

# 10 Probit. Maximum Likelihood.

## 10.1 Binary choice

Suppose we don't have a continuous dependent variable, rather it is binary:  $y_i = \{0, 1\}$ . We could still use OLS here, let's check out the assumptions:

(OLS0)  $(y_i, x_i)$  is an i.i.d. sequence

✓ Binary  $y_i$  doesn't break this, we can still have an i.i.d. sequence

(OLS1)  $E(x_i x_i')$  is finite non-singular

✓ Binary  $y_i$  doesn't affect this

(OLS2)  $E(y_i | x_i) = x_i' \beta$

?  $E(y_i | x_i) = 1 \times P(y_i = 1 | x_i) + 0 \times P(y_i = 0 | x_i) = P(y_i = 1 | x_i) \stackrel{?}{=} x_i' \beta$

Hence, for OLS2 to hold we need use the linear probability model.

(OLS3)  $\text{Var}(y_i | x_i) = \sigma^2$

×  $\text{Var}(y_i | x_i) = E(y_i^2 | x_i) - E(y_i | x_i)^2 = E(y_i | x_i) - E(y_i | x_i)^2 = x_i' \beta (1 - x_i' \beta)$

using  $y^2 = y$ . Hence OLS3 cannot hold, we do have heteroskedasticity.

(OLS4)  $E \varepsilon_i^4 < \infty$ ,  $E \|x_i\|^4 < \infty$

✓ May still hold

We can fix the heteroskedasticity with GLS or White standard errors, but the linear probability model is more of a problem. This model does not restrict predicted probabilities to be between 0 and 1, and the use of any other model will violate OLS2 meaning OLS will not be consistent.

The standard alternative is to use a function of the form

$$P(y_i = 1 | x_i) = F(x_i' \beta)$$

where  $F(\cdot)$  is a known CDF, typically assumed to be symmetric about zero, so that  $F(u) = 1 - F(-u)$ . The standard choices for  $F$  are

- Logistic:  $F(u) = \frac{e^u}{1+e^u}$ , known as the **logit** model
- Normal:  $F(u) = \Phi(u)$ , known as the **probit** model

This is identical to the latent variable model

$$\begin{aligned} y_i^* &= x_i' \beta + \varepsilon_i \\ \varepsilon_i &\sim F(\cdot) \\ y_i &= \begin{cases} 1 & \text{if } y_i^* > 0 \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

Since then

$$\begin{aligned} P(y_i = 1 | x_i) &= P(y_i^* > 0 | x_i) \\ &= P(x_i' \beta + \varepsilon_i > 0 | x_i) \\ &= P(\varepsilon_i > -x_i' \beta | x_i) \\ &= 1 - F(-x_i' \beta) \\ &= F(x_i' \beta) \end{aligned}$$

## 10.2 Maximum likelihood estimation

The probit model is typically estimated by the method of maximum likelihood (ML). Consider the typical setup:

$$\begin{aligned} z_1, \dots, z_n &\stackrel{i.i.d.}{\sim} f(\cdot|\theta) \quad \rightarrow \quad L(\theta) = \prod_{i=1}^n f(z_i|\theta) \\ \log L(\theta) = \ell(\theta) &= \sum_{i=1}^n \log f(z_i|\theta) \\ \hat{\theta}_{ML} &= \arg \max_{\theta} \ell(\theta) \end{aligned}$$

This is known as a *parametric model*, it requires the specification of the distribution of the data up to an unknown parameter  $\theta$ .

A key property is that the expected log-likelihood is maximised at the true value of the parameter vector  $\theta_0$ . Set  $Z = (z_1, \dots, z_n)$ .

**Theorem 10.2.1.**  $\theta_0 = \arg \max_{\theta} \mathbb{E}(\log L(\theta)|Z)$

The proof is presented in Lecture 11 using KL divergence. This motivates estimating  $\theta$  by finding the value which maximises log-likelihood.

