Microeconometrics

1. Regressions and Maximum Likelihood estimation

Joachim Winter

Department of Economics
University of Munich

Summer Semester 2019

Please report typos and errors to winter@lmu.de.

Overview

1.0 Notation and data structures

1.1 A general perspective on regressions
(including a review of OLS estimation)

1.2 Nonlinear models and economic choices

1.3 The likelihood function

1.4 Preview: Some examples of conditional ML models

1.5 Computation of the ML estimator

1.6 Consistency and asymptotic distribution of ML

1.7 Discussion

Textbook: C&T, sections 4.2, 4.4, 5.6; Wooldridge, sections 13.3–4

1.0 Notation and data structures

The course follows (mostly) the notation of C&T (section 1.6), which differs slightly from Wooldridge's notation.

- Vectors are always defined as column vectors and denoted by bold symbols. For the linear model with K regressors, we have

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_K \end{pmatrix} \quad \text{and} \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_K \end{pmatrix} .$$

- The scalar dependent variable is y.
- Individual observations are indexed by i. The sample size is N. The random sample is $\{(y_i, \boldsymbol{x}_i), i = 1, \ldots, N\}$.

1.0 Notation and data structures

- Parameters are denoted by greek letters.
- In general, regression models have more parameters than the K elements of $\boldsymbol{\beta}$ in the linear case.
- The general notation for parameters is $\boldsymbol{\theta}$ which is a $q \times 1$ vector.
- Estimators of parameters are functions of the data; we denote them by hats.
- The scalar error term is denoted by u.
- Matrices are denoted by uppercase bold letters.
- We consider mostly cross-sectional data, so the time index is typically suppressed.

1.0 Notation and data structures

- Stacking the N observations yields

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix} \quad \text{and} \quad \boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1' \\ \vdots \\ \boldsymbol{x}_N' \end{pmatrix} .$$

  where $\boldsymbol{y}$ is $N \times 1$ and $\boldsymbol{X}$ is $N \times K$.

- The linear regression model can be written in three ways:

$$\begin{aligned} y &= \boldsymbol{x}'\boldsymbol{\beta} + u \quad \text{or} \\ y_i &= \boldsymbol{x}_i'\boldsymbol{\beta} + u_i \,, \; i = 1, \ldots, N \quad \text{or} \\ \boldsymbol{y} &= \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u} \end{aligned}$$

1.0 Notation and data structures

- The familiar OLS estimator for the linear regression model can be written as

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \left(\sum_{i=1}^{N}\boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}\sum_{i=1}^{N}\boldsymbol{x}_i y_i\,.$$

1.0 Notation and data structures

- In microeconometrics, we typically use cross-sectional data on individuals, households, firms, . . . ("units of observation")
- Most of the time, we assume that we have a large, random sample from the relevant population
- $(y_i, \boldsymbol{x}_i)$ are independently and identically distributed (i.i.d.) random variables
- More general cases:
  - Non-random samples – touched upon in chapter 6
  - Panel data (repeated observations for the units)
  - Various reasons for dependence of observations

1.1 A general perspective on regressions

- In econometrics courses, we often motivate regression as a conditional expectation of the dependent variable given a set of explanatory variables (regressors)

$$\mathrm{E}(y|\boldsymbol{x})\,.$$

- A more general view is this (Manski, 1991, p. 34):

  A regression of y on $\boldsymbol{x}$ is any feature of the probability distribution of y conditional on $\boldsymbol{x}$, considered as a function of $\boldsymbol{x}$.

- Manski further summarizes the historical view:

  The regression of a random variable y on another such variable $\boldsymbol{x}$ was understood to be the mean of y conditional on $\boldsymbol{x}$, considered as a function of $\boldsymbol{x}$.

1.1 A general perspective on regressions

Best prediction and loss functions

- From the more general perspective, regressions can be motivated by a best prediction problem.
- Interpreting a regression as a best prediction does not imply that it has a causal interpretation. We need to think about causal interpretations separately (later).
- A best prediction problem requires the specification of a loss function:
  - $\widehat{y}$ is a predictor of y (a function of $\boldsymbol{x}$)
  - $e = y - \widehat{y}$ is the prediction error
  - $\mathrm{L}(e) = \mathrm{L}(y - \widehat{y})$ is the loss function

1.1 A general perspective on regressions

- Since y and $\boldsymbol{x}$ are random variables, the decision maker minimizes the **expected loss**,
$$\mathrm{E}[\mathrm{L}((y - \widehat{y})|\boldsymbol{x})] \,.$$
- There are many different loss functions (C&T, Table 4.1).
- In econometrics, we typically use squared error loss, $\mathrm{L}(e) = e^2$.
- Another important example is absolute error loss, $\mathrm{L}(e) = |e|$, and asymmetric version of absolute error loss, which lead to median and quantile regression (not covered in this course, see C&T, section 4.6).

