Recitation Notes 2

Konrad Menzel

September 23, 2006

1 Maximum Likelihood Estimation

We assume that we can narrow down the model of the conditional distribution of Y conditional on X to a family of probability functions (probability mass functions, density functions, or a hybrid of the two for mixed discrete/continuous data) $f(Y|X,\theta)$ which is known up to a finite-dimensional parameter vector θ, so that for the "true" value θ_0 of θ

$$Y_i|X_i \overset{iid}{\sim} F(Y_i|X_i,\theta_0)$$

The goal will now be to recover the "truth" from the data in order to obtain a complete characterization of the data generating process.

Example: normal linear regression - If we observe a sample Y_i, X_i and posit that

$$Y_i \overset{iid}{\sim} \mathcal{N}(X_i\beta_0, \sigma_0^2)$$

we might try to estimate the unknown parameter $\theta_0 := (\beta_0, \sigma_0)$. As we know, OLS will give us unbiased estimates for that.

We define the conditional likelihood function of the parameter θ as a function of the data (assuming observations are independently and identically distributed)

$$\ell(\theta;\mathbf{Y},\mathbf{X}) := \prod_{i=1}^{N} f(Y_i|X_i,\theta)$$

It turns out that it is easier to formulate the problem in terms of the conditional log-likelihood (for technical reasons that I will not get into - you'll get an idea of that in the proof of the expected log-likelihood inequality and some of the examples below)

$$L(\theta;\mathbf{Y},\mathbf{X}) = \log \ell(\theta;\mathbf{Y},\mathbf{X}) = \sum_{i=1}^{N} \log(f(Y_i|X_i,\theta)))$$

The maximum-likelihood estimator is then defined as

$$\hat{\theta}_{ML} := \arg\max_{\theta\in\Theta} L(\theta;\mathbf{Y},\mathbf{X})$$

i.e. the parameter vector that makes the data "most likely" - all estimation principles in some way "fit the data" by minimizing some notion of "distance" between the observed data and a "model" - e.g. least squares, least absolute deviations - which results in a "projection" in a very wide sense.[1] The following result is the main justification for the Maximum-Likelihood principle:

Theorem 1 ***(Expected Log-Likelihood Inequality)*** *For the conditional log-likelihood function for theta, $L(\theta; Y, X)$, under regularity conditions*

$$\mathbb{E}[L(\theta; \mathbf{Y}, \mathbf{X})|\mathbf{X}] \leq \mathbb{E}[L(\theta_0; \mathbf{Y}, \mathbf{X})|\mathbf{X}]$$

Proof: Remember that since $\log(\cdot)$ is concave, for any random variable X,

$$\mathbb{E}[\log(X)] \leq \log(\mathbb{E}X)$$

Therefore

$$\begin{aligned} \mathbb{E}[L(\theta; \mathbf{Y}, \mathbf{X})|\mathbf{X}] - \mathbb{E}[L(\theta_0; \mathbf{Y}, \mathbf{X})|\mathbf{X}] &\overset{iid}{=} \mathbb{E}\left[\log f(y|X,\theta) - \log f(Y|X,\theta_0)|\, X\right] \\ &= \mathbb{E}\left[\log\left(\frac{f(Y|X,\theta)}{f(Y|X,\theta_0)}\right)\middle| X\right] \\ &\leq \log\left(\mathbb{E}\left[\frac{f(Y|X,\theta)}{f(Y|X,\theta_0)}\middle| X\right]\right) \\ &= \log(1) = 0 \end{aligned}$$

since

$$\mathbb{E}\left[\frac{f(Y|X,\theta)}{f(Y|X,\theta_0)}\middle| X\right] = \int \frac{f(y|X,\theta)}{f(y|X,\theta_0)} f(y|X,\theta_0)dy = \int f(y|X,\theta)dy = 1$$

because $f(y|X,\theta)$ is a probability density for all values of θ. □

Since the inequality is weak, there is no guarantee that the maximizer of the log-likelihood will always be unique - if it is not, we say that θ is *not (point-)identified.*

ML estimators have a number of desirable properties: if we assumed the "right" model, they are consistent (though typically not unbiased), and (asymptotically) efficient. On the other hand, if the "truth" lies outside the range of the models we allowed for, the ML estimator will typically not be of great use (in particular because the parameter can in most cases, the interpretation of the parameter depends on the underlying model), but we'll also see some exceptions below.