**Example (OLS using MLE).**

$$\begin{aligned} f(Y_1, \dots, Y_n|X, \beta, \sigma^2) &: L = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(Y_i - X_i'\beta)^2}{2\sigma^2}} \\ \Rightarrow \ell = \log L &= \sum_{i=1}^n \log \left( \frac{1}{\sqrt{2\pi\sigma^2}} \right) - \frac{(Y_i - X_i'\beta)^2}{2\sigma^2} \\ &= n \log \left( \frac{1}{\sqrt{2\pi\sigma^2}} \right) - \frac{\sum_{i=1}^n (Y_i - X_i'\beta)^2}{2\sigma^2} \\ &= \frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{\sum_{i=1}^n (Y_i - X_i'\beta)^2}{2\sigma^2} \end{aligned}$$

Hence, the FOCs are:

$$\frac{\partial \ell}{\partial \beta} = -\frac{\sum_{i=1}^n (-X_i)(Y_i - X_i'\beta)}{\sigma^2} = 0 \quad (10.1)$$

$$\frac{\partial \ell}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{\sum_{i=1}^n (Y_i - X_i'\beta)^2}{2\sigma^4} = 0 \quad (10.2)$$

$$\begin{aligned} (10.1) \quad \Rightarrow \sum_{i=1}^n X_i Y_i - X_i X_i' \hat{\beta}_{ML} &= 0 & (10.2) \quad \Rightarrow n\sigma^2 &= \sum_{i=1}^n (Y_i - X_i' \hat{\beta}_{ML})^2 \\ \Rightarrow X'Y - X'X \hat{\beta}_{ML} &= 0 & \Rightarrow \sigma^2 &= \frac{1}{n} \sum_{i=1}^n (Y_i - X_i' \hat{\beta}_{ML})^2 \\ \Rightarrow \hat{\beta}_{ML} &= (X'X)^{-1} X'Y = \hat{\beta}_{OLS} \end{aligned}$$

Thus,  $\hat{\beta}_{OLS}$  is actually the MLE for  $\beta$ , so it has the desirable properties discussed in Lecture 11. However, the ML estimator for the variance is biased due to not correcting for the loss in degrees of freedom from estimating  $\hat{\beta}_{ML}$ .

Consider the problem of estimating  $\theta$  if you have a vector of data  $Z$  with the joint density of its elements given by  $f(z|\theta)$ .

### Definition 10.2.1: Score

The score of the likelihood function is the vector of partial derivatives with respect to the parameters.

$$\frac{\partial}{\partial \theta} \log f(Z|\theta)$$

**Theorem 10.2.2.** If  $\log f(Z|\theta)$  is second differentiable and the support of  $Z$  doesn't depend on  $\theta$  then the score has mean zero:

$$\mathbb{E} \left[ \frac{\partial}{\partial \theta} \log f(Z|\theta) \right] = 0$$

**Proof.**

$$\begin{aligned} \mathbb{E} \left[ \frac{\partial}{\partial \theta} \log f(Z|\theta) \right] &= \int_{\mathbb{R}} \frac{\frac{\partial}{\partial \theta} f(z|\theta)}{f(z|\theta)} f(z|\theta) dz \\ &= \frac{\partial}{\partial \theta} \int_{\mathbb{R}} f(z|\theta) dz \\ &= \frac{\partial}{\partial \theta} 1 \\ &= 0 \end{aligned}$$

□

### Definition 10.2.2: Fisher information

The covariance matrix of the score is known as the Fisher information

$$I(\theta) = \text{Var} \left( \frac{\partial}{\partial \theta} \log f(Z|\theta) \right) = \mathbb{E} \left[ \left( \frac{\partial}{\partial \theta} \log f(Z|\theta) \right)^2 | \theta \right]$$

### Note:-

Because the likelihood of  $\theta$  given  $Z$  is always proportional to the probability  $f(Z|\theta)$ ; their logarithms necessarily differ by a constant that is independent of  $\theta$ , and the derivatives are necessarily equal. Thus one can substitute in  $\log L(\theta) = \ell(\theta)$  for  $\log f(Z|\theta)$  in the above definitions.

