Econometric Methods (Econometrics I)

Lecture 7: Maximum Likelihood Estimation

Prof. Dr. Kai Carstensen

Kiel University

Winter Term 2023/2024

Outline of this lecture

We will study the leading binary choice models as an example for maximum likelihood estimation.

1. Introduction to Binary Choice Models
2. Maximum Likelihood Estimator
3. Hypothesis Testing
4. ML Estimation of Binary Choice Models
5. Interpretation of Binary Choice Models
6. Example

Reference: Wooldridge, Chapters 13 (maximum likelihood) and 15 (binary choice).

Example: Labor Force Participation

- We want to explain the decision whether or not married women participate in the labor market.
- The underlying decision process will supposedly depend on
  (a) the wage offer by potential employers—probably based on expected marginal product
  (b) the reservation wage of the individual women
- Offered wage may depend on education (among others)
- Reservation wage may depend on age, children in the household, other sources of income, marginal tax rates etc.
- But can we run a linear regression of the zero-one coded decision variable on all these regressors?

1. Introduction to Binary Choice Models

Models to study economic choice

- Binary choice: decision between two alternatives, e.g., whether or not to produce, consume, take up work, go to university etc
- Multinomial choice: decision between several unordered alternatives, e.g., whether to buy a Samsung, Apple, HTC or Huawei cell phone
- Ordered Choice: decision between several ordered alternatives, e.g., whether to behave as a conservative, balanced, or aggressive investor, or whether to self-assess one's well-being in a health survey as "very poor", "poor", "fair", "good", "very good".

In general: some action i is chosen from a set of alternatives; the choice reveals the individual's preference order/utility function over these alternatives

Unifying feature: dependent variable is qualitative in nature

- As the outcome variable is not quantitative, a linear regression model might not be well suited (at least in terms of efficiency).
- Instead of modeling (continuous) quantities, one effectively models the **probability** of the events/choices in question.
- The predicted probabilities may depend on individual-specific and outcome-specific covariates.
- In this course, we limit ourselves to study the simplest model: binary choice.
- More models are studied in the course "Microeconometrics".

Binary choice: setup

Suppose you want to estimate the quantitative importance of factors that influence a yes-no decision.

The data you observe are zero-one coded, e.g., as 1=yes and 0=no.

How can we sensibly model that decision?

A linear model would specify

$$\mathrm{E}(y|\mathbf{x}) = \mathbf{x}\boldsymbol{\theta}.$$

But is this a good idea?

Think of the range of values taken by y and possibly taken by $\mathbf{x}\boldsymbol{\beta}$.

How else should we specify the conditional expectation? Take into account that y is a Bernoulli variable with conditional expectation

$$\mathrm{E}(y|\mathbf{x}) = 1 \times \mathrm{P}(y = 1|\mathbf{x}) + 0 \times \mathrm{P}(y = 0|\mathbf{x}) = \mathrm{P}(y = 1|\mathbf{x}).$$

Hence, the conditional expectation is equivalent to the conditional “success” probability $\mathrm{P}(y = 1|\mathbf{x})$.

Since probabilities are restricted to the range $[0, 1]$, so are model predictions.

Challenge: how to find $\mathrm{P}(y = 1|\mathbf{x})$? We have to make (a) a distributional assumption and (b) integrate the effects of $\mathbf{x}$ into the conditional probability. This can be achieved by a latent variable representation.

A latent variable representation

Discrete dependent-variable models are often cast in the form of latent variable models, where the latent variable is allowed to be continuous.

- Idea: model the difference between benefits and costs as an unobserved variable y^*. The "net benefit" can be expressed in terms of the covariates $\mathbf{x}_i$ (which is assumed to contain unity) as
$$y_i^* = \mathbf{x}_i\boldsymbol{\theta} + e_i.$$
- The net benefit can, e.g., be interpreted as the utility difference $U_{i1} - U_{i0}$ between the two alternatives.
- The econometrician does not observe the net benefit of the choice, only whether it is made or not:
$$y_i = \begin{cases} 1 & \text{if } \; {y_i}^* > 0 \\ 0 & \text{if } \; {y_i}^* \leq 0 \end{cases}$$

Assumptions

- The error e_i is independent of $\mathbf{x}_i$.
- It has mean zero, $E[e_i] = 0$, and *known* variance. Why the latter?
- Assume e_i has variance σ^2 such that e_i/σ has variance 1.
- Then the latent models

  $$y_i^* = \mathbf{x}_i\boldsymbol{\theta} + e_i \qquad \text{and} \qquad y_i^*/\sigma = \mathbf{x}_i\boldsymbol{\theta}/\sigma + e_i/\sigma$$

  generate the same observations because whether y_i is 0 or 1 depends only on the sign of y_i^*, not on its scale.
- This means that there is no information about σ in the sample data so σ cannot be estimated.
- Hence, the parameter vector in this model is only “identified up to scale”.

The conditional probability that y_i equals one is

$$\begin{aligned} p(\mathbf{x}_i) &= \mathrm{P}(y_i = 1|\mathbf{x}_i) = \mathrm{P}({y_i}^* > 0|\mathbf{x}_i) = \mathrm{P}(\mathbf{x}_i\boldsymbol{\theta} + e_i > 0|\mathbf{x}_i) \\ &= \mathrm{P}(e_i > -\mathbf{x}_i\boldsymbol{\theta}|\mathbf{x}_i) = 1 - \mathrm{P}(e_i \leq -\mathbf{x}_i\boldsymbol{\theta}|\mathbf{x}_i) \\ &= 1 - G(-\mathbf{x}_i\boldsymbol{\theta}) = G(\mathbf{x}_i\boldsymbol{\theta}). \end{aligned}$$

Remarks:

- $G(\cdot)$ is the cdf of e_i.
- The last step only holds in case the chosen probability distribution G is symmetric.
- The regressors should include a constant to ensure $E[e_i] = 0$.

Consequently, the conditional expectation can be written as

$$\mathrm{E}(y|\mathbf{x}) = \mathrm{P}(y = 1|\mathbf{x}) = G(\mathbf{x}\boldsymbol{\theta}).$$

Remarks:

- The effect of $\mathbf{x}$ on the conditional expectation works via the scalar index $\mathbf{x}\boldsymbol{\theta}$. Models of this kind are called **index models**.
- The conditional expectation depends on the index $\mathbf{x}\boldsymbol{\theta}$ through the **link function** G.
- Typically, we would require that the range of $G(\cdot)$ be restricted to $[0, 1]$. However, as we see below, this is not strictly necessary if we are only interested in the marginal effects.
- To estimate the conditional expectation we can apply nonlinear least squares (to this end, we need to specify some link function) or maximum likelihood (to this end, we need to specify the probability distribution of e).

Linear probability model

Specify

$$G(\mathbf{x}\boldsymbol{\theta}) = \mathbf{x}\boldsymbol{\theta}$$

such that

$$\mathrm{E}(y|\mathbf{x}) = \mathrm{P}(y = 1|\mathbf{x}) = \mathbf{x}\boldsymbol{\theta}.$$

Then we can define $e = y - \mathrm{E}(y|\mathbf{x})$ and write

$$y = \mathbf{x}\boldsymbol{\theta} + e, \qquad \mathrm{E}(e|\mathbf{x}) = 0.$$

Hence, given a random sample we can run OLS to estimate the parameters.

As shown below, the disturbances are heteroscedastic,

$$\mathrm{Var}(e|\mathbf{x}) = \mathbf{x}\boldsymbol{\theta}(1 - \mathbf{x}\boldsymbol{\theta}),$$

so we need to use robust standard errors.

Advantages of the linear probability model

- If we have a random sample, the OLS regression of y on $\mathbf{x}$ produces consistent and unbiased estimators of $\boldsymbol{\theta}$.
- The interpretation of $\boldsymbol{\theta}$ is simple and straightforward:

$$\frac{\partial E(y|\mathbf{x})}{\partial x_k} = \frac{\partial \, \mathrm{P}(y = 1|\mathbf{x})}{\partial x_k} = \frac{\partial \mathbf{x}\boldsymbol{\theta}}{\partial x_k} = \theta_k .$$

- No nonlinear estimation technique has to be used.
- It is straightforward to use instrumental variables or integrate the equation into a SEM.

Disadvantages of the linear probability model

- Without some ad hoc tinkering of the disturbances we cannot constrain $\mathbf{x}\boldsymbol{\theta}$ to the 0-1 interval. This implies that within the sample some predicted probabilities

$$\mathrm{E}(y_i|\mathbf{x}_i) = \mathrm{P}(y_i = 1|\mathbf{x}_i) = \mathbf{x}_i\boldsymbol{\theta}$$

  can take values above 1 or below 0.
- Hence, the nice interpretation of $\boldsymbol{\theta}$ comes at a cost: Near the boundaries it is just not possible that a change in, say, x_k has the same effect as in the middle of the distribution if this would shift the probability outside the 0-1 interval.
- Hence, in the population the linear probability model cannot be the correct specification. It should rather be viewed as a linear projection of y on the explanatory variables.
- Consequently, the partial effects of the linear probability model are the partial effects of the linear projection. In practice, however, the differences to nonlinear models are often small unless we consider extreme values of $\mathbf{x}$.

