + \frac{1}{n+1} \theta \\ &\iff \\ \mathbb{E}[\hat{\theta}_n] &= \frac{n}{n+1} \theta + \frac{0.5\theta}{n+1} = \theta \frac{(n+0.5)}{n+1} < \theta \end{aligned}$$

**c**

For any  $\epsilon > 0$ , we have:

$$\begin{aligned}
\Pr(|\hat{\theta}_n - \theta| > \frac{\epsilon}{2}) &= \Pr(\theta - \frac{1}{2} \max X_i > \frac{\epsilon}{2}) \\
&= \Pr(\max X_i < 2(\theta - \frac{\epsilon}{2})) \\
&\leq \Pr(\max X_i \leq 2(\theta - \frac{\epsilon}{2})) \\
&= [\Pr(X_i \leq 2\theta - \epsilon)]^n \rightarrow 0
\end{aligned}$$

where we used that  $\theta \geq \frac{1}{2} \max X_i$  in the first equality, the fact that the observations are iid in the last equality, and that  $\Pr(X_i \leq 2\theta - \epsilon) = \frac{\theta - \epsilon}{\theta} < 1$ .

Thus,  $\hat{\theta}_n \xrightarrow{P} \theta$ .

## d

Let's look at the cdf of  $n(\theta - \hat{\theta}_n)$ . Since  $n(\theta - \hat{\theta}_n) \geq 0$  always, as seen in letter (a), we have that  $\Pr(n(\theta - \hat{\theta}_n) \leq x) = 0$  if  $x < 0$ . If  $x \geq 0$ :

$$\begin{aligned}
\Pr(n(\theta - \hat{\theta}_n) \leq x) &= \Pr(\hat{\theta}_n \geq \theta - \frac{x}{n}) \\
&= 1 - \Pr(\max X_i < 2(\theta - \frac{x}{n})) \\
&= 1 - [\frac{\theta - \frac{2x}{n}}{\theta}]^n \\
&= 1 - [1 - \frac{1}{n} \frac{2x}{\theta}]^n \rightarrow 1 - \exp(\frac{-x}{0.5\theta})
\end{aligned}$$

Thus we have:

$$\Pr(n(\theta - \hat{\theta}_n) \leq x) \rightarrow \begin{cases} 0, & \text{if } x < 0 \\ 1 - \exp(\frac{-x}{\lambda}), & \text{if } x \geq 0 \end{cases}$$

where  $\lambda = 0.5\theta$ .

## e

Notice first that we do not know the critical values of the exponential distribution with parameter  $\lambda$ , since we do not know  $\theta$  and  $\lambda$  depends on  $\theta$ . But we can multiply  $n(\theta - \hat{\theta}_n)$  by  $\frac{1}{0.5\hat{\theta}_n}$ . This gives us, using slusky, consistency of  $\hat{\theta}_n$  and the fact that we can scale the exponential distribution:

$$\frac{1}{0.5\hat{\theta}_n} n(\theta - \hat{\theta}_n) \xrightarrow{d} \frac{1}{0.5\theta} z = z*$$

where  $z$  is exponential with parameter  $\lambda = 0.5\theta$  and  $z^*$  is exponential with parameter 1, and we can look the critical values for  $z^*$ .

Call  $e_{0.95}$  the critical value such that  $\Pr(z^* \leq e_{0.95}) = 0.95$ . Then we have:

$$\begin{aligned}\Pr\left(\theta \in \left[\hat{\theta}_n, \hat{\theta}_n + \frac{e_{0.95}}{2n}\hat{\theta}_n\right]\right) &= \Pr\left(0 \leq \frac{1}{0.5\hat{\theta}_n}n(\theta - \hat{\theta}_n) \leq e_{0.95}\right) \\ &= \Pr\left(\frac{1}{0.5\hat{\theta}_n}n(\theta - \hat{\theta}_n) \leq e_{0.95}\right) \rightarrow \Pr(z^* \leq e_{0.95}) = 0.95\end{aligned}$$

where we used the fact that the exponential distribution only takes non negative values.