1.1 A general perspective on regressions

- The next step is to think about optimal prediction given some loss function:
$$\min_{\widehat{y}} \mathrm{E}[L(y - \widehat{y})|\boldsymbol{x}]$$
- Note that we're minimizing a criterion to find a function (since $\widehat{y}$ is a function of $\boldsymbol{x}$). The maths behind this is abstract.
- It turns out that under squared error loss, the optimal predictor is the conditional mean, $\mathrm{E}(y|\boldsymbol{x})$.
- That's why we're obsessed with conditional expectations.
- Again, a conditional expectation is a best predictor of y given $\boldsymbol{x}$, i. e., a function of the data (a "statistic"). But which function?

1.1 A general perspective on regressions

- We can try to estimate the function $\mathrm{E}(y|\boldsymbol{x})$ nonparametrically – which is easy but requires that we have
  - lots of data,
  - not too many regressors (elements of $\boldsymbol{x}$).
- If we have lots of data and lots of regressors, machine learning methods can be used within the framework classical econometrics.
- For many reasons (historical, ease of computation and interpretation), applied econometricians tend to use parametric models, so

$$\mathrm{E}(y|\boldsymbol{x}) = g(\boldsymbol{x}, \boldsymbol{\beta}) \,,$$

  where g is a function of $\boldsymbol{x}$ known up to a finite number of parameters collected in the vector $\boldsymbol{\beta}$.

1.1 A general perspective on regressions

- Most prominently, $g(\boldsymbol{x}, \boldsymbol{\beta}) = \boldsymbol{x}'\boldsymbol{\beta}$ for linear regression. Then we can estimate the parameters $\boldsymbol{\beta}$ by OLS.
- If the true conditional mean (the "data generating process") is indeed linear, then we estimate $\mathrm{E}(y|\boldsymbol{x})$ consistently by OLS, and the elements of $\boldsymbol{\beta}$ have a causal interpretation,

$$\frac{\partial \mathrm{E}(y|\boldsymbol{x})}{\partial x_k} = \beta_k \,,\; k = 1, \ldots, K.$$

- The key assumption is $\mathrm{E}(\boldsymbol{x}_i u_i) = \mathbf{0}$ (or stronger, $\mathrm{E}(u|\boldsymbol{x}) = 0$).
- If the true conditional mean is not linear in $\boldsymbol{x}$, we can still estimate $\boldsymbol{\beta}$ by OLS and interpret it as a best linear prediction under squared error loss.

1.1 A general perspective on regressions

The linear model and the OLS estimator

- The linear regression model for a typical observation i is

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + u_i \,.$$

- In matrix notation, we stack the N observations, so

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u} \,.$$

- The OLS estimator is defined to minimize the sum of squared errors

$$\sum_{i=1}^{N} u_i^2 = \boldsymbol{u}'\boldsymbol{u} = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \,.$$

1.1 A general perspective on regressions

- Solving the minimization problem yields

$$\widehat{\boldsymbol{\beta}}_{\mathsf{OLS}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\,\boldsymbol{X}'\boldsymbol{y}\,,$$

  assuming that $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ exists, i. e., $\boldsymbol{X}$ must have full column rank (no multicollinearity).

- Again, this estimator can always be given the interpretation of a best linear predictor under squared error loss.

1.1 A general perspective on regressions

Identification

- When we say that a parameter is identified, we (intuitively) mean that it can be written as a function of observable variables.
- More formally: A structure (i. e., a vector of parameters), $\boldsymbol{\theta}^0$, is identified if there is no other observationally equivalent structure in the space of admissible parameter values, $\boldsymbol{\Theta}$ (i. e., if there are no other parameters that generate the same observable data).
- See C&T, section 2.5, for details and a formal definition of observational equivalence.
- The notation $\boldsymbol{\theta}^0$ is used whenever we want to indicate explicitly the (unknown) true parameter vector in the population.

1.1 A general perspective on regressions

- In the context of OLS, identification requires two conditions:

1. $\mathrm{E}(\boldsymbol{y}|\boldsymbol{X}) = \boldsymbol{X}\boldsymbol{\beta}$

   This condition says that the conditional mean is correctly specified.

2. $\boldsymbol{X}\boldsymbol{\beta}^{(1)} = \boldsymbol{X}\boldsymbol{\beta}^{(2)} \iff \boldsymbol{\beta}^{(1)} = \boldsymbol{\beta}^{(2)}$

   This condition is equivalent to $\boldsymbol{X}'\boldsymbol{X}$ being nonsingular.

- These identification conditions, together with some additional conditions required for statistical estimation, appear in similar form as part of the standard OLS assumptions.