The Fisher information is a way of measuring the amount of information that an observable  $Z$  carries about the unknown parameter  $\theta$ . If  $f$  is sharply peaked with respect to changes in  $\theta$ , it is easy to indicate the "correct" value of  $\theta$  from the data, or equivalently, that the data  $Z$  provides a lot of information about the parameter  $\theta$ . If  $f$  is flat and spread-out, then it would take many samples of  $Z$  to estimate the true value of  $\theta$ . Note that  $I(\theta) \geq 0$ . Near the ML estimate, low Fisher information suggests the maximum appears flat, that is, there are many nearby values with similar log-likelihood. Conversely, high Fisher information indicates the maximum is sharp.

**Claim 10.2.1.** If we have  $n$  i.i.d. distributions (from  $n$  samples) then the Fisher information will be  $n$  times the Fisher information of a single sample from the common distribution.

$$I_n(\theta) = nI_1(\theta)$$

**Lemma 10.2.1 (Information equality).** The variance of the score is equal to the negative expected value of the Hessian matrix of the log-likelihood.

$$I(\theta) = \text{Var} \left( \frac{\partial}{\partial \theta} \log f(Z|\theta) \right) = -\mathbb{E} \left( \frac{\partial^2}{\partial \theta \partial \theta'} \log f(Z|\theta) \right)$$

**Proof.** Let  $\mathbf{Z}$  be an  $m$ -component column vector of random variables, not necessarily i.i.d. To ease notation, we denote their joint density as  $f(\mathbf{Z}|\theta) \equiv f$ . Also note all expectations are conditional on  $\theta$ , and integrals are multiple integrals over  $z_1, \dots, z_n$ .

$$\begin{aligned} \mathbb{E} \frac{\partial^2 \log f}{\partial \theta \partial \theta'} &= \mathbb{E} \left[ \frac{\partial}{\partial \theta} \left( \frac{\partial \log f}{\partial \theta'} \right) \right] \\ &= \mathbb{E} \left[ \frac{\partial}{\partial \theta} \left( \frac{1}{f} \frac{\partial f}{\partial \theta'} \right) \right] \\ &= \mathbb{E} \left[ -\frac{1}{f^2} \frac{\partial f}{\partial \theta} \frac{\partial f}{\partial \theta'} + \frac{1}{f} \frac{\partial^2 f}{\partial \theta \partial \theta'} \right] \\ &= -\mathbb{E} \left[ \left( \frac{1}{f} \frac{\partial f}{\partial \theta} \right) \left( \frac{1}{f} \frac{\partial f}{\partial \theta'} \right) \right] + \mathbb{E} \left[ \frac{1}{f} \frac{\partial^2 f}{\partial \theta \partial \theta'} \right] \end{aligned}$$

To obtain the information equality, we need to show the second term is zero.

$$\begin{aligned} \mathbb{E} \left[ \frac{1}{f} \frac{\partial^2 f}{\partial \theta \partial \theta'} \right] &= \int_{\mathbb{R}} f \frac{1}{f} \frac{\partial^2 f}{\partial \theta \partial \theta'} d\mathbf{Z} \\ &= \int_{\mathbb{R}} \frac{\partial^2 f}{\partial \theta \partial \theta'} d\mathbf{Z} \\ &= \int_{\mathbb{R}} \frac{\partial}{\partial \theta} \left( \frac{\partial f}{\partial \theta'} \right) d\mathbf{Z} \\ &= \frac{\partial}{\partial \theta} \int_{\mathbb{R}} \frac{\partial f}{\partial \theta'} d\mathbf{Z} \quad (\text{we can interchange these because we are economists}) \\ &= \frac{\partial}{\partial \theta} \frac{\partial}{\partial \theta'} \int_{\mathbb{R}} f d\mathbf{Z} \quad (\text{what even is a regularity condition}) \\ &= \frac{\partial}{\partial \theta} \frac{\partial}{\partial \theta'} 1 \\ &= 0 \end{aligned}$$