Thus,  $[\hat{\theta}_n, \hat{\theta}_n + \frac{e_{0.95}}{2n}\hat{\theta}_n]$  is an approximate 95% confidence interval for  $\theta$ .

## Question 4

**a**

The conditional (on  $x$ ) likelihood function is:

$$l_n(\theta) = \prod G(x'_i\theta)^{y_i}(1 - G(x'_i\theta))^{1-y_i}$$

And the conditional log-likelihood function is:

$$L_n(\theta) = \frac{1}{n} \sum y_i \log G(x'_i\theta) + (1 - y_i) \log (1 - G(x'_i\theta))$$

**b**

The general condition for a unique maximum of  $L(\theta)$  is that  $\Pr\{p_\theta(y|x) \neq p_{\theta_0}(y|x)\} > 0$  for any  $\theta \neq \theta_0$ . In this case, we can write this as follows, since  $y \in \{0, 1\}$ :

$$\begin{aligned}\Pr\{G(x'\theta)^y(1 - G(x'\theta))^{1-y} \neq G(x'\theta_0)^y(1 - G(x'\theta_0))^{1-y}\} &> 0 \\ &\iff \\ \Pr\{G(x'\theta) \neq G(x'\theta_0)\} &> 0\end{aligned}$$

If  $G(\cdot)$  is strictly increasing, this is equivalent to  $\Pr\{x'(\theta - \theta_0) \neq 0\} > 0$  for any  $(\theta - \theta_0) \neq 0$ ; that is, it is a requirement that there is no perfect colinearity in  $x$ .

**c**

Let's look at the  $L_n(\theta)$  for this dataset:

$$\begin{aligned}L_n(\theta) = \frac{1}{4} \Big[ &\log \frac{\exp(\theta_0 + \theta_1)}{1 + \exp(\theta_0 + \theta_1)} + \log \frac{1}{1 + \exp(\theta_0 + 0.8\theta_1)} \\ &+ \log \frac{\exp(\theta_0 + 2\theta_1)}{1 + \exp(\theta_0 + 2\theta_1)} + \log \frac{1}{1 + \exp(\theta_0 + 0.5\theta_1)} \Big]\end{aligned}$$

I can choose  $\theta_0$  and  $\theta_1$  to make the terms in the exponentials negative for when  $y_i = 0$  and positive otherwise (that is, to “separate” the observations according to the value of  $y_i$ ). For instance,  $\theta_0 = -0.9j$  and  $\theta_1 = j$ :

$$\begin{aligned} L_n(\theta) &= \frac{1}{4} \left[ \log \frac{\exp(0.1j)}{1 + \exp(0.1j)} + \log \frac{1}{1 + \exp(-0.1j)} \right. \\ &\quad \left. + \log \frac{\exp(1.1j)}{1 + \exp(1.1j)} + \log \frac{1}{1 + \exp(-0.4j)} \right] \\ &= \frac{1}{4} \left[ \log \frac{\exp(0.1j)}{1 + \exp(0.1j)} + \log \frac{\exp(0.1j)}{1 + \exp(0.1j)} \right. \\ &\quad \left. + \log \frac{\exp(1.1j)}{1 + \exp(1.1j)} + \log \frac{\exp(0.4j)}{1 + \exp(0.4j)} \right] \end{aligned}$$

But now we can see that, as  $j \rightarrow \infty$ ,  $L_n(\theta)$  goes to zero, since the terms inside the logs are approaching 1. Thus we can make  $L_n(\theta)$  as close to zero as we want by choosing  $j$  high enough. And, for any arbitrary values of  $\theta_0$  and  $\theta_1$  (that “separate” observations or not), the expressions inside the logs are strictly between 0 and 1, and thus  $L_n(\theta) < 0$ . Therefore, given any values for  $\theta_0^*$  and  $\theta_1^*$ , we can find  $j$  such that  $L_n(\theta)$  evaluated at  $\theta_0 = -0.9j$  and  $\theta_1 = j$  gives a higher log-likelihood than evaluated at  $\theta_0^*$  and  $\theta_1^*$ . Thus the ML estimator does not exist here.