1.1 A general perspective on regressions

Consistency and asymptotic distribution of the OLS estimator

- To show consistency and to derive the asymptotic distribution of the OLS estimator, we need to make some assumptions. Unfortunately, C&T's treatment can be confusing:
  - Proposition 4.1 (p. 73) uses assumptions that are more general than we really need at this point.
  - Section 4.4.6 provides a more specialized version that is sufficient in the cross-sectional case.
  - They key difference is whether we allow for dependence between observations.
- Those familiar with the assumptions used by Wooldridge can stick to those (but note that he assumes random sampling).

1.1 A general perspective on regressions

OLS assumptions for cross-sectional data

In the following, I leave out technical conditions, but see C&T, pp. 76–77 for discussion.

1. The data $(y_i, \boldsymbol{x}_i)$ are independent but not necessarily identically distributed over i. (This version allows for nonrandom, e. g.stratified, sampling.)

2. The model is correctly specified so that

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + u_i \,.$$

1.1 A general perspective on regressions

3. The regressor vector $\boldsymbol{x}_i$ is possibly stochastic with finite second moment. The matrix $\boldsymbol{X}$ has full column rank K in the sample being analyzed.
4. The errors have zero mean, conditional on regressors,

$$\mathrm{E}[u_i|\boldsymbol{x}_i] = 0\,.$$

5. The errors are heteroskedastic, conditional on regressors:

$$\sigma_i^2 = \mathrm{E}[u_i^2|\boldsymbol{x}_i] \text{ and}$$

$$\boldsymbol{\Omega} = \mathrm{E}[\boldsymbol{u}\boldsymbol{u}'|\boldsymbol{X}] = \mathrm{Diag}[\sigma_i^2]$$

where $\boldsymbol{\Omega}$ is an $N \times N$ positive definite matrix.

1.1 A general perspective on regressions

Unbiasedness and consistency

- Under the conditions stated above (and under the additional technical conditions stated in C&T):

- The OLS estimator is unbiased:

$$\mathrm{E}[\widehat{\boldsymbol{\beta}}_{\mathsf{OLS}}] = \boldsymbol{\beta}\,.$$

- The OLS estimator is consistent:

$$\operatorname{plim} \widehat{\boldsymbol{\beta}}_{\mathsf{OLS}} = \boldsymbol{\beta} \quad \text{or} \quad \widehat{\boldsymbol{\beta}}_{\mathsf{OLS}} \xrightarrow{p} \boldsymbol{\beta}$$

1.1 A general perspective on regressions

Asymptotic distribution

- The asymptotic distribution of the OLS estimator can be written as
$$\widehat{\boldsymbol{\beta}}_{\mathsf{OLS}} \overset{a}{\sim} \mathcal{N}(\boldsymbol{\beta}, (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Omega}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}) \,.$$
- The (asymptotic) variance matrix is given by
$$\mathrm{V}(\widehat{\boldsymbol{\beta}}_{\mathsf{OLS}}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Omega}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \,.$$
- To obtain robust standard errors, we use this formula to estimate the variance matrix:
$$\widehat{\mathrm{V}}(\widehat{\boldsymbol{\beta}}_{\mathsf{OLS}}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\,\mathrm{Diag}[\widehat{u}_i^2]\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

1.1 A general perspective on regressions

- Note that this result allows for heteroskedasticity via our definition of $\boldsymbol{\Omega}$.
- Under homoskedasticity, $\boldsymbol{\Omega} = \sigma^2 \boldsymbol{I}$, and $\widehat{\mathrm{V}}(\widehat{\boldsymbol{\beta}}_{\mathsf{OLS}})$ simplifies.

1.1 A general perspective on regressions

Endogeneity and instrumental variables

- A cornerstone of modern microeconometrics are instrumental variables (IV) estimators that address endogeneity, i. e., situations in which $\mathrm{E}(u_i|\boldsymbol{x}_i) \neq 0$ (e. g., due to omitted variables). Since these topics are covered extensively in core undergraduate econometrics courses, we won't go into details here.
- If you're interested, see C&T, sections 4.7.3, 4.7.4, and 4.8.
- In most of our following discussion of nonlinear models, we will ignore potential endogeneity problems – but there is no reason to believe that they are less common than in linear models. However, they are more difficult to deal with in nonlinear models.

1.1 A general perspective on regressions

Individual heterogeneity

- Consider the standard linear regression model,

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + u_i \,.$$

- Individuals and firms are heterogenous, so simple models like the linear regression model are unrealistic.
- One source of **unobserved heterogeneity**, mentioned above, are omitted variables. This is dealt with using IV or panel techniques.
- Individuals can also be heterogenous with respect to their marginal effects. There is increasing interest in models that allow the parameters to have a distribution in the population (instead of being fixed, as in standard models).