Conditional variance: the details (*)

(a) The disturbances have the following structure:

$$e = \begin{cases} 1 - \mathbf{x}\boldsymbol{\theta} & \text{with probability} \quad p(\mathbf{x}) \qquad (y = 1) \\ 0 - \mathbf{x}\boldsymbol{\theta} & \text{with probability} \quad 1 - p(\mathbf{x}) \quad (y = 0). \end{cases}$$

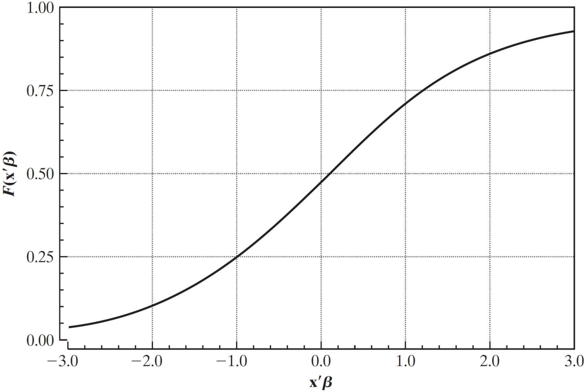
(b) Denoting $p(\mathbf{x}) = \mathrm{P}(y = 1|\mathbf{x}) = \mathbf{x}\boldsymbol{\theta}$, the expectation of e is

$$\begin{aligned} E(e|\mathbf{x}) &= (1 - \mathbf{x}\boldsymbol{\theta})p(\mathbf{x}) + (-\mathbf{x}\boldsymbol{\theta})(1 - p(\mathbf{x})) = p(\mathbf{x}) - \mathbf{x}\boldsymbol{\theta}p(\mathbf{x}) - \mathbf{x}\boldsymbol{\theta} + \mathbf{x}\boldsymbol{\theta}p(\mathbf{x}) \\ &= p(\mathbf{x}) - \mathbf{x}\boldsymbol{\theta} = \mathbf{x}\boldsymbol{\theta} - \mathbf{x}\boldsymbol{\theta} = 0. \end{aligned}$$

(c) The variance is

$$\begin{aligned} \mathrm{Var}(e|\mathbf{x}) &= E(e^2|\mathbf{x}) = (1 - \mathbf{x}\boldsymbol{\theta})^2 p(\mathbf{x}) + (-\mathbf{x}\boldsymbol{\theta})^2(1 - p(\mathbf{x})) \\ &= (1 - 2\mathbf{x}\boldsymbol{\theta} + (\mathbf{x}\boldsymbol{\theta})^2)\mathbf{x}\boldsymbol{\theta} + (\mathbf{x}\boldsymbol{\theta})^2 - (\mathbf{x}\boldsymbol{\theta})^3 = \mathbf{x}\boldsymbol{\theta} - (\mathbf{x}\boldsymbol{\theta})^2 = \mathbf{x}\boldsymbol{\theta}(1 - \mathbf{x}\boldsymbol{\theta}). \end{aligned}$$

The desired shape of the link function



Logit and probit model

Probit model: standard normal distribution for e_i

$$\mathrm{P}(y = 1|\mathbf{x}) = G(\mathbf{x}\boldsymbol{\theta}) = \Phi(\mathbf{x}\boldsymbol{\theta}) = \int_{-\infty}^{\mathbf{x}\boldsymbol{\theta}} \phi(t)dt$$

where $\Phi(\cdot)$ and $\phi(\cdot)$ are the cdf and pdf, respectively, of the standard normal distribution.

Logit model: logistic distribution for e_i

$$\mathrm{P}(y = 1|\mathbf{x}) = G(\mathbf{x}\boldsymbol{\theta}) = \Lambda(\mathbf{x}\boldsymbol{\theta}) = \frac{\exp(\mathbf{x}\boldsymbol{\theta})}{1 + \exp(\mathbf{x}\boldsymbol{\theta})},$$

where $\Lambda(\cdot)$ is the cdf of a standard logistic distribution.

Advantage of the nonlinear probability models

- The predicted values for y_i lie by construction in the 0-1 interval.

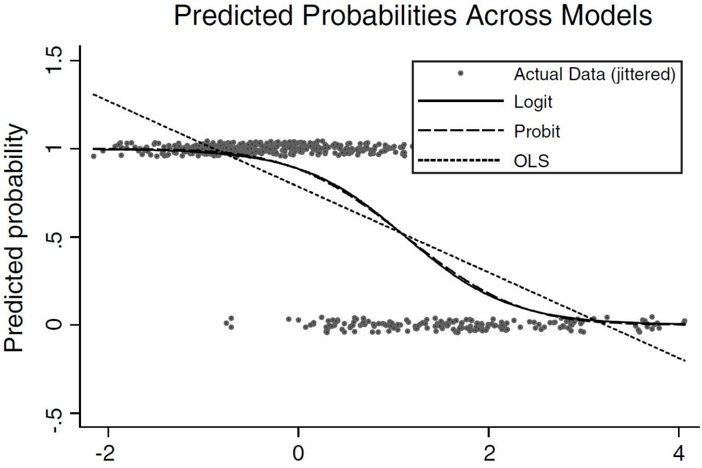
Disadvantage of the nonlinear probability models

- We need a nonlinear estimation technique. We could use NLS because the conditional expectation

  $$\mathrm{E}(y|\mathbf{x}) = G(\mathbf{x}\boldsymbol{\theta})$$

  is a nonlinear function (we called it $m(\mathbf{x}, \boldsymbol{\theta})$ in the previous lecture). However, we will apply maximum likelihood estimation.
- Interpretation of $\boldsymbol{\theta}$ is more difficult than in the linear probability model. In particular, θ_k is *not* the marginal effect of x_k on $\mathrm{P}(y = 1|\mathbf{x})$ (for the correct interpretation, see below).

Logit and Probit versus OLS



2. Introduction to Maximum Likelihood Estimation

Definition

Denote the density of a random variable y_i

$$f(y_i|\boldsymbol{\theta}_o),$$

where $\boldsymbol{\theta}_o \in \boldsymbol{\Theta}$ is the true parameter vector.

Define the **log likelihood for observation** i as

$$\ell_i(\boldsymbol{\theta}) \equiv \ell(y_i, \boldsymbol{\theta}) = \log f(y_i|\boldsymbol{\theta})$$

As shown in the textbook (p. 473-474), the true parameter vector $\boldsymbol{\theta}_o$ solves

$$\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathrm{E}[\ell_i(\boldsymbol{\theta})].$$

The **maximum likelihood estimator** (MLE) of $\boldsymbol{\theta}_o$ is the vector $\hat{\boldsymbol{\theta}}$ that solves

$$\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} N^{-1} \sum_{i=1}^{N} \ell_i(\boldsymbol{\theta}) = \max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} N^{-1} \sum_{i=1}^{N} \log f(y_i|\boldsymbol{\theta}).$$

Example: Estimating the parameter of a Bernoulli distribution

Suppose y_i follows a Bernoulli distribution, $y_i \sim \mathrm{Ber}(\theta)$.

This means that y_i is 1 with probability θ and 0 with probability $1 - \theta$.

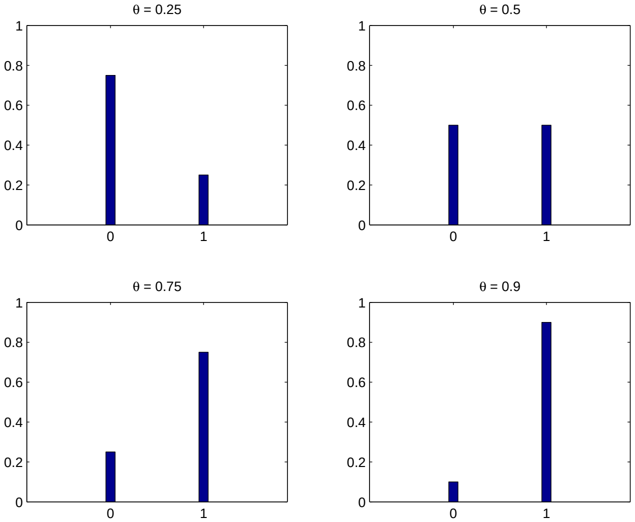
The expectation of this random variable thus is

$$\mathrm{E}(y_i) = 1 \times \mathrm{P}(y = 1) + 0 \times \mathrm{P}(y = 0) = \mathrm{P}(y = 1) = \theta.$$

The probability mass function (density = continuous variables) is

$$\mathrm{P}(y_i|\theta) = f(y_i|\theta) = \theta^{y_i}(1 - \theta)^{1 - y_i}, \qquad y_i \in \{0, 1\}, \theta \in [0, 1].$$

Probability mass functions for different choices of θ



Example: Sample of size 1

Let us think of the following setup:

- We know the parameter θ is either 0.25 or 0.75.
- We have a sample of size 1.
- We want to find the most likely value of θ.