## d

With the appropriate regularity conditions,  $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, I^{-1})$ , where  $I = -\mathbb{E} [D_{\theta, \theta'}^2 \log p_{\theta_0}(y_i | x_i)]$  is the information matrix. Let’s calculate the matrix  $I$ .

First, the gradient vector of  $L_n(\theta)$  is:

$$\begin{aligned} D_{\theta} L_n &= \frac{1}{n} \sum \frac{y_i x_i}{G(x_i' \theta)} g(x_i' \theta) - \frac{(1 - y_i) x_i}{1 - G(x_i' \theta)} g(x_i' \theta) \\ &= \frac{1}{n} \sum \frac{(1 - G(x_i' \theta)) g(x_i' \theta) y_i x_i - G(x_i' \theta) g(x_i' \theta) x_i + G(x_i' \theta) g(x_i' \theta) y_i x_i}{G(x_i' \theta) (1 - G(x_i' \theta))} \\ &= \frac{1}{n} \sum \frac{g(x_i' \theta) y_i x_i - G(x_i' \theta) g(x_i' \theta) x_i}{G(x_i' \theta) (1 - G(x_i' \theta))} \\ &= \frac{1}{n} \sum \frac{g(x_i' \theta) x_i (y_i - G(x_i' \theta))}{G(x_i' \theta) (1 - G(x_i' \theta))} = \frac{1}{n} \sum x_i (y_i - G(x_i' \theta)) \end{aligned}$$

where we used that, in the logit model,  $\frac{g(x_i' \theta)}{G(x_i' \theta) (1 - G(x_i' \theta))} = 1$ , where  $g(z) := \frac{dG(z)}{dz} = \frac{\exp(z)}{(1 + \exp(z))^2}$ .

Thus, we have (using the fact that observations are iid):

$$\begin{aligned} I &= -\mathbb{E} [D_{\theta, \theta'}^2 \log p_{\theta_0}(y_i|x_i)] \\ &= \mathbb{E} \left[ \frac{1}{n} \sum x_i g(x'_i \theta) x'_i \right] \\ &= \mathbb{E} [g(x'_i \theta) x_i x'_i] = \mathbb{E} \left[ \frac{\exp(x'_i \theta)}{(1 + \exp(x'_i \theta))^2} x_i x'_i \right] \end{aligned}$$

**e**

Since  $\Pr(y_i = 1|x_i = x) = G(x'\theta_0)$ ,  $D_x \Pr(y_i = 1|x_i = x) = g(x'\theta_0)\theta_0$ . This is equal  $\frac{\exp(x'_i \theta_0)}{(1 + \exp(x'_i \theta_0))^2} \theta_0$  in the logit case.

Using the ML estimator for  $\theta_0$  ( $\hat{\theta}_n$ ) we can estimate  $D_x \Pr(y_i = 1|x_i = x)$  by  $g(x'\hat{\theta}_n)\hat{\theta}_n$ . In the logit case:  $\frac{\exp(x'_i \hat{\theta}_n)}{(1 + \exp(x'_i \hat{\theta}_n))^2} \hat{\theta}_n$ .

Due to the invariance property, this estimator is actually the same as the ML of  $D_x \Pr(y_i = 1|x_i = x)$ . That is, due to invariance, if we are interested in a function of  $\theta_0$  (such as  $g(x'\theta_0)\theta_0$  here), the ML estimator of this function will be the function applied to the ML estimator of  $\theta_0$ .

**f**

In part (d) we saw that  $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, I^{-1})$ , where  $I$  is the information matrix. Since now we are interested in a function ( $f(\theta) = g(x'\theta)\theta$ ) of the estimator, we can use the delta method.

