M1 INTERMEDIATE ECONOMETRICS

MAXIMUM LIKELIHOOD ESTIMATION

Koen Jochmans

November 13, 2025

1. Intuition and examples

1.1. Binary data

If the binary variable $Y \in \{0,1\}$ is Bernoulli distributed with $\theta = \mathbb{P}(Y=1)$ its probability mass function at $y \in \{0,1\}$ is

$$f(y;\theta) = \mathbb{P}(Y=y) = \theta^y \times (1-\theta)^{1-y}.$$

Now consider a random sample of size n from this distribution, $Y_1, Y_2, \ldots, Y_n$. The probability that this sequence is equal to $(y_1, y_2, \ldots, y_n) \in \{0,1\}^n$ equals

$$\mathbb{P}(Y_1 = y_1, \ldots, Y_n = y_n) = \prod_{i=1}^n f(y_i;\theta) = \prod_{i=1}^n \theta^{y_i} \times (1-\theta)^{1-y_i}.$$

Drawing a sample that consists of only successes, that is $y_i = 1$ for all $1 \leq i \leq n$, happens with probability θ^n.

Now suppose that we do not know the actual value θ_0 of the success probability that actually generated our sample. We can compute the above probability for any value of $\theta \in (0,1)$. As a function of θ, this probability is

$$L_n(\theta) = \prod_{i=1}^n f(Y_i;\theta) = \prod_{i=1}^n \theta^{Y_i} \times (1-\theta)^{1-Y_i}.$$

We call L_n the likelihood function. The value $L_n(\theta)$ is the probability of observing the sample that has been drawn if the success probability were

equal to θ.

A sensible estimator of the success probability θ_0 based on the data is

$$\hat{\theta} = \arg\max_{\theta} L_n(\theta).$$

This is the maximum-likelihood estimator. Notice that, here, $\hat{\theta}$ is equal to the value of the success probability that makes it most likely to generate the sample that we have in front of us.

It is easy to calculate $\hat{\theta}$ in this example. It is helpful to work with the monotone transformation

$$\ell_n(\theta) = \log L_n(\theta).$$

This is the log-likelihood function. Clearly, the maximiser of the log-likelihood function is the same as the maximiser of the likelihood function. However, the log-likelihood function is easier to work with. Indeed, here,

$$\ell_n(\theta) = \sum_{i=1}^{n} Y_i \log(\theta) + (1 - Y_i) \log(1 - \theta),$$

because the log of a product becomes the sum of the logs. In this example the log-likelihood can be further simplifies, as it depends on the data only through the mean

$$\bar{Y}_n = n^{-1} \sum_{i=1}^{n} Y_i.$$

Indeed,

$$n^{-1}\ell_n(\theta) = \bar{Y}_n \log(\theta) + (1 - \bar{Y}_n) \log(1 - \theta).$$

This is a globally-concave function of θ. To find its maximizer we solve the

first-order condition

$$n^{-1}\frac{\partial \ell_n(\theta)}{\partial \theta} = \frac{\bar{Y}_n}{\theta} - \frac{1-\bar{Y}_n}{1-\theta} = \frac{\bar{Y}_n - \theta}{\theta(1-\theta)} = 0,$$

to find $\hat{\theta} = \bar{Y}_n$ as maximum-likelihood estimator. Thus, we estimate the actual success probability by the proportion of successes observed in the sample.

1.2. Exponential data

Our next example deals with a continuous variable. The exponential distribution with mean θ has probability density function

$$f(y;\theta) = \frac{e^{-y/\theta}}{\theta}, \qquad y \geq 0.$$

The log-likelihood function is

$$\ell_n(\theta) = \log\left(\prod_{i=1}^{n} f(Y_i;\theta)\right) = \sum_{i=1}^{n} \log f(Y_i;\theta) = -n\left(\log\theta + \frac{\bar{Y}_n}{\theta}\right).$$

The maximum-likelihood estimator solves

$$n^{-1}\frac{\partial \ell_n(\theta)}{\partial \theta} = -\frac{1}{\theta} + \frac{\bar{Y}_n}{\theta^2} = 0$$

and so again equals $\hat{\theta} = \bar{Y}_n$, the sample mean.