Possible outcomes:

	True $\theta_o = 0.25$	True $\theta_o = 0.75$
Sample $y_1 = 0$	$f(y_1 = 0 \mid \theta_o = 0.25) = 0.75$	$f(y_1 = 0 \mid \theta_o = 0.75) = 0.25$
Sample $y_1 = 1$	$f(y_1 = 1 \mid \theta_o = 0.25) = 0.25$	$f(y_1 = 1 \mid \theta_o = 0.75) = 0.75$

Hence, if $y_1 = 0$, we choose $\hat{\theta} = 0.25$; if $y_1 = 1$, we choose $\hat{\theta} = 0.75$. This is a maximum likelihood estimator.

Now let us change the setup to:

- We know the parameter is between 0 and 1: $\theta \in [0, 1]$.
- We have a sample of size 1.

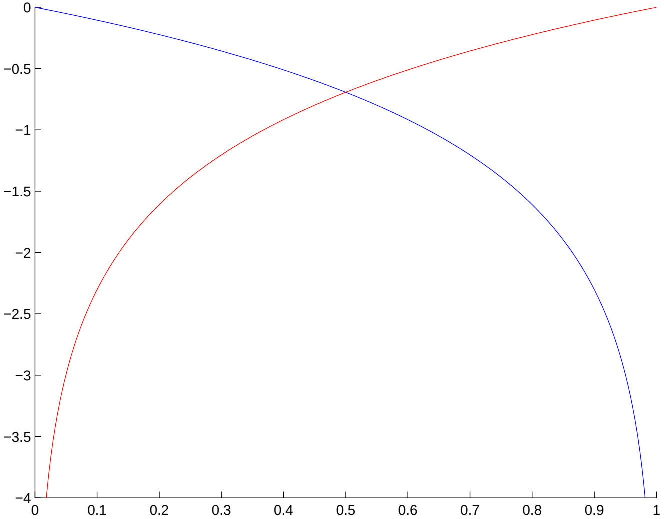
This is more complicated. Let us use the log likelihood function to find the estimator:

$$\ell_1(\theta) = \log f(y_1|\theta) = y_1 \log(\theta) + (1 - y_1) \log(1 - \theta)$$

$$\begin{array}{ll} \text{Sample } y_1 = 0: & \ell_1(\theta) = \log(1 - \theta) \\ \text{Sample } y_1 = 1: & \ell_1(\theta) = \log(\theta) \end{array}$$

We choose that value as $\hat{\theta}$ which maximizes the log likelihood function.

Log likelihood function for sample of size 1 given $y_1 = 0$ (blue) and given $y_1 = 1$ (red)



Example: Sample of size 2

Now let us again change the setup to:

- We know the parameter $\theta \in [0, 1]$.
- We have a sample of size 2.

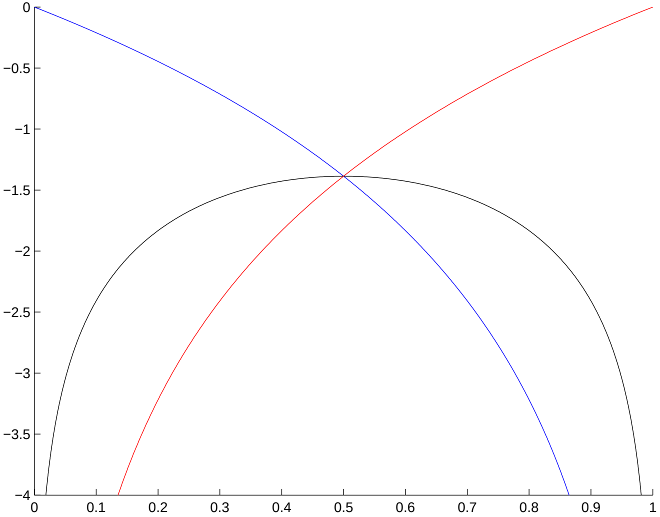
The log likelihood function becomes:

$$\begin{aligned}\mathcal{L}(\theta) &= \ell_1(\theta) + \ell_2(\theta) \\ &= [y_1 \log(\theta) + (1 - y_1) \log(1 - \theta)] + [y_2 \log(\theta) + (1 - y_2) \log(1 - \theta)]\end{aligned}$$

$$\begin{array}{ll}\text{Sample } y_1 = y_2 = 0: & \mathcal{L}(\theta) = 2 \log(1 - \theta) \\ \text{Sample } y_1 = 0, y_2 = 1: & \mathcal{L}(\theta) = \log(1 - \theta) + \log(\theta) \\ \text{Sample } y_1 = 1, y_2 = 0: & \mathcal{L}(\theta) = \log(\theta) + \log(1 - \theta) \\ \text{Sample } y_1 = y_2 = 1: & \mathcal{L}(\theta) = 2 \log(\theta)\end{array}$$

We again choose that value as $\hat{\theta}$ which maximizes the log likelihood function.

Log likelihood function for sample of size 2 given $y_1 = y_2 = 0$ (blue), $y_1 = 1, y_2 = 0$ or vice versa (black), and $y_1 = y_2 = 1$ (red)



Example: general case

Now let us consider the general case of some sample size N.

The log likelihood function becomes:

$$\begin{aligned}\mathcal{L}(\theta) &= \sum_{i=1}^{N} \ell_i(\theta) = \sum_{i=1}^{N} [y_i \log(\theta) + (1-y_i)\log(1-\theta)] \\ &= \log(\theta)\sum_{i=1}^{N} y_i + \log(1-\theta)\sum_{i=1}^{N}(1-y_i) = \log(\theta)N\bar{y} + \log(1-\theta)N(1-\bar{y})\end{aligned}$$

First derivative on the log likelihood function:

$$\frac{\partial \mathcal{L}}{\partial \theta} = \frac{N\bar{y}}{\theta} - \frac{N(1-\bar{y})}{1-\theta}$$

FOC for a maximum defines the estimator:

$$\frac{N\bar{y}}{\hat{\theta}} - \frac{N(1-\bar{y})}{1-\hat{\theta}} \overset{!}{=} 0 \qquad \Rightarrow \qquad \hat{\theta} = \bar{y}$$

3. Maximum Likelihood Estimator

General setup (including covariates)

Denote the (correctly specified) conditional density for a random vector $\mathbf{y}_i$ given $\mathbf{x}_i$ as

$$f(\mathbf{y}_i|\mathbf{x}_i;\boldsymbol{\theta}_o),$$

where $\boldsymbol{\theta}_o \in \boldsymbol{\Theta}$ is the true parameter vector.

Define the **conditional log likelihood for observation** i as

$$\ell_i(\boldsymbol{\theta}) \equiv \ell(\mathbf{y}_i, \mathbf{x}_i, \boldsymbol{\theta}) = \log f(\mathbf{y}_i|\mathbf{x}_i;\boldsymbol{\theta})$$

Remarks:

- This is a random function of $\boldsymbol{\theta}$ since it depends on the random vector $(\mathbf{y}_i, \mathbf{x}_i)$.
- It is conditional on $\mathbf{x}_i$. Hence, we talk about conditional maximum likelihood estimation (CML) even though we will typically call it just maximum likelihood estimation (ML).

As shown in the textbook (p. 473-474), the true parameter vector $\boldsymbol{\theta}_o$ solves

$$\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathrm{E}[\ell_i(\boldsymbol{\theta})].$$

The **conditional maximum likelihood estimator** (CMLE) of $\boldsymbol{\theta}_o$ is the vector $\hat{\boldsymbol{\theta}}$ that solves the sample analogue

$$\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} N^{-1}\sum_{i=1}^{N}\ell_i(\boldsymbol{\theta}) = \max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} N^{-1}\sum_{i=1}^{N}\log f(\mathbf{y}_i|\mathbf{x}_i;\boldsymbol{\theta}).$$

This has important consequences:

- The CMLE is an M-estimator. Define $\mathbf{w}_i = (\mathbf{x}_i, \mathbf{y}_i)$ and $q(\mathbf{w}_i, \boldsymbol{\theta}) \equiv -\ell_i(\boldsymbol{\theta})$ to see this.
- Hence, we can use all the results and estimation procedures we obtained for the M-estimator.
- In particular, the CMLE is consistent and asymptotically normally distributed.

Consistency of the CMLE

The CMLE is consistent if the following key assumptions are satisfied:

- the parametric model is correctly specified (this includes that the distributional assumption is correct),
- $\boldsymbol{\theta}_o$ is identified (in just the same way as discussed for the M-estimator),
- the log likelihood function is continuous in $\boldsymbol{\theta}$.