First notice that:

$$D_\theta f(\theta) = g'(x'\theta)\theta x' + Id_{k+1}g(x'\theta)$$

Thus, by the delta method:

$$\sqrt{n}(g(x'\hat{\theta}_n)\hat{\theta}_n - g(x'\theta_0)\theta_0) \xrightarrow{d} N(0, D_\theta f(\theta)I^{-1}D_\theta f(\theta)')$$

## Question 5

**a**

$$\begin{aligned} \mathbb{E}[X_i] &= (1 + \theta) \int_0^1 x^{\theta+1} dx \\ &= (1 + \theta) \left[ \frac{x^{\theta+2}}{\theta+2} \right]_0^1 = \frac{\theta+1}{\theta+2} = \mu \end{aligned}$$



Expressing  $\theta$  as a function of  $\mu$ :

$$\begin{aligned}\frac{\theta + 1}{\theta + 2} &= \mu \\ \theta + 1 &= \theta\mu + 2\mu \\ \theta(1 - \mu) &= 2\mu - 1 \\ \theta &= \frac{2\mu - 1}{1 - \mu}\end{aligned}$$

## b

The log-likelihood function as a function of  $\theta$  is:

$$\begin{aligned}L_n(\theta) &= \frac{1}{n} \sum_i \log(1 + \theta) + \theta \log(x_i) \\ &= \log(1 + \theta) + \theta \frac{\sum_i \log(x_i)}{n} \\ &= \log(1 + \theta) + \theta \overline{\ln X_n}\end{aligned}$$

where  $\overline{\ln X_n}$  is the average of  $\log(x_i)$ .

Thus, the log-likelihood function as a function of  $\mu$  is:

$$\begin{aligned}L_n(\mu) &= \log\left(1 + \frac{2\mu - 1}{1 - \mu}\right) + \frac{2\mu - 1}{1 - \mu} \overline{\ln X_n} \\ &= \log\left(\frac{1 - \mu + 2\mu - 1}{1 - \mu}\right) + \frac{2\mu - 1}{1 - \mu} \overline{\ln X_n} \\ &= \log(\mu) - \log(1 - \mu) + \frac{2\mu - 1}{1 - \mu} \overline{\ln X_n}\end{aligned}$$

## c

Let's look at the FOC:

$$\begin{aligned}\frac{dL_n(\mu)}{d\mu} &= \frac{1}{\mu} + \frac{1}{1 - \mu} + \frac{2 - 2\mu + 2\mu - 1}{(1 - \mu)^2} \overline{\ln X_n} = 0 \\ \frac{(1 - \mu)^2 + \mu(1 - \mu) + \mu \overline{\ln X_n}}{\mu(1 - \mu)^2} &= 0 \\ (1 - \mu)^2 + \mu(1 - \mu) + \mu \overline{\ln X_n} &= 0 \\ (1 - \mu)(1 - \mu + \mu) + \mu \overline{\ln X_n} &= 0 \\ \hat{\mu}_n &= \frac{1}{1 - \overline{\ln X_n}}\end{aligned}$$

Looking at the second order condition:

$$\begin{aligned}\frac{dL_n(\mu)}{d\mu^2} &= -\frac{1}{\mu^2} + \frac{1}{(1-\mu)^2} + \frac{2}{(1-\mu)^3} \overline{\ln X_n} \\ &= \frac{-(1-\mu)^2 + \mu^2}{\mu^2(1-\mu)^2} + \frac{2}{(1-\mu)^3} \overline{\ln X_n} \\ &= \frac{2\mu - 1}{\mu^2(1-\mu)^2} + \frac{2}{(1-\mu)^3} \overline{\ln X_n}\end{aligned}$$

The second term is always negative, because  $0 < x_i < 1$  for all  $i$  gives  $\overline{\ln X_n}$ , and  $0 < \mu < 1$ . But the first term can be positive for  $\mu > 0.5$ . Thus whether or not we can find the maximum likelihood estimator through the first order conditions may depend on the sample and the true  $\mu$ .