1.3. The probit model

We now consider a conditional model where, in addition to binary Y, we now also have a vector of covariates X. We wish to allow the success probability

to depend on X. A parsimonious way to do so is to let

$$\mathbb{P}(Y = 1|X = x) = F(x'\theta)$$

for a chosen cumulative distribution function F, such as the standard-normal distribution in case of the probit model. In this case we know the distribution of Y conditional on X,

$$f(y|x;\theta) = \mathbb{P}(Y = y|X = x) = F(x'\theta)^y \times (1 - F(x'\theta))^{1-y}.$$

The conditional likelihood is $L_n(\theta_n) = \prod_{i=1}^n f(Y_i|X_i;\theta)$. The log-likelihood is

$$\ell_n(\theta) = \sum_{i=1}^n Y_i \log(F(X_i'\theta)) + (1 - Y_i) \log(1 - F(X_i'\theta)).$$

Here, the first-order condition for the parameter θ is the nonlinear system,

$$\frac{\partial \ell_n(\theta)}{\partial \theta} = \sum_{i=1}^n X_i f(X_i'\theta) \frac{Y_i - F(X_i'\theta)}{F(X_i'\theta)\,(1 - F(X_i'\theta))} = 0,$$

where f is the density function associated with F. The maximum-likelihood estimator is thus defined as an implicit function here. The likelihood is nonetheless globally concave in θ and so a numerical optimisation routine such as Newton-Raphson, for example, will find the global optimum in few iterations.

Figure 1 shows the screen output from a probit fit in Stata. The data are from Mroz (1987) and concern the decision of a sample of 753 married women to participate to the labor force as a function of various characteristics. These characteristics include the number of children the household has aged less than 6, 6 or up, the women's age, education, and experience, as well as the

Figure 1: Probit model for female labor-force participation

```
Iteration 0:  Log likelihood =  -514.8732
Iteration 1:  Log likelihood = -408.37815
Iteration 2:  Log likelihood = -407.46153
Iteration 3:  Log likelihood = -407.46038
Iteration 4:  Log likelihood = -407.46038

Probit regression                                     Number of obs =    753
                                                      LR chi2(6)    = 214.83
                                                      Prob > chi2   = 0.0000
Log likelihood = -407.46038                           Pseudo R2     = 0.2086
```

inlf	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
kidslt6	-.871957	.116976	-7.45	0.000	-1.101226	-.6426883
kidsge6	.0274219	.042933	0.64	0.523	-.0567253	.1115691
age	-.0575521	.008267	-6.96	0.000	-.0737552	-.0413491
educ	.1269323	.0247196	5.13	0.000	.0784828	.1753817
exper	.0719188	.0074972	9.59	0.000	.0572247	.0866129
huswage	-.0229368	.0125263	-1.83	0.067	-.0474878	.0016143
_cons	.6771326	.4930208	1.37	0.170	-.2891704	1.643435

husbands' wage. The top of the output shows how the point estimates were obtained by maximizing the log-likelihood function; the global maximum was obtained after 3 iterations.

Figure 2: Marginal effects in the probit model

```
Predictive margins                                    Number of obs = 753
Model VCE: OIM

Expression: Pr(inlf), predict()
1._at: kidslt6 = 0
2._at: kidslt6 = 1
3._at: kidslt6 = 2
4._at: kidslt6 = 3
5._at: kidslt6 = 4
```

	Margin	Delta-method std. err.	z	P>|z|	[95% conf.	interval]
_at						
1	.6373647	.0173223	36.79	0.000	.6034136	.6713158
2	.3611582	.0303623	11.89	0.000	.3016492	.4206672
3	.1407216	.0383346	3.67	0.000	.065587	.2158561
4	.0347353	.0208998	1.66	0.097	-.0062276	.0756982
5	.0050725	.0055501	0.91	0.361	-.0058055	.0159504

In the binary-choice model marginal effects are functions of the covariate values. In Figure 2 we report the estimated average (over the covariates) participation probability for different numbers of young children. Clearly, this probability is estimated to go down monotonically. For example, having one young child is estimated to bring down the participation probability from 0.637 to 0.361, on average. That is a decrease of 27.6 percentage points. Having additional young children brings the participation rate down further, although the increments decrease. A simple least-squares fit (not reported here) gives a coefficient estimate on `kidslt6` of -0.274. While this is close to our estimated impact of going from none to a single young child, it ignores the nonlinear response to the number of children, thereby largely overestimating the impact of having additional children on the decision to participate in the labor market.