For the details, see Theorem 13.1 on p. 475, or the discussion of the M-estimator.

Asymptotic normality of the CMLE

The CMLE is asymptotically normally distributed if the following additional assumptions are satisfied:

- the log likelihood function is twice continuously differentiable with respect to $\boldsymbol{\theta}$ and
- the score of the log likelihood function has conditional mean zero,

$$\mathrm{E}[\mathbf{s}_i(\boldsymbol{\theta})|\mathbf{x}_i] = \mathbf{0}.$$

Note that the conditional mean zero assumption is satisfied in all cases we consider. It requires interchangeability of an integral and a derivative (recall that we assumed that for the M-estimator, too). For the details, see equation (13.21) in the textbook on p. 477 and the discussion on p. 479.

Score and Hessian

The score of the log likelihood function is

$$\mathbf{s}_i(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta})' = \left[\frac{\partial \ell_i}{\partial \theta_1}(\boldsymbol{\theta}), \ldots, \frac{\partial \ell_i}{\partial \theta_P}(\boldsymbol{\theta})\right]'.$$

The Hessian of the log likelihood function is

$$\mathbf{H}_i(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \mathbf{s}_i(\boldsymbol{\theta}) = \nabla^2_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}) = \begin{pmatrix} \frac{\partial^2 \ell_i(\boldsymbol{\theta})}{\partial \theta_1 \partial \theta_1} & \frac{\partial^2 \ell_i(\boldsymbol{\theta})}{\partial \theta_1 \partial \theta_2} & \cdots & \frac{\partial^2 \ell_i(\boldsymbol{\theta})}{\partial \theta_1 \partial \theta_P} \\ \frac{\partial^2 \ell_i(\boldsymbol{\theta})}{\partial \theta_2 \partial \theta_1} & \frac{\partial^2 \ell_i(\boldsymbol{\theta})}{\partial \theta_2 \partial \theta_2} & \cdots & \frac{\partial^2 \ell_i(\boldsymbol{\theta})}{\partial \theta_2 \partial \theta_P} \\ \vdots & \vdots & \ddots & \\ \frac{\partial^2 \ell_i(\boldsymbol{\theta})}{\partial \theta_P \partial \theta_1} & \frac{\partial^2 \ell_i(\boldsymbol{\theta})}{\partial \theta_P \partial \theta_2} & \cdots & \frac{\partial^2 \ell_i(\boldsymbol{\theta})}{\partial \theta_P \partial \theta_P} \end{pmatrix}$$

Asymptotic distribution

Define

$$\mathbf{A}_o = -\,\mathrm{E}[\mathbf{H}_i(\boldsymbol{\theta}_o)]$$

as minus the Hessian because we have a maximization problem (in contrast to our formulation of M-estimation as a minimization problem), and the **Fisher information matrix**

$$\mathbf{B}_o = \mathrm{Var}[\mathbf{s}_i(\boldsymbol{\theta}_o)] = \mathrm{E}[\mathbf{s}_i(\boldsymbol{\theta}_o)\mathbf{s}_i(\boldsymbol{\theta}_o)'].$$

Then we can apply the result obtained for the M-estimator: the CMLE has limiting distribution

$$\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o) \xrightarrow{\mathrm{d}} \mathrm{Normal}(\mathbf{0}, \mathbf{A}_o^{-1}\mathbf{B}_o\mathbf{A}_o^{-1})$$

with asymptotic variance

$$\mathbf{V} = \mathbf{A}_o^{-1}\mathbf{B}_o\mathbf{A}_o^{-1}.$$

Information matrix equality

Under fairly general conditions, in the maximum likelihood context the **conditional information matrix equality (CIME)**

$$-\,\mathrm{E}[\mathbf{H}_i(\boldsymbol{\theta}_o)|\mathbf{x}_i] = \mathrm{E}[\mathbf{s}_i(\boldsymbol{\theta}_o)\mathbf{s}_i(\boldsymbol{\theta}_o)'|\mathbf{x}_i]$$

holds. Taking expectations with respect to the distribution of $\mathbf{x}_i$ and using the law of iterated expectations yields the **unconditional information matrix equality (UIME)**

$$-\,\mathrm{E}[\mathbf{H}_i(\boldsymbol{\theta}_o)] = \mathrm{E}[\mathbf{s}_i(\boldsymbol{\theta}_o)\mathbf{s}_i(\boldsymbol{\theta}_o)'] \quad \Leftrightarrow \quad \mathbf{A}_o = \mathbf{B}_o.$$

This simplifies the limiting distribution of the CMLE to

$$\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o) \xrightarrow{\mathrm{d}} \mathrm{Normal}(\mathbf{0}, \mathbf{A}_o^{-1}).$$

Hence,

$$\mathrm{Avar}(\hat{\boldsymbol{\theta}}) = \mathbf{A}_o^{-1}/N = \mathbf{V}/N.$$

Proof: (*)

Consider a continuous conditional distribution for $\mathbf{y}$. Then the expectation of the score (which is zero by assumption) is (for all $\boldsymbol{\theta}$)

$$\mathrm{E}_{\boldsymbol{\theta}}[\mathbf{s}_i(\boldsymbol{\theta})|\mathbf{x}_i] \equiv \int_{\mathcal{Y}} \mathbf{s}_i(\boldsymbol{\theta}) f(\mathbf{y}_i|\mathbf{x}_i;\boldsymbol{\theta}) d\mathbf{y} = \mathbf{0}.$$

Taking the first derivative on both sides yields

$$\nabla_{\boldsymbol{\theta}} \int_{\mathcal{Y}} \mathbf{s}_i(\boldsymbol{\theta}) f(\mathbf{y}_i|\mathbf{x}_i;\boldsymbol{\theta}) d\mathbf{y} = \mathbf{0}.$$

Interchanging differentiation and integration (assuming we are allowed to do so) yields

$$\int_{\mathcal{Y}} \nabla_{\boldsymbol{\theta}} \left[\mathbf{s}_i(\boldsymbol{\theta}) f(\mathbf{y}_i|\mathbf{x}_i;\boldsymbol{\theta})\right] d\mathbf{y} = \mathbf{0}.$$

By the product rule for differentiation we obtain

$$\int_{\mathcal{Y}} \left[\nabla_{\boldsymbol{\theta}} \mathbf{s}_i(\boldsymbol{\theta}) f(\mathbf{y}_i|\mathbf{x}_i;\boldsymbol{\theta}) + \mathbf{s}_i(\boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}} f(\mathbf{y}_i|\mathbf{x}_i;\boldsymbol{\theta})\right] d\mathbf{y} = \mathbf{0}.$$

Recall that

$$\mathbf{s}_i(\boldsymbol{\theta}) = \nabla'_{\boldsymbol{\theta}}\ell_i(\boldsymbol{\theta}) = \nabla'_{\boldsymbol{\theta}}[\log f(\mathbf{y}_i|\mathbf{x}_i;\boldsymbol{\theta})] = \frac{\nabla'_{\boldsymbol{\theta}} f(\mathbf{y}_i|\mathbf{x}_i;\boldsymbol{\theta})}{f(\mathbf{y}_i|\mathbf{x}_i;\boldsymbol{\theta})}.$$

Hence,

$$\nabla'_{\boldsymbol{\theta}} f(\mathbf{y}_i|\mathbf{x}_i;\boldsymbol{\theta}) = \mathbf{s}_i(\boldsymbol{\theta}) f(\mathbf{y}_i|\mathbf{x}_i;\boldsymbol{\theta}).$$

Substituting the transpose in the above expression yields

$$\mathbf{0} = \int_{\mathcal{Y}} \left[\nabla_{\boldsymbol{\theta}}\mathbf{s}_i(\boldsymbol{\theta}) f(\mathbf{y}_i|\mathbf{x}_i;\boldsymbol{\theta}) + \mathbf{s}_i(\boldsymbol{\theta})\mathbf{s}_i(\boldsymbol{\theta})' f(\mathbf{y}_i|\mathbf{x}_i;\boldsymbol{\theta})\right] d\mathbf{y}$$

$$\mathbf{0} = \int_{\mathcal{Y}} \left[\mathbf{H}_i(\boldsymbol{\theta}) f(\mathbf{y}_i|\mathbf{x}_i;\boldsymbol{\theta}) + \mathbf{s}_i(\boldsymbol{\theta})\mathbf{s}_i(\boldsymbol{\theta})' f(\mathbf{y}_i|\mathbf{x}_i;\boldsymbol{\theta})\right] d\mathbf{y}$$

$$\mathbf{0} = \underbrace{\int_{\mathcal{Y}} \mathbf{H}_i(\boldsymbol{\theta}) f(\mathbf{y}_i|\mathbf{x}_i;\boldsymbol{\theta}) d\mathbf{y}}_{\mathrm{E}_{\boldsymbol{\theta}}[\mathbf{H}_i(\boldsymbol{\theta})|\mathbf{x}_i]} + \underbrace{\int_{\mathcal{Y}} \mathbf{s}_i(\boldsymbol{\theta})\mathbf{s}_i(\boldsymbol{\theta})' f(\mathbf{y}_i|\mathbf{x}_i;\boldsymbol{\theta}) d\mathbf{y}}_{\mathrm{E}_{\boldsymbol{\theta}}[\mathbf{s}_i(\boldsymbol{\theta})\mathbf{s}_i(\boldsymbol{\theta})'|\mathbf{x}_i]}$$

which can be evaluated at $\boldsymbol{\theta} = \boldsymbol{\theta}_o$ which yields

$$-\mathrm{E}[\mathbf{H}_i(\boldsymbol{\theta}_o)|\mathbf{x}_i] = \mathrm{E}[\mathbf{s}_i(\boldsymbol{\theta}_o)\mathbf{s}_i(\boldsymbol{\theta}_o)'|\mathbf{x}_i].$$