**d**

Let's first show that  $\mathbb{E}[\log(x_i)] = -\frac{1}{1+\theta}$ , using the hint given in the question in the third equality:

$$\begin{aligned}\mathbb{E}[\log(x_i)] &= (1+\theta) \int_0^1 \log(x_i) x_i^\theta dx \\ &= (1+\theta) \left\{ \left[ \frac{x_i^{\theta+1}}{1+\theta} \log(x_i) \right]_{x \rightarrow 0}^1 - \int_0^1 \frac{1}{x_i} \frac{x_i^{\theta+1}}{1+\theta} dx \right\} \\ &= -(1+\theta) \int_0^1 \frac{x_i^\theta}{1+\theta} dx = -\frac{1}{1+\theta}\end{aligned}$$

Now, WLLN gives us that  $\overline{\ln X_n} \xrightarrow{p} \mathbb{E}[\log(x_i)] = -\frac{1}{1+\theta}$ , since the  $x_i$  (and thus also the  $\log(x_i)$ ) are iid. Thus, using the continuous mapping theorem (we can use, since the function  $f(x) = \frac{1}{1-x}$  is continuous when  $x < 0$  and  $-\frac{1}{1+\theta} < 0$ ) we have:

$$\hat{\mu}_n = \frac{1}{1 - \overline{\ln X_n}} \xrightarrow{p} \frac{1}{1 + \frac{1}{1+\theta}} = \frac{\theta + 1}{\theta + 2} = \mu$$

**e**

Let's first show that  $\mathbb{E}[\log(x_i)^2] = \frac{2}{(1+\theta)^2}$ , using the hint given in the question and the result of letter (d):

$$\begin{aligned}\mathbb{E}[\log(x_i)^2] &= (1+\theta) \int_0^1 \log(x_i)^2 x_i^\theta dx \\ &= (1+\theta) \left\{ \left[ \frac{x_i^{\theta+1}}{1+\theta} \log(x_i)^2 \right]_{x \rightarrow 0}^1 - \int_0^1 2 \log(x_i) \frac{1}{x_i} \frac{x_i^{\theta+1}}{1+\theta} dx \right\} \\ &= -2 \int_0^1 \log(x_i) x_i^\theta dx = -2 \left( -\frac{1}{(1+\theta)^2} \right) = \frac{2}{(1+\theta)^2}\end{aligned}$$

From the CLT we know that  $\sqrt{n}(\overline{\ln X_n} - \mathbb{E}[\log(x_i)]) \xrightarrow{d} N(0, \text{Var}[\log(x_i)])$ . We can then use the Delta method with the function  $f(x) = \frac{1}{1-x}$  ( $\frac{df(x)}{dx} = \frac{1}{(1-x)^2}$ ) to obtain:

$$\begin{aligned} \sqrt{n}(f(\overline{\ln X_n}) - f(-\frac{1}{1+\theta})) &\xrightarrow{d} \left. \frac{df(x)}{dx} \right|_{x=-\frac{1}{1+\theta}} * N(0, \text{Var}[\log(x_i)]) \\ \sqrt{n}(\hat{\mu}_n - \mu) &\xrightarrow{d} \frac{1}{(1 + \frac{1}{1+\theta})^2} * N(0, \text{Var}[\log(x_i)]) \\ \sqrt{n}(\hat{\mu}_n - \mu) &\xrightarrow{d} N\left(0, \frac{(\theta+1)^4}{(\theta+2)^4} \text{Var}[\log(x_i)]\right) \\ \sqrt{n}(\hat{\mu}_n - \mu) &\xrightarrow{d} N\left(0, \frac{(\theta+1)^2}{(\theta+2)^4}\right) = N(0, \mu^2(1-\mu)^2) \end{aligned}$$

where we used, in the last line, that  $\text{Var}[\log(x_i)] = \frac{2}{(1+\theta)^2} - [-\frac{1}{1+\theta}]^2 = \frac{1}{(\theta+1)^2}$ , and that  $1 - \mu = \frac{1}{\theta+2}$ .