1.4. The classical linear regression model

When

$$Y|X \sim N(X'\beta, \sigma^2)$$

the conditional density function of $Y|X$ is

$$f(y|x;\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(y-x'\beta)^2}{\sigma^2}\right).$$

Here, $\theta = (\beta, \sigma^2)$ and

$$L_n(\beta, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(Y_i - X_i'\beta)^2}{\sigma^2}\right)$$

so that

$$\ell_n(\beta, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2}\sum_{i=1}^{n}\frac{(Y_i - X_i'\beta)^2}{\sigma^2}.$$

The first-order conditions are

$$\frac{\partial \ell_n(\beta, \sigma^2)}{\partial \beta} = \sum_{i=1}^{n}\frac{X_i(Y_i - X_i'\beta)}{\sigma^2} = 0,$$

and

$$\frac{\partial \ell_n(\beta, \sigma^2)}{\partial \sigma^2} = \frac{1}{2}\left(\frac{\sum_{i=1}^{n}(Y_i - X_i'\beta)^2}{\sigma^4} - \frac{n}{\sigma^2}\right) = 0.$$

The solution is

$$\hat{\beta} = \left(\sum_{i=1}^{n} X_i X_i'\right)^{-1}\left(\sum_{i=1}^{n} X_i Y_i\right), \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i'\hat{\beta})^2,$$

which are by now familiar. Indeed, $\hat{\beta}$ is simply the least-squares estimator and $\hat{\sigma}^2$ is the traditional variance estimator, without any degrees-of-freedom correction.

1.5. The Tobit model

The Tobit model is the censoring problem with normal errors, that is,

$$Y = \max(Y^*, 0).$$

with

$$Y^* = X'\beta + e, \qquad e|X \sim N(0, \sigma^2).$$

We have that

$$\mathbb{P}(Y = 0|X) = \mathbb{P}(Y^* < 0|X) = \Phi(-X'\beta/\sigma) = 1 - \Phi(X'\beta/\sigma)$$

while, for any $y > 0$

$$\mathbb{P}(Y \leq y|X) = \mathbb{P}(Y = 0|X) + \mathbb{P}(Y \leq y|X, Y > 0)\,\mathbb{P}(Y > 0|X)$$

with

$$\mathbb{P}(Y \leq y|X, Y > 0)\,\mathbb{P}(Y > 0|X) = \mathbb{P}(Y \leq y|X) - \mathbb{P}(Y \leq 0|X)$$

so that

$$\mathbb{P}(Y \leq y|X) = \Phi\left(\frac{y - X'\beta}{\sigma}\right)$$

with density

$$\frac{1}{\sigma}\phi\left(\frac{y - X'\beta}{\sigma}\right).$$

The likelihood function thus is

$$L_n(\beta, \sigma^2) = \prod_{i=1}^{n} \left(1 - \Phi\left(\frac{X'\beta}{\sigma}\right)\right)^{\{Y_i=0\}} \left(\frac{1}{\sigma}\phi\left(\frac{Y_i - X_i'\beta}{\sigma}\right)\right)^{\{Y_i>0\}}$$

which has both a probit component and a normal-regression component. Again, finding the maximum-likelihood estimator in the Tobit model requires numerical optimisation.

2. Maximum likelihood

2.1. The population problem

Consider the population problem where we maximize the expected log-likelihood,

$$\mathbb{E}(\log f(Y|X;\theta)).$$

This function achieves its global maximum at θ_0. To see this let $h(x)$ be the density/mass function of X at x and observe that

$$\begin{aligned}
\mathbb{E}(\log f(Y|X;\theta)) - \mathbb{E}(\log f(Y|X;\theta_0)) &= \mathbb{E}\left(\log\left(\frac{f(Y|X;\theta\ )}{f(Y|X;\theta_0)}\right)\right) \\
&\leq \log \mathbb{E}\left(\frac{f(Y|X;\theta\ )}{f(Y|X;\theta_0)}\right) \\
&= \log \iint \frac{f(y|x;\theta\ )}{f(y|x;\theta_0)} f(y|x;\theta_0)\, dy h(x) dx \\
&= \log \iint f(y|x;\theta\ )\, dy\, h(x)\, dx \\
&= 0,
\end{aligned}$$

where we have used the fact that taking logs is a concave operation (so that we can apply Jensen's inequality) and the fact that $\int f(y|x;\theta)\, dy = 1$ for any θ.