Three estimators of the variance

(1) Direct estimate of the Hessian:

$$\hat{\mathbf{A}}_o = N^{-1} \sum_{i=1}^{N} -\mathbf{H}_i(\hat{\boldsymbol{\theta}})$$

(2) Estimate of the conditional expectation $\mathbf{A}(\mathbf{x}_i, \boldsymbol{\theta}_o) \equiv -\mathrm{E}[\mathbf{H}(\mathbf{y}_i, \mathbf{x}_i, \boldsymbol{\theta}_o)|\mathbf{x}_i]$ of the Hessian:

$$\hat{\mathbf{A}}_o = N^{-1} \sum_{i=1}^{N} \mathbf{A}(\mathbf{x}_i, \hat{\boldsymbol{\theta}})$$

(3) Outer product of the score:

$$\hat{\mathbf{A}}_o = N^{-1} \sum_{i=1}^{N} \mathbf{s}_i(\hat{\boldsymbol{\theta}})\mathbf{s}_i(\hat{\boldsymbol{\theta}})'$$

4. Hypothesis Testing

Types of tests

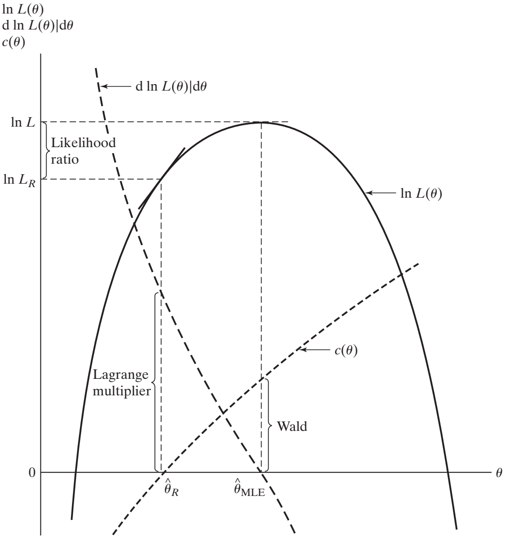
Suppose you want to test the null hypothesis concerning $\boldsymbol{\theta}$,

$$H_0 : \mathbf{c}(\boldsymbol{\theta}) = \mathbf{0}.$$

This includes linear hypotheses like $\mathbf{R}\boldsymbol{\theta} - \mathbf{r} = \mathbf{0}$, and nonlinear hypotheses like $\theta_1 - \theta_2/\theta_3 = 0$.

In the maximum likelihood context, we can use three different types of tests:

- Wald test: test statistic is based on the distance of $\mathbf{c}(\hat{\boldsymbol{\theta}})$ from $\mathbf{0}$.
- Likelihood ratio test: test statistic is based on the ratio of the unrestricted likelihood value (under H_1) to the restricted likelihood value (under H_0).
- Lagrange multiplier test (score test): test statistic is based on the slope of the score under H_0.



Source: Greene (2012) Econometric Analysis, p. 525.

Wald tests

Nothing new compared to the M-estimation context.

Given the asymptotic normal distribution of the CMLE, Wald tests of linear hypotheses of the kind $H_0 : \mathbf{R}\boldsymbol{\theta} = \mathbf{r}$ can be performed as in the linear regression case.

The limiting distribution of the Wald statistic for Q restrictions is

$$W_N \equiv \left[\mathbf{R}\hat{\boldsymbol{\theta}} - \mathbf{r}\right]' \left[\mathbf{R}(\hat{\mathbf{V}}/N)\mathbf{R}'\right]^{-1} \left[\mathbf{R}\hat{\boldsymbol{\theta}} - \mathbf{r}\right] \overset{a}{\sim} \chi^2_Q.$$

For nonlinear hypotheses $\mathbf{c}(\boldsymbol{\theta}_o) = \mathbf{0}$, the Wald statistic is

$$W_N \equiv \mathbf{c}(\hat{\boldsymbol{\theta}})' \left[\mathbf{C}(\hat{\boldsymbol{\theta}})(\hat{\mathbf{V}}/N)\mathbf{C}(\hat{\boldsymbol{\theta}})'\right]^{-1} \mathbf{c}(\hat{\boldsymbol{\theta}}) \overset{a}{\sim} \chi^2_Q,$$

where $\mathbf{C}(\cdot)$ is the $Q \times P$ Jacobian

$$\mathbf{C}(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}}\mathbf{c}(\boldsymbol{\theta})$$

Likelihood ratio tests

Define the sample log likelihood function

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{i=1}^{N} \ell_i(\boldsymbol{\theta}) = \sum_{i=1}^{N} \log f(\mathbf{y}_i|\mathbf{x}_i;\boldsymbol{\theta})$$

and the estimators

$\tilde{\boldsymbol{\theta}}$ restricted estimator (estimated under $H_0 : \mathbf{c}(\boldsymbol{\theta}_o) = \mathbf{0}$)

$\hat{\boldsymbol{\theta}}$ unrestricted estimator (estimated under $H_1 : \mathbf{c}(\boldsymbol{\theta}_o) \neq \mathbf{0}$)

Then the **likelihood ratio (LR) statistic**

$$LR \equiv 2[\mathcal{L}(\hat{\boldsymbol{\theta}}) - \mathcal{L}(\tilde{\boldsymbol{\theta}})]$$

is distributed asymptotically as χ^2_Q under H_0, where Q is the number of restrictions. (A proof can be found in the textbook on p. 429.)

Lagrange multiplier tests (or score tests)

The Lagrange multiplier (LM) statistic offers a way to test $H_0 : \mathbf{c}(\boldsymbol{\theta}_o) = \mathbf{0}$ by only estimating the model under the null. This is often convenient if estimation of the unrestricted model is complicated.

Let $\tilde{\mathbf{s}}_i = \mathbf{s}_i(\tilde{\boldsymbol{\theta}})$ be the $P \times 1$ score evaluated at the restricted estimate $\tilde{\boldsymbol{\theta}}$. This implies that you

(1) have to set up the likelihood function of the unconstrained model,

(2) compute the partial derivatives of $\ell_i(\boldsymbol{\theta})$ with respect to *each* of the P parameters (also with respect to those which might be restricted under H_0), and

(3) evaluate this vector of partial derivatives at the *restricted* estimates.

Hence, you still have to write down the unrestricted model but you only have to estimate the restricted one.

Then the **lagrange multiplier (LM) statistic**

$$LM \equiv \left(N^{-1/2} \sum_{i=1}^{N} \tilde{\mathbf{s}}_i \right)' \tilde{\mathbf{A}}^{-1} \left(N^{-1/2} \sum_{i=1}^{N} \tilde{\mathbf{s}}_i \right)$$

is distributed asymptotically as χ^2_Q under H_0, where Q is the number of restrictions. (A proof can be found in the textbook on p. 422-424.)

Note that $\tilde{\mathbf{A}}$ is an estimator of $\mathbf{A}_o$ evaluated under H_0. Due to the information matrix equality, three different estimators can be applied:

(1) Direct estimate of the Hessian:

$$\tilde{\mathbf{A}} = N^{-1} \sum_{i=1}^{N} -\mathbf{H}_i(\tilde{\boldsymbol{\theta}})$$

(2) Estimate of the conditional expectation $\mathbf{A}(\mathbf{x}_i, \boldsymbol{\theta}_o)$ of the Hessian:

$$\tilde{\mathbf{A}} = N^{-1} \sum_{i=1}^{N} \mathbf{A}(\mathbf{x}_i, \tilde{\boldsymbol{\theta}})$$

(3) Outer product of the score:

$$\tilde{\mathbf{A}} = N^{-1} \sum_{i=1}^{N} \tilde{\mathbf{s}}_i \tilde{\mathbf{s}}_i'$$

Outer product of the score LM statistic

While the outer product of the score is sometimes a bit unreliable (unless N is really large) as an estimator for $\mathbf{A}_o$, using it leads to a particularly simple expression of the LM statistic.