□

**Note:-**

All we are assuming here is that we can interchange the order of differentiation and integration; a set of sufficient conditions for this are:

1. The function  $\frac{\partial}{\partial \theta} f(\mathbf{Z}|\theta)$  is continuous in  $\mathbf{Z}$  and in  $\theta \in \Theta$  where  $\Theta$  is an open set.
2. The integral  $\int f(\mathbf{Z}|\theta) d\mathbf{Z}$  exists.
3.  $\int \left| \frac{\partial}{\partial \theta} f(\mathbf{Z}|\theta) \right| d\mathbf{Z} < M < \infty$  for all  $\theta \in \Theta$

**Misspecification and the information equality**

Suppose our random variables have joint density  $f$  as before, but we specify that they have joint density  $g$  instead. As before

$$\mathbb{E}_f \frac{\partial^2 \log g}{\partial \theta \partial \theta'} = -\mathbb{E}_f \left[ \left( \frac{1}{g} \frac{\partial g}{\partial \theta} \right) \left( \frac{1}{g} \frac{\partial g}{\partial \theta'} \right) \right] + \mathbb{E}_f \left[ \frac{1}{g} \frac{\partial^2 g}{\partial \theta \partial \theta'} \right]$$

where the  $f$  subscript denotes the fact that we are taking the expectation with respect to the true distribution. Previously we made progress because the integrand contained  $f \frac{1}{f} = 1$ , however we now have  $f \frac{1}{g}$  which doesn't simplify. In general, under misspecification

$$\mathbb{E}_f \left[ \frac{1}{g} \frac{\partial^2 g}{\partial \theta \partial \theta'} \right] \neq 0$$

and the IE doesn't hold. Note: this does not exclude the possibility that this expected value is after all zero and the IE holds, it just generally isn't.

**Theorem 10.2.3 (Cramer-Rao lower bound).** If  $\tilde{\theta}$  is an unbiased estimator of  $\theta$ , then we have the following bound on its variance

$$\text{Var}(\tilde{\theta}|\mathbf{Z}) \geq [I(\theta)]^{-1}$$

These are both matrices, meaning this inequality tells us the difference between the left and right hand sides is positive semi-definite.

This result is similar to the Gauss-Markov theorem which established a lower bound for unbiased estimators in homoskedastic linear regression.

**Example (Information bound for normal regression).** We will apply the CRLB conditionally on  $\mathbf{X}$ . Define the expected Hessian

$$\mathbb{E}(H) = \begin{bmatrix} \mathbb{E} \left( \frac{\partial^2 \ell}{\partial \beta \partial \beta'} | X \right) & \mathbb{E} \left( \frac{\partial^2 \ell}{\partial \beta \partial \sigma^2} | X \right) \\ \mathbb{E} \left( \frac{\partial^2 \ell}{\partial \sigma^2 \partial \beta'} | X \right) & \mathbb{E} \left( \frac{\partial^2 \ell}{\partial \sigma^2 \partial \sigma^2} | X \right) \end{bmatrix}$$