2 MLE - Examples

In general, it will not be possible to write out the maximum-likelihood estimator for a problem as a closed-form expression, but in practise, we'd need to use numerical optimization techniques (e.g. the Newton algorithm or grid search) to find the maximum of the log-likelihood. However, there are a few simple examples that have an analytic solution, so that we can directly see what maximum likelihood actually does.

[1] The notion of "distance" maximum likelihood estimation is based on is the *Kullback-Leibler divergence*, which for two probability functions $p(x)$ and $q(x)$ is defined as

$$D_{KL}(p|q) = \int \log\left(\frac{q(x)}{p(x)}\right) p(x)dx$$

Note that this is not a distance metric in the strict sense since $D_{KL}(\cdot,\cdot)$ is not symmetric in its two arguments, but $p(x)$ assumes the role of the "reference" distribution.

2.1 Bernoulli Distribution

Suppose that there is a constant, but unknown, probability p that a certain event occurs, and we observe independent realizations of an indicator

$$Y_i = \begin{cases} 1 & \text{if } A \text{ occurs} \\ 0 & \text{otherwise} \end{cases}$$

This means that Y_i follows a Bernoulli distribution with probability p, in symbols

$$Y_i \overset{iid}{\sim} \mathcal{B}(p)$$

Of course it is straightforward to estimate the only parameter of the model, p by the sample mean,

$$\hat{p}_{LS} = \frac{1}{N}\sum_{i=1}^{N} Y_i$$

but let's now see what the maximum-likelihood estimator for this problem looks like. The log-likelihood for this problem is

$$L(p;\mathbf{Y}) = \frac{1}{N}\log\left(\prod_{i=1}^{N} p^{Y_i}(1-p)^{1-Y_i}\right) = \frac{1}{N}\sum_{i=1}^{N}\Big[\,Y_i\log(p) + (1-Y_i)\log(1-p)\Big]$$

Maximizing this with respect to p gives us the first-order conditions

$$0 = \frac{\partial}{\partial p}L(p;\mathbf{Y}) = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{Y_i}{p} - \frac{1-Y_i}{1-p}\right] = \frac{1}{N}\sum_{i=1}^{N}\frac{Y_i(1-p)-(1-Y_i)p}{p(1-p)} = \frac{1}{N}\sum_{i=1}^{N}\frac{Y_i - p}{p(1-p)}$$

Solving for p gives us the ML estimator

$$\hat{p}_{ML} = \frac{1}{N}\sum_{i=1}^{N} Y_i = \hat{p}_{LS}$$

This means that the sample average of Y_i is not only an unbiased estimator for p, but that it's also maximum likelihood.

2.2 Normal Linear Regression

Now we want to estimate the parameters in a linear regression

$$Y_i = X_i\beta + \varepsilon_i$$

and we know that

$$\varepsilon_i \overset{iid}{\sim} \mathcal{N}(0,\sigma^2)$$

for some (unknown) $\sigma^2 > 0$. Since the normal density is

$$f_\varepsilon(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{\varepsilon^2}{2\sigma^2}\right)$$

can compute the log-likelihood for this regression

$$
\begin{aligned}
L(\beta, \sigma^2; Y_i, X_i) &= \frac{1}{N}\sum_{i=1}^{N} \log f_\varepsilon(Y_i - X_i\beta) \\
&= -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{N}\sum_{i=1}^{N}\frac{(Y_i - X_i\beta)^2}{2\sigma^2}
\end{aligned}
$$

The first-order conditions for a maximum are

$$
\begin{aligned}
0 = \frac{\partial}{\partial\beta}L(\beta, \sigma^2; X, Y) &= \frac{1}{N}\sum_{i=1}^{N} X_i' \frac{(Y_i - X_i\beta)}{2\sigma^2} \\
0 = \frac{\partial}{\partial\sigma^2}L(\beta, \sigma^2; X, Y) &= -\frac{1}{2\sigma^2} + \frac{1}{N}\sum_{i=1}^{N}\frac{(Y_i - X_i\beta)^2}{2\sigma^4}
\end{aligned}
$$

solving the first equation for β, get

$$
\hat{\beta}_{ML} = \left(\frac{1}{N}\sum_{i=1}^{N} X_i'X_i\right)^{-1}\frac{1}{N}\sum_{i=1}^{N} X_i'y_i = (\mathbf{X'X})^{\mathbf{-1}}\mathbf{X'y} = \hat{\beta}_{LS}
$$

and for the variance parameter, obtain from second equation

$$
\hat{\sigma}^2_{ML} = \frac{1}{N}\sum_{i=1}(Y_i - X_i\beta)^2 = \frac{N-k}{N}\hat{\sigma}^2_{LS}
$$