Substituting yields

$$LM = \left(\sum_{i=1}^{N} \tilde{\mathbf{s}}_i \right)' \left(\sum_{i=1}^{N} \tilde{\mathbf{s}}_i \tilde{\mathbf{s}}_i' \right)^{-1} \left(\sum_{i=1}^{N} \tilde{\mathbf{s}}_i \right).$$

This statistic equals N times the uncentered R-squared, say R_0^2, from the regression

$$1 \text{ on } \tilde{\mathbf{s}}_i', \qquad i = 1, \ldots, N.$$

This particularly easy to compute.

5. ML Estimation of Binary Choice Models

Setup of the likelihood function

Let us use CML estimation to find the parameters of the probit and logit models introduced above.

Since y_i is a 0-1 variable and follows conditional on $\mathbf{x}_i$ a Bernoulli distribution, its conditional density is

$$f(y_i|\mathbf{x}_i;\boldsymbol{\theta}) = \mathrm{P}(y_i = 1|\mathbf{x}_i)^{y_i}[1 - \mathrm{P}(y_i = 1|\mathbf{x}_i)]^{1-y_i} = G(\mathbf{x}_i\boldsymbol{\theta})^{y_i}[1 - G(\mathbf{x}_i\boldsymbol{\theta})]^{1-y_i}.$$

The log likelihood function for observation i thus is

$$\ell_i(\boldsymbol{\theta}) = \log f(y_i|\mathbf{x}_i;\boldsymbol{\theta}) = y_i \log[G(\mathbf{x}_i\boldsymbol{\theta})] + (1 - y_i)\log[1 - G(\mathbf{x}_i\boldsymbol{\theta})].$$

Finally, the log likelihood function for a sample of size N is

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{i=1}^{N} \ell_i(\boldsymbol{\theta}) = \sum_{i=1}^{N} \{y_i \log[G(\mathbf{x}_i\boldsymbol{\theta})] + (1 - y_i)\log[1 - G(\mathbf{x}_i\boldsymbol{\theta})]\}.$$

The Score

$$\begin{aligned}
\mathbf{s}_i(\boldsymbol{\theta}) = \nabla'_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}) &= \frac{\partial \ell_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \\
&= y_i \frac{\partial \log[G(\mathbf{x}_i \boldsymbol{\theta})]}{\partial \boldsymbol{\theta}} + (1 - y_i) \frac{\partial \log[1 - G(\mathbf{x}_i \boldsymbol{\theta})]}{\partial \boldsymbol{\theta}} \\
&= \frac{y_i}{G(\mathbf{x}_i \boldsymbol{\theta})} \frac{\partial G(\mathbf{x}_i \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} + \frac{1 - y_i}{1 - G(\mathbf{x}_i \boldsymbol{\theta})} \frac{\partial [1 - G(\mathbf{x}_i \boldsymbol{\theta})]}{\partial \boldsymbol{\theta}} \\
&= \frac{y_i}{G(\mathbf{x}_i \boldsymbol{\theta})} G'(\mathbf{x}_i \boldsymbol{\theta}) \mathbf{x}'_i - \frac{1 - y_i}{1 - G(\mathbf{x}_i \boldsymbol{\theta})} G'(\mathbf{x}_i \boldsymbol{\theta}) \mathbf{x}'_i \\
&= \left[\frac{y_i}{G(\mathbf{x}_i \boldsymbol{\theta})} - \frac{1 - y_i}{1 - G(\mathbf{x}_i \boldsymbol{\theta})} \right] g(\mathbf{x}_i \boldsymbol{\theta}) \mathbf{x}'_i,
\end{aligned}$$

where $g(\cdot) \equiv G'(\cdot)$ is the first derivative of the link function $G(\cdot)$.

Note: the link function is a cdf, hence $g(\cdot)$ is a pdf.

Defining $u_i \equiv y_i - \mathrm{E}(y_i|\mathbf{x}_i) = y_i - G(\mathbf{x}_i\boldsymbol{\theta}_o)$, the score can be further simplified to

$$\begin{aligned}
\mathbf{s}_i(\boldsymbol{\theta}) &= \left[\frac{y_i}{G(\mathbf{x}_i\boldsymbol{\theta})} - \frac{1-y_i}{1-G(\mathbf{x}_i\boldsymbol{\theta})}\right] g(\mathbf{x}_i\boldsymbol{\theta})\mathbf{x}_i' \\
&= \frac{y_i - G(\mathbf{x}_i\boldsymbol{\theta})}{G(\mathbf{x}_i\boldsymbol{\theta})[1-G(\mathbf{x}_i\boldsymbol{\theta})]}\, g(\mathbf{x}_i\boldsymbol{\theta})\mathbf{x}_i' \\
&= \frac{g(\mathbf{x}_i\boldsymbol{\theta})}{G(\mathbf{x}_i\boldsymbol{\theta})[1-G(\mathbf{x}_i\boldsymbol{\theta})]}\, \mathbf{x}_i' u_i .
\end{aligned}$$

First order condition

The CMLE $\hat{\boldsymbol{\theta}}$ solves the FOC

$$\sum_{i=1}^{N} \mathbf{s}_i(\hat{\boldsymbol{\theta}}) = \sum_{i=1}^{N} \frac{g(\mathbf{x}_i\hat{\boldsymbol{\theta}})}{G(\mathbf{x}_i\hat{\boldsymbol{\theta}})[1 - G(\mathbf{x}_i\hat{\boldsymbol{\theta}})]}\, \mathbf{x}_i' \hat{u}_i = \mathbf{0}.$$

Since this FOC cannot be solved explicitly for $\hat{\boldsymbol{\theta}}$, we need a numerical optimization procedure (Newton-Raphson or similar) to obtain the CMLE $\hat{\boldsymbol{\theta}}$.

For inference, we need the asymptotic variance matrix. To this end, we calculate the Hessian next.

The Hessian

Finding the Hessian turns out to be a bit tedious. Starting from

$$\mathbf{s}_i(\boldsymbol{\theta}) = \left[\frac{y_i}{G(\mathbf{x}_i\boldsymbol{\theta})} - \frac{1-y_i}{1-G(\mathbf{x}_i\boldsymbol{\theta})} \right] g(\mathbf{x}_i\boldsymbol{\theta})\mathbf{x}_i'$$

and defining $G_i \equiv G(\mathbf{x}_i\boldsymbol{\theta})$ and $g_i \equiv g(\mathbf{x}_i\boldsymbol{\theta})$, the product rule of differentiation yields

$$\mathbf{H}_i(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}}\mathbf{s}_i(\boldsymbol{\theta}) = -\left[\frac{y_i g_i}{G_i^2} + \frac{(1-y_i)g_i}{[1-G_i]^2} \right] g_i \mathbf{x}_i'\mathbf{x}_i + \left[\frac{y_i}{G_i} - \frac{1-y_i}{1-G_i} \right] g'(\mathbf{x}_i\boldsymbol{\theta})\mathbf{x}_i'\mathbf{x}_i.$$

This expression is simplified considerably if we take conditional expectation at $\boldsymbol{\theta}_o$,

$$\mathbf{A}(\mathbf{x}_i, \boldsymbol{\theta}_o) = -\operatorname{E}[\mathbf{H}_i(\boldsymbol{\theta}_o)|\mathbf{x}],$$

because $\operatorname{E}[y_i|\mathbf{x}, \boldsymbol{\theta}_o] = G(\mathbf{x}_i\boldsymbol{\theta}_o)$ so that the term in the second square brackets becomes

$$\operatorname{E}\left[\frac{y_i}{G_i} - \frac{1-y_i}{1-G_i} \middle| \mathbf{x}, \boldsymbol{\theta}_o \right] = \frac{G(\mathbf{x}_i\boldsymbol{\theta}_o)}{G(\mathbf{x}_i\boldsymbol{\theta}_o)} - \frac{1-G(\mathbf{x}_i\boldsymbol{\theta}_o)}{1-G(\mathbf{x}_i\boldsymbol{\theta}_o)} = 0.$$

Hence,

$$
\begin{aligned}
\mathbf{A}(\mathbf{x}_i, \boldsymbol{\theta}_o) &= -\operatorname{E}\left\{ -\left[\frac{y_i g_i}{G_i^2} + \frac{(1-y_i) g_i}{[1-G_i]^2} \right] g_i \mathbf{x}_i' \mathbf{x}_i \middle| \mathbf{x}, \boldsymbol{\theta}_o \right\} \\
&= \left[\frac{G_i g_i}{G_i^2} + \frac{(1-G_i) g_i}{[1-G_i]^2} \right] g_i \mathbf{x}_i' \mathbf{x}_i \\
&= \left[\frac{1}{G_i} + \frac{1}{1-G_i} \right] g_i^2 \mathbf{x}_i' \mathbf{x}_i \\
&= \frac{g(\mathbf{x}_i \boldsymbol{\theta}_o)^2}{G(\mathbf{x}_i \boldsymbol{\theta}_o)[1-G(\mathbf{x}_i \boldsymbol{\theta}_o)]} \mathbf{x}_i' \mathbf{x}_i
\end{aligned}
$$