The function

$$n^{-1}\ell_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log f(Y_i|X_i;\theta)$$

is a sample analog of the population problem. Under some suitable technical conditions

$$\frac{1}{n}\sum_{i=1}^{n} \log f(Y_i|X_i;\theta) \underset{p}{\to} \mathbb{E}(\log f(Y_i|X_i;\theta))$$

in a uniform sense. If the maximizer of the population problem is unique

this then provides an argument for consistency,

$$\hat{\theta} \underset{p}{\to} \theta_0$$

as $n \to \infty$. This is the intuition behind the 'argmax' theorem for maximum likelihood.

The uniqueness of the maximiser is important here. For example, in the classical linear model, the maximiser is not unique when the matrix $\mathbb{E}(XX')$ is singular, as we have seen. In this case the limit problem is flat in certain directions and so it has many global maximisers that cannot be distinguished from θ_0.

2.2. The score

An instructive alternative way to think about maximum likelihood is to see it as a device that delivers a good estimating equation. The score is defined as

$$s(y|x;\theta) = \frac{\partial \log f(y|x;\theta)}{\partial \theta}.$$

Note that, provided that the support of $f(y|x;\theta)$ does not change with θ, and assuming that $f(y|x;\theta)$ is differentiable,

$$\begin{aligned}
\mathbb{E}_\theta(s(Y|X;\theta)|X=x) &= \int \frac{\partial \log f(y|x;\theta)}{\partial \theta} f_\theta(y|x)\, dy \\
&= \int \frac{\partial f(y|x;\theta)}{\partial \theta}\, dy \\
&= \frac{\partial \int f(y|x;\theta)\, dy}{\partial \theta} = 0,
\end{aligned}$$

because $\int f(y|x;\theta)\, dy = 1$ for any θ. By iterated expectations,

$$\mathbb{E}(s(Y|X;\theta_0)) = 0,$$

that is, the score is also mean zero unconditionally. All our examples above were regular in this sense.

In our examples we found $\hat{\theta}$ by solving

$$n^{-1}\frac{\partial \ell_n(\hat{\theta})}{\partial \theta} = \frac{1}{n}\sum_{i=1}^{n} s(Y_i|X_i;\hat{\theta}) = 0.$$

This is a sample version of the moment condition just given. In this sense, we can think about maximum likelihood as a method-of-moment estimator, just like least squares or two-stage least squares.

In many problems other possible estimating equations exist. For example, in the (conditional) binary-choice problem looked at earlier the model states that

$$\mathbb{E}(Y|X = x) = F(x'\theta_0).$$

This implies that $\mathbb{E}(Y - F(X'\theta_0)|X = x) = 0$ and so, in turn, that, for any function G,

$$\mathbb{E}(G(X;\theta_0)\,(Y - F(X'\theta_0))) = 0.$$

We can thus in principle construct an estimator by solving

$$\frac{1}{n}\sum_{i=1}^{n} G(X_i;\hat{\theta})\,(Y_i - F(X_i'\hat{\theta})) = 0.$$

The score of maximum likelihood essentially gives us a specific G. In the binary choice problem we had

$$G(x;\theta_0) = x\,\frac{f(x'\theta_0)}{F(x'\theta_0)\,(1 - F(x'\theta_0))}$$

It turns out that the suggestion of maximum likelihood leads to optimality in large samples.

2.3. The information matrix

The information matrix is defined as the variance of the score, i.e.,

$$I = \text{var}\left(s(Y|X;\theta_0)\right) = \mathbb{E}\left(s(Y|X;\theta_0)\, s(Y|X;\theta_0)'\right).$$

The information equality states that

$$I = -\mathbb{E}\left(\frac{\partial^2 \log f(Y|X;\theta_0)}{\partial\theta\partial\theta'}\right),$$

so, up to its sign, the variance of the score (which cannot be negative) is equal to the Hessian of the population problem (which must be negative in regular problems). The information equality is important in understanding where the optimality of maximum likelihood comes from.