There are two things in the relationship between ML and OLS estimators that are worth pointing out:

- The maximum-likelihood estimator for the *slope coefficients* β is the same as OLS which, as we know from the Gauss-Markov theorem, is unbiased, consistent, and even efficient under much weaker assumptions. So in some cases, maximum-likelihood will give us consistent estimates even under some particular forms of misspecification.
- The estimators for the *variance* in ε_i differ only in the degrees-of-freedom correction $\frac{N}{N-k}$ which will converge to 1 as the sample size increases. Since the OLS estimator with the degrees-of-freedom correction is unbiased in finite samples, this means that ML estimators typically have finite-sample bias even though they are consistent.

3 The Probit Model

The probit model assumes that there exists an (unobserved) latent variable

$$
Y_i^* = X_i\beta + \varepsilon_i
$$

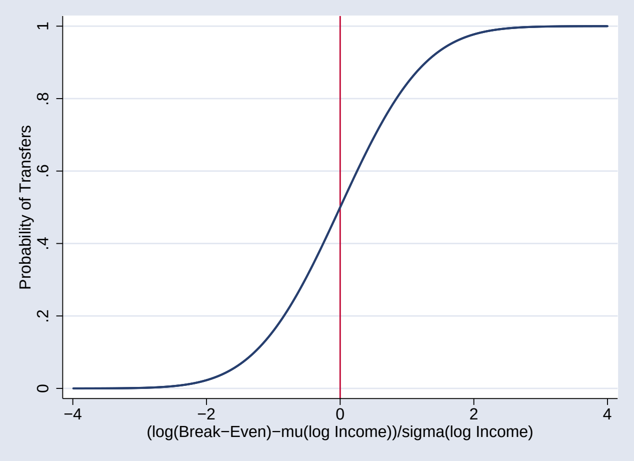
where

$$
\varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)
$$

and we can only observe

$$
Y_i = \mathbb{1}\{Y_i^* \geq 0\} = \mathbb{1}\{X_i\beta > -\varepsilon_i\}
$$

Figure 1: Illustration of the Probit model in Ashenfelter (1983) - for the control group, the likelihood of receiving transfers is the standard normal cdf evaluated at $(\log(G/t) - \mu_y)/\sigma$.



In the Ashenfelter (1983) paper, Y_i^* is equal to the log of the break-even level $\frac{G}{t}$ minus log earned income, and Y_i is a dummy for receiving transfers from the SIME/DIME program.
From these assumptions, it follows that

$$\mathbb{E}[Y_i|X_i] = \mathbb{P}(Y_i|X_i) = \Phi\left(\frac{X_i\beta}{\sigma}\right)$$

where $\Phi(z)$ is the standard normal cdf. Since any cdf is monotone, one way of backing out $X_i\beta/\sigma$ from the data (of course only up to sampling error) is to calculate the sample average of Y_i conditional on $X_i = x$ (which will be a conditional probability since Y_i is binary), and plot $\Phi^{-1}\left(\hat{\mathbb{E}}[Y_i|X_i = x]\right)$, which should be linear in x if the model assumptions are right. Taking the inverse of $\Phi(\cdot)$ is nothing but a nonlinear transformation of the y-axis, and if we want to do this analysis graphically, we can directly plot the conditional probabilities against x on specially prepared normal probability paper (see figure 2). Without sampling uncertainty (and if the linear model with normal errors is actually correct), the coefficient β/σ could be read off directly as the slope of the resulting straight line. If we didn't get a linear graph from this procedure despite a relatively large sample, we should start to worry whether either linearity or the distributional assumption (or, of course both) might have been wrong.
Back to how a Probit is actually estimated in practise, let's now write down the likelihood of the sample