Estimator of the asymptotic variance

A consistent estimator of the asymptotic variance thus is

$$\hat{\mathbf{V}} = \left[N^{-1} \sum_{i=1}^{N} \frac{g(\mathbf{x}_i\hat{\theta})^2}{G(\mathbf{x}_i\hat{\theta})[1 - G(\mathbf{x}_i\hat{\theta})]} \mathbf{x}_i'\mathbf{x}_i \right]^{-1} .$$

Hence, we obtain

$$\widehat{\text{Avar}}(\hat{\theta}) = \hat{\mathbf{V}}/N = \left[\sum_{i=1}^{N} \frac{g(\mathbf{x}_i\hat{\theta})^2}{G(\mathbf{x}_i\hat{\theta})[1 - G(\mathbf{x}_i\hat{\theta})]} \mathbf{x}_i'\mathbf{x}_i \right]^{-1} .$$

Probit

The probit model chooses $G(\mathbf{x}_i\boldsymbol{\theta}) = \Phi(\mathbf{x}_i\boldsymbol{\theta})$ and $g(\mathbf{x}_i\boldsymbol{\theta}) = \phi(\mathbf{x}_i\boldsymbol{\theta})$.

This yields the FOC

$$\sum_{i=1}^{N} \mathbf{s}_i(\hat{\boldsymbol{\theta}}) = \sum_{i=1}^{N} \frac{\phi(\mathbf{x}_i\hat{\boldsymbol{\theta}})}{\Phi(\mathbf{x}_i\hat{\boldsymbol{\theta}})[1-\Phi(\mathbf{x}_i\hat{\boldsymbol{\theta}})]}\, \mathbf{x}_i'[y_i - \Phi(\mathbf{x}_i\hat{\boldsymbol{\theta}})] = \mathbf{0}$$

and the estimated variance

$$\widehat{\mathrm{Avar}}(\hat{\boldsymbol{\theta}}) = \left[\sum_{i=1}^{N} \frac{\phi(\mathbf{x}_i\hat{\boldsymbol{\theta}})^2}{\Phi(\mathbf{x}_i\hat{\boldsymbol{\theta}})[1-\Phi(\mathbf{x}_i\hat{\boldsymbol{\theta}})]}\mathbf{x}_i'\mathbf{x}_i\right]^{-1}.$$

Logit

The logit model chooses $G(\mathbf{x}_i\boldsymbol{\theta}) = \Lambda(\mathbf{x}_i\boldsymbol{\theta})$ and $g(\mathbf{x}_i\boldsymbol{\theta}) = \lambda(\mathbf{x}_i\boldsymbol{\theta})$.

Note that $\Lambda(z) = \exp(z)/[1+\exp(z)]$ and thus $1-\Lambda(z) = 1/[1+\exp(z)]$.

The first derivative is

$$\lambda(z) = \frac{\exp(z)[1+\exp(z)] - \exp(z)\exp(z)}{[1+\exp(z)]^2} = \frac{\exp(z)}{[1+\exp(z)]^2} = \Lambda(z)[1-\Lambda(z)].$$

This simplifies both the FOC

$$\sum_{i=1}^{N} \mathbf{s}_i(\hat{\boldsymbol{\theta}}) = \sum_{i=1}^{N} \frac{\lambda(\mathbf{x}_i\hat{\boldsymbol{\theta}})}{\Lambda(\mathbf{x}_i\hat{\boldsymbol{\theta}})[1-\Lambda(\mathbf{x}_i\hat{\boldsymbol{\theta}})]}\,\mathbf{x}_i'[y_i - \Lambda(\mathbf{x}_i\hat{\boldsymbol{\theta}})] = \sum_{i=1}^{N} \mathbf{x}_i'[y_i - \Lambda(\mathbf{x}_i\hat{\boldsymbol{\theta}})] = \mathbf{0}$$

and the estimated variance

$$\widehat{\text{Avar}}(\hat{\boldsymbol{\theta}}) = \left[\sum_{i=1}^{N} \frac{\lambda(\mathbf{x}_i\hat{\boldsymbol{\theta}})^2}{\Lambda(\mathbf{x}_i\hat{\boldsymbol{\theta}})[1-\Lambda(\mathbf{x}_i\hat{\boldsymbol{\theta}})]}\mathbf{x}_i'\mathbf{x}_i\right]^{-1} = \left[\sum_{i=1}^{N} \lambda(\mathbf{x}_i\hat{\boldsymbol{\theta}})\mathbf{x}_i'\mathbf{x}_i\right]^{-1}.$$

6. Interpretation of Binary Choice Models

Partial effects for continuous variables

Recall that the the conditional mean is a nonlinear function of $\mathbf{x}_i$:

$$\mathrm{E}(y_i|\mathbf{x}_i) = \mathrm{P}(y_i = 1|\mathbf{x}_i) = G(\mathbf{x}_i\boldsymbol{\theta}) = \begin{cases} \Phi(\mathbf{x}_i\boldsymbol{\theta}) & \text{for probit} \\ \Lambda(\mathbf{x}_i\boldsymbol{\theta}) & \text{for logit} \end{cases}$$

Hence, the parameters $\boldsymbol{\theta}$ do *not* measure the partial effects.

Instead, the partial effect of a continuous variable, say, $x_{i,k}$,

$$\frac{\partial\,\mathrm{E}(y_i|\mathbf{x}_i)}{\partial x_{i,k}} = \frac{\partial G(\mathbf{x}_i\boldsymbol{\theta})}{\partial x_{i,k}} = g(\mathbf{x}_i\boldsymbol{\theta})\theta_k,$$

is a nonlinear function of *all* parameters and *all* regressors. This means that the impact of a change in a single variable on the “success” probability depends on the *entire characteristics* of the individual and thus typically differs between individuals.

At least, $g(\mathbf{x}_i\boldsymbol{\theta}) > 0$ such that the *sign* of the partial effect is given by the sign of θ_k.

The partial effect $g(\mathbf{x}_i\boldsymbol{\theta})\theta_k$ is maximal at the mode of $g(\cdot)$ and becomes gradually smaller towards the tails.

This ensures that all predicted probabilities lie in the 0-1 interval.

In the probit model, the maximum effect is at $\mathbf{x}_i\boldsymbol{\theta} = 0$ which yields

$$\frac{\partial \operatorname{E}(y_i|\mathbf{x}_i)}{\partial x_{i,k}} = \frac{\partial \operatorname{P}(y_i = 1|\mathbf{x}_i)}{\partial x_{i,k}} = \phi(0)\theta_k \approx 0.4\ \theta_k.$$

In the logit model, the maximum effect is again at $\mathbf{x}_i\boldsymbol{\theta} = 0$ which yields

$$\frac{\partial \operatorname{E}(y_i|\mathbf{x}_i)}{\partial x_{i,k}} = \frac{\partial \operatorname{P}(y_i = 1|\mathbf{x}_i)}{\partial x_{i,k}} = \lambda(0)\theta_k = \frac{\exp(0)}{[1+\exp(0)]^2}\ \theta_k = 0.25\ \theta_k.$$

Note that relative effects do not depend on **x**: for continuous variables x_k and x_h, the ratio of the partial effects is constant across individuals:

$$\frac{\partial \operatorname{E}(y_i|\mathbf{x}_i)/\partial x_{i,k}}{\partial \operatorname{E}(y_i|\mathbf{x}_i)/\partial x_{i,h}} = \frac{\partial \operatorname{P}(y_i = 1|\mathbf{x}_i)/\partial x_{i,k}}{\partial \operatorname{P}(y_i = 1|\mathbf{x}_i)/\partial x_{i,h}} = \theta_k/\theta_h.$$

Partial effects for discrete variables

For discrete explanatory variables such as 0-1 dummies, first derivatives are not defined.

Suppose you are interested in the partial effect of a dummy variable, say, $x_{i,P}$, i.e., in the effect a change in $x_{i,P}$ from 0 to 1 has on $\mathrm{P}(y_i = 1|\mathbf{x}_i)$.

This effect is computed as follows:

$$\Delta P_i = \mathrm{P}(y_i = 1|x_{i,P} = 1) - \mathrm{P}(y_i = 1|x_{i,P} = 0),$$

where the $x_{i,1}, \ldots, x_{i,P-1}$ are as observed. The probabilities are computed using the link function $G(\cdot)$:

$$\Delta P_i = G([x_{i,1}, \ldots, x_{i,P-1}, 1]\boldsymbol{\theta}) - G([x_{i,1}, \ldots, x_{i,P-1}, 0]\boldsymbol{\theta}).$$

Averaging partial effects

In practice, we like to report a kind of average effect because we are typically not interested in the partial effects of any particular individual.