To see where the equality itself comes from, we start again from the conditional zero-mean condition of the score, i.e.,

$$\int \frac{\partial \log f(y|x;\theta)}{\partial\theta}\, f_\theta(y|x)\, dy = 0.$$

Differentiating this equation with respect to θ gives

$$\int \frac{\partial^2 \log f(y|x;\theta)}{\partial\theta\partial\theta'}\, f(y|x;\theta)\, dy + \int \frac{\partial \log f(y|x;\theta)}{\partial\theta}\, \frac{\partial f(y|x;\theta)}{\partial\theta'}\, dy = 0.$$

So, the terms on the left-hand side of this expression must be equal to each other up to sign. It is immediate that the first of these terms is the conditional expectation of the second derivative of the log density function. To see that the other term is the conditional variance of the score we substitute the simple equality

$$\frac{\partial f(y|x;\theta)}{\partial\theta'} = \frac{\log \partial f(y|x;\theta)}{\partial\theta'}\, f(y|x;\theta)$$

inside the second integral to get

$$\int \frac{\partial \log f_\theta(y|x)}{\partial \theta} \frac{\partial \log f_\theta(y|x)}{\partial \theta'} f_\theta(y|x)\, dy = \mathbb{E}_\theta(s(Y|X;\theta)\, s(Y|X;\theta)'|X = x).$$

Thus, we have shown that

$$\text{var}_\theta(s(Y|X;\theta)|X = x) = -\mathbb{E}_\theta\left(\frac{\partial^2 \log f(Y|X;\theta)}{\partial \theta \partial \theta'} \middle| X = x \right).$$

To finalise our calculations note that, by the law of total variance, we have

$$I = \text{var}\left(s(Y|X;\theta_0)\right) = \mathbb{E}\left(\text{var}\left(s(Y|X;\theta_0)|X\right)\right) + \text{var}\left(\mathbb{E}\left(s(Y|X;\theta_0)|X\right)\right)$$

and that the second right-hand side term is zero by the mean-zero property of the score. Therefore,

$$I = \mathbb{E}(\text{var}(s(Y|X;\theta_0)|X)) = -\mathbb{E}\left(\frac{\partial^2 \log f(Y|X;\theta_0)}{\partial \theta \partial \theta'} \right),$$

as claimed.

3. Asymptotic distribution

We consider only regular problems where the mass/density function $f(y|x;\theta)$ is twice continuously differentiable, the information matrix is non-singular, and some technical conditions collected below hold.

3.1. Derivation

In regular cases we can start from an expansion of the first-order condition

$$n^{-1} \frac{\partial \ell_n(\hat{\theta})}{\partial \theta} = 0$$

around θ_0 to write

$$n^{-1}\frac{\partial \ell_n(\theta_0)}{\partial \theta} + n^{-1}\frac{\partial^2 \ell_n(\theta_*)}{\partial \theta \partial \theta'}(\hat{\theta} - \theta_0) = 0,$$

where θ_* is determined by the mean-value theorem and satisfies $\theta_* \underset{p}{\to} \theta_0$. Now we can re-arrange the above expression to arrive at

$$\sqrt{n}(\hat{\theta} - \theta_0) = \left(-\frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2 \log f(Y_i|X_i;\theta_*)}{\partial \theta \partial \theta'} \right)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial \log f(Y_i|X_i;\theta_0)}{\partial \theta},$$

where we have already substituted in the expressions for the first and second derivatives in terms of the log density. Under certain technical conditions we have

$$-n^{-1}\sum_{i=1}^{n} \frac{\partial^2 \log f(Y_i|X_i;\theta_*)}{\partial \theta \partial \theta'} \underset{p}{\to} -\mathbb{E}\left(\frac{\partial^2 \log f(Y_i|X_i;\theta_0)}{\partial \theta \partial \theta'} \right) = I$$

while, also,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial \log f(Y_i|X_i;\theta_0)}{\partial \theta} \underset{d}{\to} N(0, I).$$

It then follows from Slutsky's theorem that

$$\sqrt{n}(\hat{\theta} - \theta_0) \underset{d}{\to} N(0, I^{-1})$$

as $n \to \infty$. The simplification of the asymptotic variance is a consequence of the information equality.