1.1 A general perspective on regressions

- Consider again the linear model, but now with heterogenous parameters:

$$y_i = \boldsymbol{x}_i' \boldsymbol{\beta}_i + u_i \,.$$

- If we use OLS, the model that we effectively estimate becomes

$$y_i = \boldsymbol{x}_i' \boldsymbol{\beta} + (u_i + \boldsymbol{x}_i'(\boldsymbol{\beta}_i - \boldsymbol{\beta})) \,.$$

- An OLS regression of y_i on $\boldsymbol{x}_i$ is consistent for $\boldsymbol{\beta}$ (the mean of $\boldsymbol{\beta}_i$) if the $\boldsymbol{\beta}_i$ and the $\boldsymbol{x}_i$ are independent. (Why?)
- If the $\boldsymbol{\beta}_i$ and the $\boldsymbol{x}_i$ are not independent, panel methods may still allow consistent estimation of the mean $\boldsymbol{\beta}$ in a linear model even if the $\boldsymbol{\beta}_i$ are heterogenous.

1.1 A general perspective on regressions

- In nonlinear models, similar results typically don't hold. Therefore, **random parameter models** or **mixture models** which specify the distribution of the β_i's explicitly are now commonly used.
- We will discuss such models at the end of the semester (if we have time).

1.1 A general perspective on regressions

- The situation becomes even more complicated when the linear index and separability assumptions are given up. A general model would be

$$y = \phi(\boldsymbol{x}, \boldsymbol{a}) \,,$$

  where $\boldsymbol{x}$ is a vector of observed covariates while the possibly infinite-dimensional vector $\boldsymbol{a}$ is unobserved. ϕ is a fixed function defined on $\mathcal{R}^K \times \mathcal{A}$.

- Such models allow for arbitrary forms of heterogeneity since $\boldsymbol{a}$ can be of infinite dimension.

1.1 A general perspective on regressions

- One object of interest is the **local average structural derivative** with respect to some element of $\boldsymbol{x}$, evaluated at $\boldsymbol{x}^*$

$$\mathrm{E}[\partial_{x_k}\phi(\boldsymbol{x},\boldsymbol{a})|\boldsymbol{x}=\boldsymbol{x}^\star]\,.$$

- Note that in the special case of the standard linear regression model, this derivative would simply be β_k (and thus a constant).
- Such nonparametric models are commonly analyzed in theoretical econometrics, but rarely used in practice.
- We won't discuss such models in this course, but see, for example, the survey by Rosa Matzkin in the *Handbook of Econometrics* (chapter 73).

1.1 A general perspective on regressions

M-estimators

- The OLS estimator minimizes the sum of the squared residuals (which is a function of the data).
- The value that an estimator takes is the function of the data in a specific sample.
- General case: An estimator that maximizes (or minimizes) some function of the data is called an **M-estimator**.
- Objective functions can be motivated in various ways (as in OLS, ML, . . .)

1.1 A general perspective on regressions

- Unknown population parameter: $\boldsymbol{\theta}_0$ ($q \times 1$ vector)
- The M-estimator maximizes an objective function:

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\arg\max}\ Q_N(\boldsymbol{\theta}) \quad \text{with}$$

$$Q_N(\boldsymbol{\theta}) = \frac{1}{N} \sum_{i=1}^{N} q(y_i, \boldsymbol{x}_i, \boldsymbol{\theta})$$

where $q(\cdot)$ is the contribution of each observation to the objective function (e. g., the squared residual).

1.1 A general perspective on regressions

- The value of the estimator for a given sample can be computed using the FOC:

$$\frac{1}{N}\sum_{i=1}^{N}\left.\frac{\partial q(y_i, \boldsymbol{x}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\widehat{\boldsymbol{\theta}}} = 0$$

- This is a system of q equations in q unknowns that generally has no explicit solution for $\widehat{\boldsymbol{\theta}}$, so we need to solve it numerically.
- Examples:
  - OLS, WLS
  - GMM, 2SLS/IV
  - Maximum likelihood

1.2 Nonlinear models and economic choices

- Consider the regression model

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + u_i$$

- What happens if y is not continuous and we use OLS?
- OLS will still give the best linear prediction under square error loss, but this may not be appropriate.
- Examples: $y \in \{0, 1\}$ or $y = \max[y^*, 0]$
- See the Stata code used in the lecture.

1.2 Nonlinear models and economic choices

- Most outcomes resulting from economics choice are not continuous and unbounded random variables, but
  - **discrete** (binary, multinomial, etc.) or
  - continuous but **limited** in the data we observe (because of censoring, selection, etc.).
- Thus, we need nonlinear models.

1.3 The likelihood function

Introduction and notation

- First, we consider unconditional maximum likelihood estimation (i. e., we do not have any explanatory variables).