To this end, we can evaluate the partial effect at the sample means $\bar{\mathbf{x}}$ of the data (partial effect of the average, PEA) or calculate the mean of the partial effects for every observation (average partial effect, APE).

In large samples these generally yield similar results but this does not need to hold in small and moderate-sized samples.

For continuous explanatory variables, this yields in population

$$PEA = g\left(\mathrm{E}[\mathbf{x}_i]\boldsymbol{\theta}\right)\theta_k$$

$$APE = \mathrm{E}\left[g(\mathbf{x}_i\boldsymbol{\theta})\right]\theta_k$$

The sample analogues are given by

$$\widehat{PEA} = g\left(\bar{\mathbf{x}}\hat{\boldsymbol{\theta}}\right)\hat{\theta}_k$$

$$\widehat{APE} = N^{-1}\sum_{i=1}^{N} g(\mathbf{x}_i\hat{\boldsymbol{\theta}})\hat{\theta}_k$$

For discrete explanatory variables, we compute

$$\widehat{PEA} = G([\bar{x}_1, \ldots, \bar{x}_{P-1}, 1]\hat{\boldsymbol{\theta}}) - G([\bar{x}_1, \ldots, \bar{x}_{P-1}, 0]\hat{\boldsymbol{\theta}})$$

$$\widehat{APE} = N^{-1} \sum_{i=1}^{N} \left[G([x_{i,1}, \ldots, x_{i,P-1}, 1]\hat{\boldsymbol{\theta}}) - G([x_{i,1}, \ldots, x_{i,P-1}, 0]\hat{\boldsymbol{\theta}}) \right]$$

7. Example: Labor Force Participation

Economic question and variables

- Recall: We want to explain the decision whether married women participate in the labor market ($inlf = 1$) or not ($inlf = 0$).
- The underlying decision process will supposedly depend on
  (a) the wage offer by potential employer—probably based on expected marginal product
  (b) the reservation wage of the individual women
- Suppose the offered wage depends on education ($educ$) and quadratically on experience ($exper$).
- Suppose the reservation wage depends on age (age), the number of young kids ($kidslt6$), the number of older kids ($kidsge6$) and nonwife income in thousands of dollars ($nwifeinc$).
- The sample comprises $N = 753$ observations and is assumed to be a random sample.

Stata program

```
use mroz.dta
summarize inlf nwifeinc edu exper age
tab kidslt6
tab kidsge6
regress inlf nwifeinc edu exper expersq age kidslt6 kidsge6, robust
logit inlf nwifeinc edu exper expersq age kidslt6 kidsge6, vce(oim)
probit inlf nwifeinc edu exper expersq age kidslt6 kidsge6, vce(oim)
```

Summary statistics

```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
        inlf |        753    .5683931    .4956295          0          1
    nwifeinc |        753    20.12896     11.6348   -.0290575         96
        educ |        753    12.28685    2.280246          5         17
       exper |        753    10.63081     8.06913          0         45
         age |        753    42.53785    8.072574         30         60
```

Summary statistics

```
 # kids < 6 |
      years |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        606       80.48       80.48
          1 |        118       15.67       96.15
          2 |         26        3.45       99.60
          3 |          3        0.40      100.00
------------+-----------------------------------

      Total |        753      100.00
```

Summary statistics

```
# kids 6-18 |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        258       34.26       34.26
          1 |        185       24.57       58.83
          2 |        162       21.51       80.35
          3 |        103       13.68       94.02
          4 |         30        3.98       98.01
          5 |         12        1.59       99.60
          6 |          1        0.13       99.73
          7 |          1        0.13       99.87
          8 |          1        0.13      100.00
------------+-----------------------------------
      Total |        753      100.00
```

Estimation results (Wooldridge, p. 580)

Independent Variable	LPM (OLS)	Logit (MLE)	Probit (MLE)
nwifeinc	−.0034 (.0015)	−.021 (.008)	−.012 (.005)
educ	.038 (.007)	.221 (.043)	.131 (.025)
exper	.039 (.006)	.206 (.032)	.123 (.019)
$exper^2$	−.00060 (.00019)	−.0032 (.0010)	−.0019 (.0006)
age	−.016 (.002)	−.088 (.015)	−.053 (.008)
kidslt6	−.262 (.032)	−1.443 (0.204)	−.868 (.119)
kidsge6	.013 (.013)	.060 (.075)	.036 (.043)
constant	.586 (.151)	.425 (.860)	.270 (.509)
Number of observations	753	753	753
Percent correctly predicted	73.4	73.6	73.4
Log-likelihood value	—	−401.77	−401.30
Pseudo R-squared	.264	.220	.221

Average partial effects

	OLS	logit	probit
nwifeinc	−.0034 (.002)	−.0038 (.001)	−.0036 (.001)
educ	.0380 (.007)	.0395 (.007)	.0394 (.007)
exper	.0395 (.006)	.0368 (.005)	.0371 (.005)
expersq	−.0006 (.000)	−.0006 (.000)	−.0006 (.000)
age	−.0161 (.002)	−.0157 (.002)	−.0159 (.002)
kidslt6	−.2618 (.032)	−.2578 (.032)	−.2612 (.032)
kidsge6	.0130 (.013)	.0107 (.013)	.0108 (.013)

Note: estimated standard errors in brackets below the estimates.

Do you think the APEs for *exper* and *expersq* make sense?

To get the results, we ran the Stata command:

```
margins, dydx(nwifeinc edu exper expersq age kidslt6 kidsge6)
```

Note: we have not told Stata that *kidslt*6 and *kidsge*6 are discrete variables. Hence, Stata calculates APE's assuming they are continuous.

What can we do to assess the average partial effect of having one more child? We may compare the "success" probability of (a) the sample as it is with (b) adding one more child to each household.

Assuming this is the last variable $x_{i,P}$, we compute

$$\widehat{APE} = N^{-1} \sum_{i=1}^{N} [G([x_{i,1}, \dots, x_{i,P-1}, x_{i,P} + 1]\boldsymbol{\theta}) - G([x_{i,1}, \dots, x_{i,P-1}, x_{i,P}]\boldsymbol{\theta})]$$

Here are the average partial effects on the probability to participate in the labor market of adding young kids to the household based on the logit model:

		before	after	APE
one more kid	under continuity			$-.2578$
	discrete change	.568	.311	$-.2576$
two more kids	discrete change	.568	.122	$-.4463$
three more kids	discrete change	.568	.036	$-.5324$

APE of going from zero to one young child

Difference between household types

Here we use the effect calculated under the assumption that *kidslt6* is continuous. In all cases: age = 30.

	no experience	half experience	full experience
high income, high education	$-.303$	$-.211$	$-.137$
median income, median education	$-.361$	$-.259$	$-.143$
low income, low education	$-.357$	$-.272$	$-.144$

Notes:

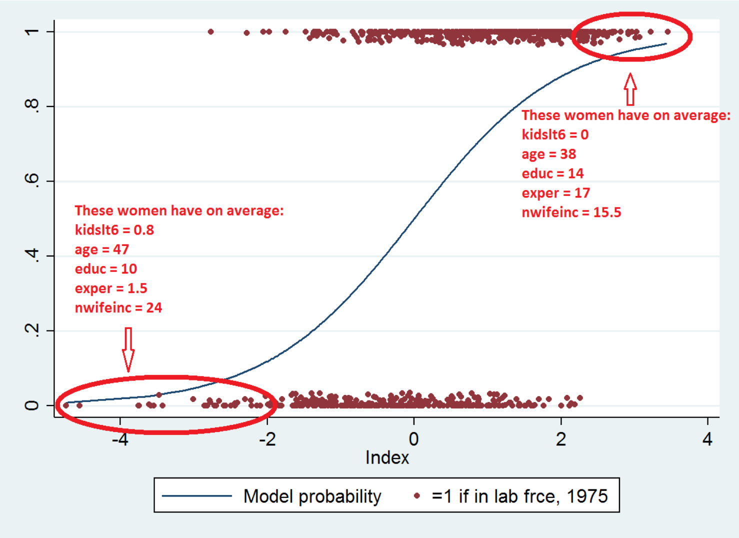
High income = 32.7K dollars (90% quantile), median income = 17.7K dollars (50% quantile), low income = 9K dollars (10% quantile)

High education = 16 years (90% quantile), median education = 12 years (50% quantile), low education = 10 years (10% quantile)

No experience = 0 years; half experience = (30 years - 6 year - years of education)/2; full experience = (30 years - 6 year - years of education)

Data and model prediction

Index $= x\theta$



These women have on average:
kidslt6 = 0.8
age = 47
educ = 10
exper = 1.5
nwifeinc = 24

These women have on average:
kidslt6 = 0
age = 38
educ = 14
exper = 17
nwifeinc = 15.5