$$\begin{aligned}\ell(\beta,\sigma;y_1,\dots,y_N,X_1,\dots,X_N) &:= \mathbb{P}\left\{ Y_1 = y_1, Y_2 = y_2,\dots \middle| X_1, X_2,\dots,\beta,\sigma\right\} \\ &\overset{iid}{=} \prod_{i=1}^{N} \mathbb{P}\{X_i\beta > \varepsilon_i\}^{y_i}\mathbb{P}\{X_i\beta \le \varepsilon_i\}^{1-y_i} \\ &= \prod_{i=1}^{N} \Phi\left(\frac{X_i\beta}{\sigma}\right)^{y_i}\left[1 - \Phi\left(\frac{X_i\beta}{\sigma}\right)\right]^{1-y_i}\end{aligned}$$

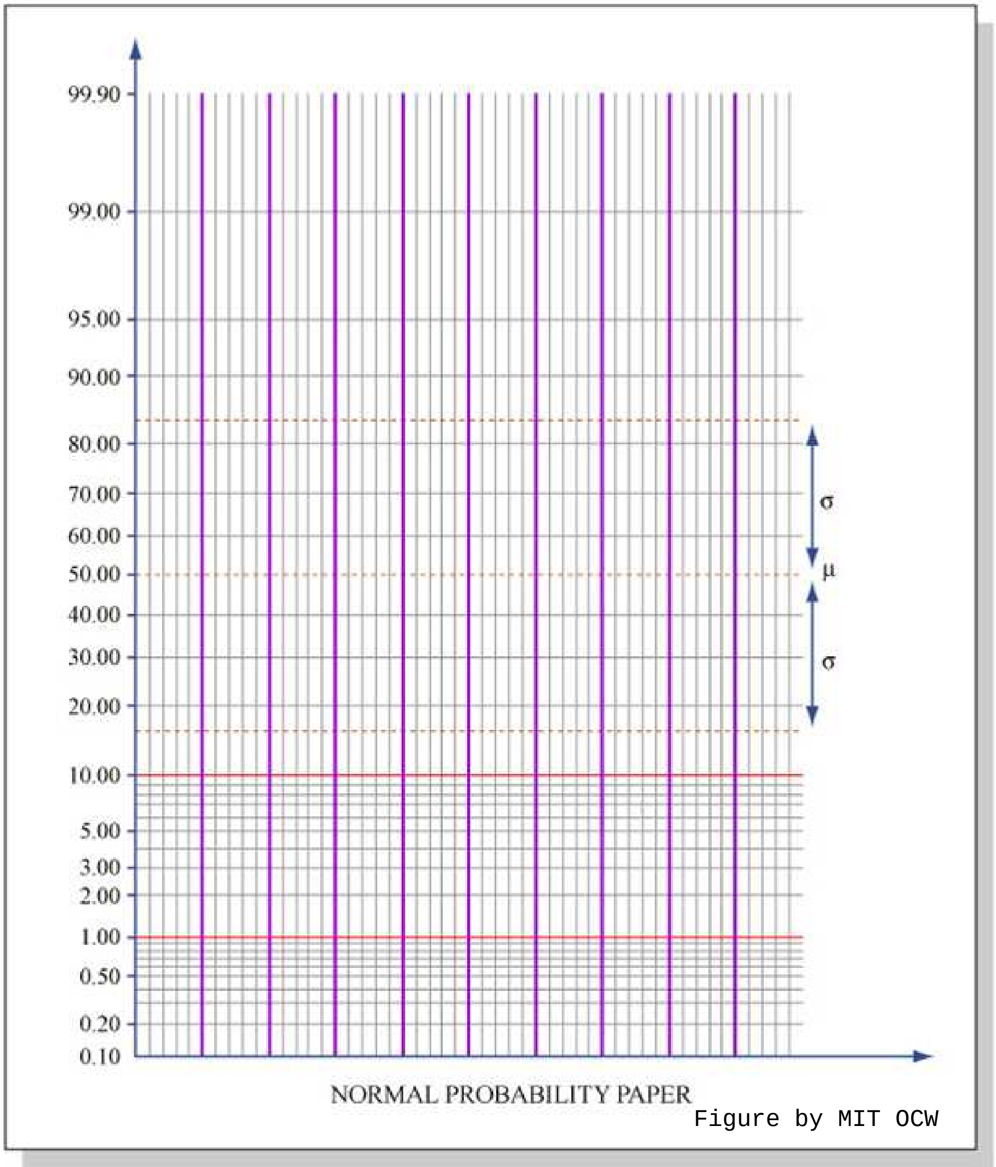


Figure 2: Normal Probability Paper

Notice that this looks exactly like the Bernoulli example above, except that we now assume that p is a particular function of X_i. Taking logs, we get the log-likelihood function