The asymptotic variance can be estimated using either of two methods. The first is as the outer product of the score,

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\partial \log f(Y_i|X_i;\hat{\theta})}{\partial \theta} \frac{\partial \log f(Y_i|X_i;\hat{\theta})}{\partial \theta'}.$$

The second is through the information equality, as

$$-\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \log f(Y_i|X_i;\hat{\theta})}{\partial\theta\partial\theta'}.$$

The second estimator is often easier to obtain as the second derivative has usually already been computed as part of the numerical optimisation of the log-likelihood function in calculating $\hat{\theta}$.

With the limit distribution available we can perform inference on θ_0 in the usual way.

3.2. Regularity conditions

Consistency and asymptotic normality as established above implicitly uses the following conditions.

C1, The function $\mathbb{E}(\log f(Y|X;\theta))$ has a unique global maximiser on Θ, equal to θ_0, and the global maximum is achieved on the interior of this parameter space Θ.

C2. The information matrix I has full rank.

C3, The function $f(y|x;\theta)$ is twice continuously differentiable in θ on a known compact set Θ and

$$\mathbb{E}\left(\sup_{\theta\in\Theta}\log f(Y|X;\theta)\right) < +\infty, \qquad \mathbb{E}\left(\sup_{\theta\in\Theta}\log\frac{\partial^2 f(Y|X;\theta)}{\partial\theta\partial\theta'}\right) < +\infty.$$

Further, its support does not depend on θ.

Here, Assumption C1 ensures global identification of θ_0. Assumption C2 is a local identification condition. This is not needed for consistency but is used in deriving the limit distribution. When the information matrix is singular the rate of convergence of the maximum-likelihood estimator will be slower

than $n^{-1/2}$ and its limit distribution will generally not be normal. Assumption C3 finally collects regularity conditions that allow to take derivatives and the use of a uniform law of large numbers.

4. Optimality

To explain where the notion of optimality of maximum likelihood comes from it is useful to consider a generic situation where an estimator $\hat{\theta}$ is unbiased for θ. That is

$$\mathbb{E}_\theta(\hat{\theta} - \theta) = 0.$$

Introducing bias only complicates the notation but does not alter the main message as long as bias vanishes at a rate faster than $n^{-1/2}$, which it does quite generally.

We consider the simplest case where θ is a scalar and we abstract from conditioning variables (regressors). The zero-bias condition can be written as

$$\int \cdots \int (\hat{\theta} - \theta) \left(\prod_{i=1}^{n} f(y_i; \theta) \right) dy_1 \cdots dy_n = 0,$$

where the structure of the distribution is a consequence of random sampling. Differentiating with respect to θ yields

$$\int \cdots \int (\hat{\theta} - \theta) \frac{\partial \prod_{i=1}^{n} f(y_i; \theta)}{\partial \theta} dy_1 \cdots dy_n = \int \cdots \int \prod_{i=1}^{n} f(y_i; \theta) dy_1 \cdots dy_n.$$

The term on the right-hand side is simply equal to one as it is the integral of an n-variate density. The term on the right-hand side is equal to

$$\int \cdots \int (\hat{\theta} - \theta) \left(\sum_{i=1}^{n} \frac{\partial \log f(y_i; \theta)}{\partial \theta} \right) \left(\prod_{j} f(y_j; \theta) \right) dy_1 \cdots dy_n$$