- Let $\{y_1, y_2, \ldots, y_N\}$ be a random sample drawn from the population distribution $f(y|\boldsymbol{\theta})$.

- Thus, the observations, $i = 1, \ldots, N$, are i.i.d.

- The distribution $f(y|\boldsymbol{\theta})$ can be discrete or continuous.

1.3 The likelihood function

- Because of the i.i.d. sampling process, the likelihood function – the joint density of the data in the sample – can be written as the product of the individual densities,

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{N} f(y_i|\boldsymbol{\theta}) \,.$$

- Typically, we work with the log likelihood function, where $\ell_i(\boldsymbol{\theta}) = \ln f(y_i|\boldsymbol{\theta})$ is the log likelihood contribution of observation i,

$$\mathcal{L}(\boldsymbol{\theta}) = \ln L(\boldsymbol{\theta}) = \ln \prod_{i=1}^{N} f(y_i|\boldsymbol{\theta}) = \sum_{i=1}^{N} \ell_i(\boldsymbol{\theta}) \,.$$

1.3 The likelihood function

- The maximum likelihood estimator of $\boldsymbol{\theta}$ is given by

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\arg\max}\ \mathcal{L}(\boldsymbol{\theta}) = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\arg\max} \sum_{i=1}^{N} \ell_i(\boldsymbol{\theta})\,.$$

- This is a special case of an M-estimator; the objective function is the sum of the log likelihood contributions.

- The log transformation is a monotone transformation and thus the objective function $\mathcal{L}(\boldsymbol{\theta})$ has the same maximum as the product of the likelihood contributions, $L(\boldsymbol{\theta})$.

1.3 The likelihood function

Example 1: An unconditional maximum likelihood estimator

- Imagine a large urn, filled with red and blue balls.
- We want to estimate the fraction of red balls in this urn, denoted by p.
- Thus, the parameter vector $\boldsymbol{\theta}$ has just one element, p.
- We draw a random sample of N balls from this urn.
- In other words, our sample consists of N realizations of a binary random variable.
- Denote by N_1 the number of red balls in our sample of size N.

1.3 The likelihood function

- Since we have an i.i.d. sampling process, this sample can be characterized by a single number – the number of red balls, N_1.
- Such numbers (statistics) that fully characterize a sample are called **sufficient statistics**.
- The ML estimator of p can be derived by maximizing the probability of observing exactly N_1 red balls and $N - N_1$ blue balls in a random sample of size N.
- This probability is proportional to $p^{N_1} \cdot (1-p)^{N-N_1}$.
- Note that features of our sample enter only via the sample size, N, and the sufficient statistic, N_1.

1.3 The likelihood function

- The log likelihood function is, ignoring a constant that does not involve p,

$$\mathcal{L}(p) = N_1 \ln(p) + (N - N_1) \ln(1 - p) \,.$$

- Maximizing this log likelihood function with respect to $p \in [0, 1]$ yields the first-order condition $\frac{N_1}{p} - \frac{N - N_1}{1 - p} = 0$ which can be solved for p to yield the maximum likelihood estimator

$$\widehat{p} = \frac{N_1}{N} \,.$$

- Obviously, we do not have any explanatory variables here, so this is an unconditional maximum likelihood estimator.

1.3 The likelihood function

Conditional maximum likelihood estimators

- In econometrics, we are typically interested in situations in which we have explanatory variables, and therefore we use a **conditional maximum likelihood estimator** (CMLE). Most of the time, we simply call this the maximum likelihood (ML) estimator.

- Regressions are specified conditional models by writing $\mathrm{E}(y|\boldsymbol{x}) = m(\boldsymbol{x}, \boldsymbol{\theta})$. Typically, $m(\cdot)$ is a linear function (in $\boldsymbol{\theta}$).

- Now, we specify a conditional density $f(y|\boldsymbol{x}, \boldsymbol{\theta})$, where $\boldsymbol{x}$ is again a $1 \times K$ vector of explanatory variables and $\boldsymbol{\theta}$ is a $P \times 1$ vector of parameters with $P \geq K$.

1.3 The likelihood function

- Under an i.i.d. sampling assumption, and using similar notation as before, we can write the sample likelihood function as

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{N} f(y_i|\boldsymbol{x}_i, \boldsymbol{\theta}) \,.$$

- The sample log likelihood function is

$$\mathcal{L}(\boldsymbol{\theta}) = \ln L(\boldsymbol{\theta}) = \sum_{i=1}^{N} \ell_i(\boldsymbol{\theta}) \,.$$

  where $\ell_i(\boldsymbol{\theta}) = \ln f(y_i|\boldsymbol{x}_i, \boldsymbol{\theta})$ is the log likelihood contribution of observation i.