## f

The information matrix is:

$$\begin{aligned} -\mathbb{E}\left[\frac{dL_n(\mu)}{d\mu^2}\right] &= -\frac{2\mu-1}{\mu^2(1-\mu)^2} - \frac{2}{(1-\mu)^3} \mathbb{E}[\overline{\ln X_n}] \\ &= -\frac{2\mu-1}{\mu^2(1-\mu)^2} - \frac{2}{(1-\mu)^3} \left(-\frac{1}{1+\theta}\right) \\ &= -\frac{2\mu-1}{\mu^2(1-\mu)^2} - \frac{2}{(1-\mu)^3} \left(-\frac{1-\mu}{1-\mu+2\mu-1}\right) \\ &= -\frac{2\mu-1}{\mu^2(1-\mu)^2} - \frac{2}{(1-\mu)^3} \left(-\frac{1-\mu}{\mu}\right) \\ &= -\frac{2\mu-1}{\mu^2(1-\mu)^2} + \frac{2}{\mu(1-\mu)^2} \\ &= \frac{-2\mu+1+2\mu}{\mu^2(1-\mu)^2} = \frac{1}{\mu^2(1-\mu)^2} \end{aligned}$$

Thus we can see that the asymptotic variance of the maximum likelihood estimator of  $\mu$  found in letter (e) is the inverse of the information matrix.

## g

Consider the function  $f(\mu) = \mu - \frac{2}{3}$ . Under the null, its value is zero. The test statistic for the score test is: The score test statistic is:

$$\begin{aligned} T_n &= nS' \left\{ D_\mu f(\hat{\mu}_n) \hat{\Omega}_n D_\mu f(\hat{\mu}_n)' \right\}^{-1} S \\ S &= D_\mu f(\hat{\mu}_n) (\hat{B}_n)^{-1} D_\mu L_n(\hat{\mu}) \end{aligned}$$

In this case,  $D_\mu f(\hat{\mu}_n) = 1$ , and under the null:

$$D_\mu L_n(\hat{\mu}) = \frac{3}{2} + \frac{3}{1} + \frac{9}{1} \overline{\ln X_n} = \frac{9}{2} + 9 \overline{\ln X_n}$$

And we saw that  $B = (-\mu^2(1 - \mu)^2)^{-1}$ , so we can estimate it consistently as  $\hat{B}_n^{-1} = -\hat{\mu}^2(1 - \hat{\mu})^2$  (and also  $\hat{\Omega}_n = \hat{\mu}^2(1 - \hat{\mu})^2$ , since we saw above that  $\Omega = \mu^2(1 - \mu)^2$ ).

Then,  $T_n$  becomes:

$$\begin{aligned} T_n &= n \left( \frac{9}{2} + 9 \overline{\ln X_n} \right) (-\mu^2(1 - \mu)^2) (\hat{\mu}^2(1 - \hat{\mu})^2)^{-1} (-\mu^2(1 - \mu)^2) \left( \frac{9}{2} + 9 \overline{\ln X_n} \right) \\ &= n \left( \frac{9}{2} + 9 \overline{\ln X_n} \right)^2 \left( \frac{4}{9} * \frac{1}{9} \right) = n \left( \frac{81}{4} + 2 \frac{81 \overline{\ln X_n}}{2} + 81 (\overline{\ln X_n})^2 \right) \frac{4}{81} = n(1 + 4 \overline{\ln X_n} + 4 (\overline{\ln X_n})^2) \end{aligned}$$

We know from class that, under the null, this converges to a chi-square with one degree of freedom (the dimension of the restriction). Thus we can compare this value to the critical value to this distribution such that the probability of a draw from the distribution being below this critical value is  $1 - \alpha$ . If our test statistic is above this value, we reject the null. This test has level  $\alpha$ .