This follows from the observation that

$$\begin{aligned}\sum_{i=1}^{n} \frac{\partial \log f(y_i;\theta)}{\partial \theta} &= \sum_{i=1}^{n} \frac{1}{f(y_i;\theta)} \frac{\partial f(y_i;\theta)}{\partial \theta} \\ &= \sum_{i=1}^{n} \left(\frac{\prod_{j \neq i} f(y_j;\theta)}{\prod_j \quad f(y_j;\theta)} \right) \frac{\partial f(y_i;\theta)}{\partial \theta} \\ &= \frac{1}{\prod_j f(y_j;\theta)} \sum_{i=1}^{n} \prod_{j \neq i} f(y_j;\theta) \frac{\partial f(y_i;\theta)}{\partial \theta} \\ &= \frac{1}{\prod_j f(y_j;\theta)} \frac{\partial \prod_i f(y_i;\theta)}{\partial \theta},\end{aligned}$$

so that we can write

$$\frac{\partial \prod_i f(y_i;\theta)}{\partial \theta} = \left(\sum_{i=1}^{n} \frac{\partial \log f(y_i;\theta)}{\partial \theta} \right) \left(\prod_j f(y_j;\theta) \right).$$

The derivative of the zero-bias condition then can be written as

$$\mathbb{E}_\theta \left((\hat{\theta} - \theta) \left(\sum_{i=1}^{n} \frac{\partial \log f(y_i;\theta)}{\partial \theta} \right) \right) = 1.$$

Because

$$\mathbb{E}_\theta \left(\sum_{i=1}^{n} \frac{\partial \log f(y_i;\theta)}{\partial \theta} \right) = \sum_{i=1}^{n} \mathbb{E}_\theta \left(\frac{\partial \log f(y_i;\theta)}{\partial \theta} \right) = 0$$

the left-hand side term is equal to the covariance between the estimator $\hat{\theta}$ and the sum of the scores $\sum_{i=1}^{n} \partial \log f(y_i;\theta)/\partial\theta$, and so we have that

$$\operatorname{cov}_\theta \left(\hat{\theta}, \sum_{i=1}^{n} \frac{\partial f(y_i;\theta)}{\partial \theta} \right) = 1.$$

By the Cauchy-Shwarz inequality,

$$\mathrm{cov}_\theta\left(\hat{\theta}, \sum_{i=1}^{n} \frac{\partial f(y_i;\theta)}{\partial\theta}\right)^2 \leq \mathrm{var}_\theta(\hat{\theta}) \times \mathrm{var}_\theta\left(\sum_{i=1}^{n} \frac{\partial f(y_i;\theta)}{\partial\theta}\right) = \mathrm{var}_\theta(\hat{\theta}) \times (n\, I)$$

Hence, we must have that

$$\frac{I^{-1}}{n} \leq \mathrm{var}_\theta(\hat{\theta}).$$

This lower bound on the variance of any unbiased estimator is called the Cramer-Rao bound. Any estimator that attains it must thus satisfy the Cauchy-Shwarz inequality above with a strict equality. It must thus be that $\hat{\theta} - \theta$ is a linear transformation of the $\sum_{i=1}^{n} \partial \log f(Y_i|X_i;\theta_0)/\partial\theta$; that is, for constants A and B

$$\hat{\theta} - \theta = A + B \sum_{i=1}^{n} \frac{\partial \log f(y_i;\theta)}{\partial\theta}.$$

By unbiasedness of the estimator and the zero-mean property of the score we must have that $A = 0$. Because the variance of the right-hand side must equal I^{-1}/n and the variance of the right-hand side is nB^2I we must further have that $B = I^{-1}/n$. Therefore, for an estimator to satisfy the bound for a given n it must be

$$\hat{\theta} - \theta = \frac{1}{n} \sum_{i=1}^{n} I^{-1} \frac{\partial \log f(y_i;\theta)}{\partial\theta}.$$

In many cases such an estimator does not exist. However, the maximum likelihood estimator satisfies

$$\hat{\theta} - \theta = \frac{1}{n} \sum_{i=1}^{n} I^{-1} \frac{\partial \log f(y_i;\theta)}{\partial\theta} + o_p(n^{-1/2}),$$

quite generally, and so attains the Cramer-Rao bound asymptotically. Here, the term $o_p(n^{-1/2})$ indicates a random variable that converges to zero in probability at a rate faster than $n^{-1/2}$

You can verify that the maximum-likelihood estimator attains the bound for fixed n in the examples from Sections 1.1 and 1.2. It also achieves the bound for fixed n for the regression slopes in Section 1.4, but not for the variance, nor for the coefficients in the probit or Tobit model. In all cases, though, it does achieve the bound in large samples.