1.3 The likelihood function

- The (conditional) maximum likelihood estimator maximizes the log likelihood function,

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\arg\max} \sum_{i=1}^{N} \ell_i(\boldsymbol{\theta}) \,.$$

- To prove the consistency and other asymptotic properties of this estimator, one needs to show that this estimator is a special case of an M estimator so that the earlier consistency result applies. See Wooldridge's sections 13.3 and 13.4 for details.
- Next, we present two simple examples of models that can be estimated by ML.

1.3 The likelihood function

Example 2: The linear regression model

- Consider a linear regression model,

$$y = \boldsymbol{x}'\boldsymbol{\beta} + u\,.$$

- We assume that $u|\boldsymbol{x} \sim \mathrm{N}(0, \sigma^2)$, which is stronger than our usual assumption $\mathrm{E}(u|\boldsymbol{x}) = 0$.

- The parameter vector is given by $\boldsymbol{\theta} = (\boldsymbol{\beta}', \sigma^2)'$. The number of parameters is $P = K + 1$.

- This model is called the **normal regression model**.

1.3 The likelihood function

- To derive the conditional density of observation i of an i.i.d. random sample, note that $u_i = y_i - \boldsymbol{x}_i'\boldsymbol{\beta}$. Thus,

$$f(y_i|\boldsymbol{x}_i, \boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2}{2\sigma^2}\right) ,$$

  which is derived from the well-known formula for the p.d.f. of a normally distributed random variable.

1.3 The likelihood function

- Next, we compute the sample likelihood function:

$$\begin{aligned} L(\boldsymbol{\theta}) &= \prod_{i=1}^{N} f(y_i|\boldsymbol{\theta}) \\ &= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2}{2\sigma^2}\right) \\ &= \left(2\pi\sigma^2\right)^{-\frac{N}{2}} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\right) , \end{aligned}$$

where $\boldsymbol{y}$ is the $N \times 1$ vector of dependent variables and $\boldsymbol{X}$ is the $N \times K$ matrix of explanatory variables in our sample.

1.3 The likelihood function

- Finally, we derive the sample log likelihood function of the normal regression model:

$$\mathcal{L}(\boldsymbol{\theta}) = \ln L(\boldsymbol{\theta}) = -\frac{N}{2}\ln 2\pi - \frac{N}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})\,.$$

- To be continued. . .

1.4 Some examples of conditional ML models

Dependent variable is continuous and unbounded

- $y \in (-\infty, \infty)$
- Distribution: typically normal, $N(\mu, \sigma^2)$
- Two parameters: $\mu = \boldsymbol{x}'\boldsymbol{\beta}$
  σ^2 (unrestricted)
- Normal linear regression model, as in example 2 above.

1.4 Some examples of conditional ML models

Dependent variable is binary

- $y \in \{0, 1\}$
- Distribution: Bernoulli, $p^y(1-p)^{1-y}$, $p \in (0, 1)$
- One parameter: $p = \frac{e^{\boldsymbol{x}'\boldsymbol{\beta}}}{1+e^{\boldsymbol{x}'\boldsymbol{\beta}}}$

  $p = \Phi(\boldsymbol{x}'\boldsymbol{\beta})$
- Logit and probit models, respectively.
- More on these models in chapter 3.

1.4 Some examples of conditional ML models

Dependent variable is continuous and bounded

- $y \in (0, \infty)$
- Distribution: exponential, $\lambda e^{-\lambda y}$, $\lambda > 0$
- One parameter: $\lambda = e^{\boldsymbol{x}'\boldsymbol{\beta}}$
- Exponential regression model
- Not to be confused with censored regression.

1.4 Some examples of conditional ML models

Dependent variable is a count

- $y \in \{0, 1, 2, 3 \dots\}$
- Distribution: Poisson, $\frac{e^{-\lambda}\lambda^y}{y!}$, $\lambda > 0$
- One parameter: $\lambda = e^{\boldsymbol{x}'\boldsymbol{\beta}}$
- Poisson regression model
- More on these models in chapter 6.

1.5 Computation of the ML estimator

- In order to compute the value of the ML estimator of $\boldsymbol{\theta}$, we need to find the first-order conditions:

$$\frac{\partial \mathcal{L}_N(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{i=1}^{N} \frac{\partial \ln f(y_i|\boldsymbol{x}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{0}$$

- The gradient vector (left-hand side) is called the **score vector** or just the score.
- This system of equations sometimes has an analytic solution (e. g., in the linear regression case), but typically needs to be solved numerically – you will learn how to do this in the PC class.
- The number of equations is equal to the number of unknown parameters (elements of $\boldsymbol{\theta}$).