Recall the log likelihood

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

Thus we have second derivatives

$$\begin{aligned} \frac{\partial^2 \ell}{\partial \beta \partial \beta'} &= \frac{\partial}{\partial \beta'} \frac{\sum_{i=1}^n X_i (Y_i - X_i' \beta)}{\sigma^2} = -\frac{1}{\sigma^2} \sum_{i=1}^n X_i X_i' = -\frac{1}{\sigma^2} X' X \\ \frac{\partial^2 \ell}{\partial \beta \partial \sigma^2} &= \frac{\partial}{\partial \sigma^2} \frac{\sum_{i=1}^n X_i (Y_i - X_i' \beta)}{\sigma^2} = -\frac{\sum_{i=1}^n X_i (Y_i - X_i' \beta)}{\sigma^4} = -\frac{1}{\sigma^4} X' (Y - X \beta) \\ \frac{\partial^2 \ell}{\partial \sigma^2 \partial \sigma^2} &= \frac{n}{2} \frac{1}{\sigma^4} - \frac{\sum_{i=1}^n (Y_i - X_i' \beta)^2}{\sigma^6} = \frac{n}{2} \frac{1}{\sigma^4} - \frac{1}{\sigma^6} (Y - X \beta)' (Y - X \beta) \end{aligned}$$

$$\begin{aligned} \Rightarrow \mathbb{E}(H) &= \begin{bmatrix} \mathbb{E} \left[ \frac{1}{\sigma^2} X' X | X \right] & \mathbb{E} \left[ -\frac{1}{\sigma^4} X' (Y - X \beta) | X \right] \\ \mathbb{E} \left[ -\frac{1}{\sigma^4} X' (Y - X \beta) | X \right] & \mathbb{E} \left[ \frac{n}{2} \frac{1}{\sigma^4} - \frac{1}{\sigma^6} (Y - X \beta)' (Y - X \beta) | X \right] \end{bmatrix} \\ &= \begin{bmatrix} \frac{1}{\sigma^2} X' X & 0 \\ 0 & \frac{n}{2} \frac{1}{\sigma^4} - \frac{n \sigma^2}{\sigma^6} \end{bmatrix} \\ &= \begin{bmatrix} \frac{1}{\sigma^2} X' X & 0 \\ 0 & -\frac{n}{2} \frac{1}{\sigma^4} \end{bmatrix} \end{aligned}$$

The block diagonal matrix can be inverted to find the lower bound on asymptotic conditional variance

$$[I(\theta)]^{-1} = \begin{bmatrix} \sigma^2(X'X)^{-1} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{bmatrix}$$

The variance of  $\hat{\beta}_{OLS} = \hat{\beta}_{ML}$  meets the CRLB. Thus we have the following theorem

**Theorem 10.2.4.** In the normal regression, OLS is the Best Unbiased Estimator (BUE).

This result should be distinguished from the Gauss-Markov Theorem that  $\hat{\beta}_{OLS}$  is minimum variance among those estimators that are unbiased and linear in  $y$ . Theorem 10.2.4 says that  $\hat{\beta}_{OLS}$  is minimum variance in a larger class of estimators that includes non-linear unbiased estimators. This stronger statement is obtained under the normality assumption which is not assumed in the Gauss-Markov Theorem. Put differently, the Gauss-Markov Theorem does not exclude the possibility of some non-linear estimator beating OLS, but this possibility is ruled out by the normality assumption.

As we have already seen, the ML estimator of  $\sigma^2$  is biased, so the CRLB does not apply. But the OLS estimator  $\hat{\sigma}^2$  of  $\sigma^2$  is unbiased, does it achieve the bound? We know  $\frac{(n-k)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-k)$ , and  $Var(\chi^2(p)) = 2p$ . Thus

$$\begin{aligned} Var\left(\frac{(n-k)\hat{\sigma}^2}{\sigma^2}\right) &= 2(n-k) \\ \Rightarrow \frac{(n-k)^2}{\sigma^4} Var(\hat{\sigma}^2) &= 2(n-k) \\ \Rightarrow Var(\hat{\sigma}^2) &= \frac{2\sigma^4}{n-k} \end{aligned}$$

Therefore  $\hat{\sigma}^2$  does not attain the CRLB  $2\sigma^4/n$ . However it can be shown that an unbiased estimator with variance lower than  $\hat{\sigma}^2$  does not exist.