$$L(\beta,\sigma;\mathbf{y},\mathbf{X}) = \sum_{i=1}^{N}\left\{y_i\log\left[\Phi\left(\frac{X_i\beta}{\sigma}\right)\right]+(1-y_i)\log\left[1-\Phi\left(\frac{X_i\beta}{\sigma}\right)\right]\right\}$$

At this point you may have noticed that the parameters we want to estimate always appear in ratios $\frac{1}{\sigma}\beta$ which means that if we replace the "true" value of σ, σ_0, with, say, $2\sigma_0$, scaling up β by the same factor would leave the value of the log-likelihood unchanged. Therefore the maximum can't be unique unless we normalize σ to some positive value, typically $\sigma = 1$. We say that σ is *not identified* from the data. Such a normalization (often it's made only implicitly) is necessary for any type of discrete choice model (multiple choice, ordered choice, other distributions for the error term etc.).

If you think of $u_i = -X_i\beta+\varepsilon_i$ as the utility differential between working (and being paid) and not working, the fundamental reason for the identification problem is that, as you may recall from an intermediate micro class, a preference relation is only ordinal. I.e. any strictly monotone transformation of a given utility function represents exactly the same preferences, so that the scale of β has no empirical meaning.

The MLE solves the first-order conditions to the maximization problem (recall that we normalized σ to one)

$$\begin{aligned}0=\frac{\partial}{\partial\beta}L(\beta,1;\mathbf{y},\mathbf{X}) &= \sum_{i=1}^{N}\left\{\frac{y_i}{\Phi(X_i\beta)}-\frac{1-y_i}{1-\Phi(X_i\beta)}\right\}\varphi(X_i\beta)X_i\\ &= \sum_{i=1}^{N}\frac{y_i[1-\Phi(X_i\beta)]-(1-y_i)\Phi(X_i\beta)}{\Phi(X_i\beta)[1-\Phi(X_i\beta)]}\varphi(X_i\beta)X_i\\ &= \sum_{i=1}^{N}\frac{y_i-\Phi(X_i\beta)}{\Phi(X_i\beta)[1-\Phi(X_i\beta)]}\varphi(X_i\beta)X_i\end{aligned}$$

Now, instead of estimating the parameters via maximum likelihood, consider an alternative approach: you could always try to approximate the conditional mean $\mathbb{E}[Y_i|X_i]$ by just fitting a standard normal cdf $\Phi(X_i\gamma)$ via weighted nonlinear least squares, i.e.

$$\hat{\gamma}_{NLS} := \arg\min_{\gamma\in\mathbb{R}^k}\sum_{i=1}^{N}w_i\Big(\,y_i-\Phi(X_i\gamma)\Big)^2$$

where w_i are weights you may want to use for estimation. The first-order conditions for this problem are

$$0=\sum_{i=1}^{N}w_i(y_i-\Phi(X_i\gamma))\varphi(X_i\gamma)X_i$$

Now, plugging in

$$w_i := \frac{1}{\Phi(X_i\gamma)[1-\Phi(X_i\gamma)]}$$

these are exactly the same first-order conditions as for the maximum-likelihood estimator, so that for these particular weights, $\hat{\gamma}_{NLS} = \hat{\beta}_{ML}$. Since we don't know the value of γ beforehand, we could start with a preliminary estimate $\gamma^{(0)}$ and update the weights using the estimate from the last stage as we go along.

The main insight from this is that - regardless of whether the specification of the error distribution is correct - maximum-likelihood estimators for discrete-choice models approximate the conditional mean of y_i with a weighted least-squares fit of the cdf chosen by the researcher. Since we didn't use any specific properties of the normal distribution for this derivation - except that it has a density - an analogous result will hold for all binary outcome models based on a differentiable cdf $G(z)$.