1.5 Computation of the ML estimator

- We also need the second derivatives (to convince us that we have a global maximum and to find it using numerical methods).
- This matrix is also called the information matrix (the expected value of the outer product matrix of the score):

$$I = \mathrm{E}\left(\frac{\partial \mathcal{L}_N(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial \mathcal{L}_N(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\right)$$

- Intuitively, it summarizes (co-)variances of the score vector. (Recall that $E(\boldsymbol{z}\boldsymbol{z}') = \mathrm{var}(\boldsymbol{z})$.)
- More on this below.

1.5 Computation of the ML estimator

Example 2, continued

- In the linear regression model, the first derivatives of the sample log likelihood function are

$$\begin{aligned}\frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}} &= \frac{1}{\sigma^2}\boldsymbol{X}'(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) \\ \frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \sigma^2} &= -\frac{N}{2\sigma^2}+\frac{1}{2\sigma^4}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})\,.\end{aligned}$$

1.5 Computation of the ML estimator

- Setting these to zero yields two expressions that look familiar from OLS estimation,

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}\boldsymbol{y} \\ \widehat{\sigma}^2 &= \frac{(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})}{N} = \frac{\boldsymbol{u}'\boldsymbol{u}}{N} \end{aligned}$$

- The only difference to the OLS expressions is that the estimator of σ^2 does not involve a degrees-of-freedom adjustment.
- This implies that ML estimation yields a smaller variance of u than OLS, but the difference vanishes as the sample size increases.

1.6 Consistency and asymptotic distribution of ML

- The formal derivation of the ML estimator is a bit involved, so we reproduce only the main insights.
- First, note that if the parametric model is correctly specified, then it must hold that

$$\mathrm{E}\{\ln[f(y_i|\boldsymbol{x}_i,\boldsymbol{\theta}_0)]|\boldsymbol{x}_i\} \geq \mathrm{E}\{\ln[f(y_i|\boldsymbol{x}_i,\boldsymbol{\theta})]|\boldsymbol{x}_i\}\,.$$

- Put very intuitively, the probability that observation i was generated by the true value of the parameter vector, $\boldsymbol{\theta}_0$, is (weakly) larger than the probability that it was generated by any other value of $\boldsymbol{\theta}$.

1.6 Consistency and asymptotic distribution of ML

- Using our short-hand notation $\ell_i(\boldsymbol{\theta}) = \ln f(y_i|\boldsymbol{x}_i, \boldsymbol{\theta})$, we can write this condition as

$$\mathrm{E}\{\ell_i(\boldsymbol{\theta}_0)|\boldsymbol{x}_i\} \geq \mathrm{E}\{\ell_i(\boldsymbol{\theta})|\boldsymbol{x}_i\}\,, \quad \forall \boldsymbol{\theta} \in \boldsymbol{\Theta}\,.$$

- This implies that

$$\boldsymbol{\theta}_0 = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\arg\max}\ \mathrm{E}\{\ell_i(\boldsymbol{\theta})\}\,,$$

  where the expectation is now with respect to the joint distribution of the data, $(\boldsymbol{x}_i, y_i)$.

1.6 Consistency and asymptotic distribution of ML

- By using the sample analogue of this expectation, we obtain the maximum likelihood estimator,

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\arg\max} \; \frac{1}{N} \sum_{i=1}^{N} \ln f(y_i | \boldsymbol{x}_i, \boldsymbol{\theta}) \, .$$

- The ML estimate is the maximizer of the (average) sample log likelihood function. Consistency follows since it's a special case of an M estimator; see Wooldridge's theorem 13.1 on p. 475.
- We still need to find the maximum of the likelihood function. In the first example, there exists an analytic solution. In most other cases, the solution to the maximization problem must be obtained by numerical methods. Examples are discussed in the PC class.

1.6 Consistency and asymptotic distribution of ML

- Next, we sketch the asymptotic normality result for ML estimators.
- The first derivatives of the log likelihood function are crucial here since they characterize its maximum.
- Define the **score (vector)** of the log likelihood function for observation i as

$$\boldsymbol{s}_i(\boldsymbol{\theta}) = \nabla_\theta \ell_i(\theta)' = \left(\frac{\partial \ell_i}{\partial \theta_1}(\boldsymbol{\theta}), \frac{\partial \ell_i}{\partial \theta_2}(\boldsymbol{\theta}), \ldots, \frac{\partial \ell_i}{\partial \theta_P}(\boldsymbol{\theta}) \right)' .$$

- Thus, the score is a $P \times 1$ vector containing the first partial derivatives of the log likelihood contributions of observation i.

1.6 Consistency and asymptotic distribution of ML

- Note that the expectation of an individual observation i's score, evaluated at the true value of the parameter vector, and conditional on the explanatory variables, is zero,

$$\mathrm{E}[\boldsymbol{s}_i(\boldsymbol{\theta}_0)|\boldsymbol{x}_i] = \mathbf{0}\,.$$

- The derivation is a bit too complicated for our purposes.

1.6 Consistency and asymptotic distribution of ML

- Next, we define the **Hessian matrix** for observation i (the $P \times P$ matrix of second partial derivatives of $\ell_i(\boldsymbol{\theta})$), assuming that $\ell_i(\boldsymbol{\theta})$ is twice continuously differentiable, as

$$\boldsymbol{H}_i(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \boldsymbol{s}_i(\boldsymbol{\theta}) = \nabla^2_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}) \,.$$

- Define the (negative of) the expectation of the Hessian matrix, also called the **information matrix**, as

$$\boldsymbol{A}_0 = -\,\mathrm{E}[(\boldsymbol{H}_i(\boldsymbol{\theta}_0)] \,.$$

- Since we solve a maximization problem, the expected value of the Hessian matrix is negative definite, and the information matrix is positive definite at $\boldsymbol{\theta}_0$.

1.6 Consistency and asymptotic distribution of ML

- One can further show that

$$\boldsymbol{A}_0 = -\,\mathrm{E}[\boldsymbol{H}_i(\boldsymbol{\theta}_0)] = \mathrm{E}[\boldsymbol{s}_i(\boldsymbol{\theta}_0)\boldsymbol{s}_i(\boldsymbol{\theta}_0)'] \,.$$

- The proof of this "unconditional information matrix equality" is non-trivial. Wooldridge, pp. 478–9, provides some intuition; a formal proof can be found in Newey and McFadden (1994), pp. 2146.

- We can now state the asymptotic normality result for the ML estimator (Wooldridge, Theorem 13.2 on p. 479):

$$\sqrt{N}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \xrightarrow{d} \mathrm{N}(\mathbf{0}, \boldsymbol{A}_0^{-1}) \,.$$

1.6 Consistency and asymptotic distribution of ML

- A final step is to estimate the asymptotic variance matrix $\mathrm{Avar}(\widehat{\boldsymbol{\theta}}) = \frac{1}{N}\boldsymbol{A}_0^{-1}$.

- The following three matrices can serve as an estimator of $\mathrm{Avar}(\widehat{\boldsymbol{\theta}})$. Each estimator is based on the sample average of a sample analogue of $\boldsymbol{A}_0$. Note that the expression $\frac{1}{N}$ cancels out.

1. The first estimator is based on the Hessian of the log likelihood and requires computation of the second derivatives:

$$\left[-\sum_{i=1}^{N} \boldsymbol{H}_i(\widehat{\boldsymbol{\theta}}) \right]^{-1}$$

1.6 Consistency and asymptotic distribution of ML

2. The second estimator is based on the outer product matrix of the score and thus requires only first derivatives:

$$\left[\sum_{i=1}^{N} \boldsymbol{s}_i(\widehat{\boldsymbol{\theta}})\boldsymbol{s}_i(\widehat{\boldsymbol{\theta}})'\right]^{-1}$$

- This estimator was proposed by Berndt, Hall, Hall, and Hausman (1974) as part of their algorithm for estimation of ML models (known as the **BHHH algorithm**). While it is easier to compute than the first estimator, it can have poor small-sample properties.

1.6 Consistency and asymptotic distribution of ML

3. The third estimator is attractive when $\boldsymbol{A}$ can be derived analytically, for instance in the Probit model.

$$\left[\sum_{i=1}^{N} \boldsymbol{A}(\boldsymbol{x}_i, \widehat{\boldsymbol{\theta}})\right]^{-1}$$

- The technical details of asymptotic variance estimation are somewhat involved; a careful discussion can be found in Newey and McFadden (1994), pp. 2153–2160.

1.6 Consistency and asymptotic distribution of ML

Efficiency of the ML estimator

- One can further show that the ML estimator has the smallest asymptotic variance among $\sqrt{N}$-consistent estimators (i. e., estimators that converging at that rate).
- The variance of the ML estimator attains the Cramer-Rao lower bound.
- But recall that consistency required correct specification of the model, i. e., of the distributional assumptions such as normality.

1.7 Discussion

- OLS and IV estimation of the linear regression model do not require distributional assumptions other than conditional expectation (or conditional covariance) assumptions on the explanatory variables, instruments, and the error term.

- Maximum likelihood estimation, in contrast, requires the (parametric) specification of the joint distribution of the data.

- In econometrics, ML models typically (but not always) assume normally distributed error terms.

1.7 Discussion

- If the distributional assumptions are correct, then maximum likelihood is the efficient estimator (in a broad class of estimators).
- If they are incorrect, the ML parameter and variance estimates are typically inconsistent. Thus, other estimation methods are more robust than ML.
- For linear models, OLS and IV are generally preferred because of their robustness.
- For nonlinear models, ML estimation has historically been the preferred estimation method (even though moment-based alternatives exist).