Cotrolling for unobserved heterogeneity

B.D. in Business Administration and Economics
Course in Quantitative Methods III

Rosario Barone
University of Rome "Tor Vergata"

rosario.barone@uniroma2.it

Omitted variable

Let assume that the dependent variable y depends on the covariates x and z, such that

$$Y = \alpha + \beta_x X + \beta_z Z + u$$

Let now consider a sample of n individuals. By defininig the linear regression problem:

$$E(y_i) = \alpha + \beta_x x_i$$

we aim to model the expected value of y as function of the covariate x.

In this formulation, we are excluding (omitting) the variable z.

Omitted variable bias

Two conditions must hold true for omitted-variable bias to exist in linear regression:

- $\hat{\beta}_z > 0$ (the omitted variable must be a determinant of the dependent variable);
- $cov(x, z) \neq 0$ (the omitted variable must be correlated with an independent variable specified in the regression).

Unobserved heterogeneity

The question is: what if the omitted variable is not even observed/observable?

Unobserved heterogeneity is a term that describes the existence of unmeasured (unobserved) differences between study participants or samples that are associated with the (observed) variables of interest.

Such unobserved heterogeneity leads to omitted variables bias: statistical findings based on the observed data may be incorrect.

Control for unobserved heterogeneity

Three solutions for dealing with unobserved heterogeneity:

- **Instrumental variables**
- Mixture models
- Panel data models

Instrumental Variables

Instrumental variable methods allow for consistent estimation when the explanatory variables (covariates) are correlated with the error terms in a regression model. Such correlation may occur when:

- changes in the dependent variable change the value of at least one of the covariates ("reverse" causation);
- there are omitted variables that affect both the dependent and independent variables;
- the covariates are subject to non-random measurement error.

Instrumental Variables

In linear models we select an "instrument" Z, which must satisfy two main requirements:

- It must be correlated with the explanatory variable X:
  - strong correlation $\rightarrow$ strong first stage;
  - weak correlation $\rightarrow$ misleading inferences about parameter estimates and standard errors.
- It must satisfy the exclusion restriction: it cannot be correlated with the error term, conditionally on the other covariates.

Estimation of instrumental variables

Intrumental variable models may be estimated via two-stage least squares:

- first stage:
  - $x_i = Z_i\gamma + v_i$
  - $\hat{x}_i = Z_i * \hat{\gamma}$
- second stage: $y_i = \hat{x}_i\beta + u_i$.

Control for unobserved heterogeneity

Three solutions for dealing with unobserved heterogeneity:

- Instrumental variables
- **Mixture models**
- Panel data models

Mixture models: why?

When analyzing a data set we assume that each observation comes from one specific distribution.

$$Y_i \sim N(\mu, \sigma^2) \quad \text{for } i = 1, \ldots, n$$

Then we proceed to estimate parameters of this distribution using maximum likelihood estimation, i.e.:

$$\frac{\partial \mathcal{L}(\mu, \sigma^2; Y)}{\partial \mu \partial \sigma^2} = 0 \quad \text{and} \quad \left(\hat{\mu}_{ML}, \hat{\sigma}^2_{ML}\right)$$

The assumption that each observation comes from one specific distribution may often be inadequate.

Mixture models: when?

In many cases, assuming that each sample comes from the same unimodal distribution is too restrictive and may not make intuitive sense. Often the data we are trying to model are more complex:

- Single or groups of observed individuals may have unobserved effects which may affect the estimates.

$\downarrow$

- Multimodality: multiple regions with high probability mass.

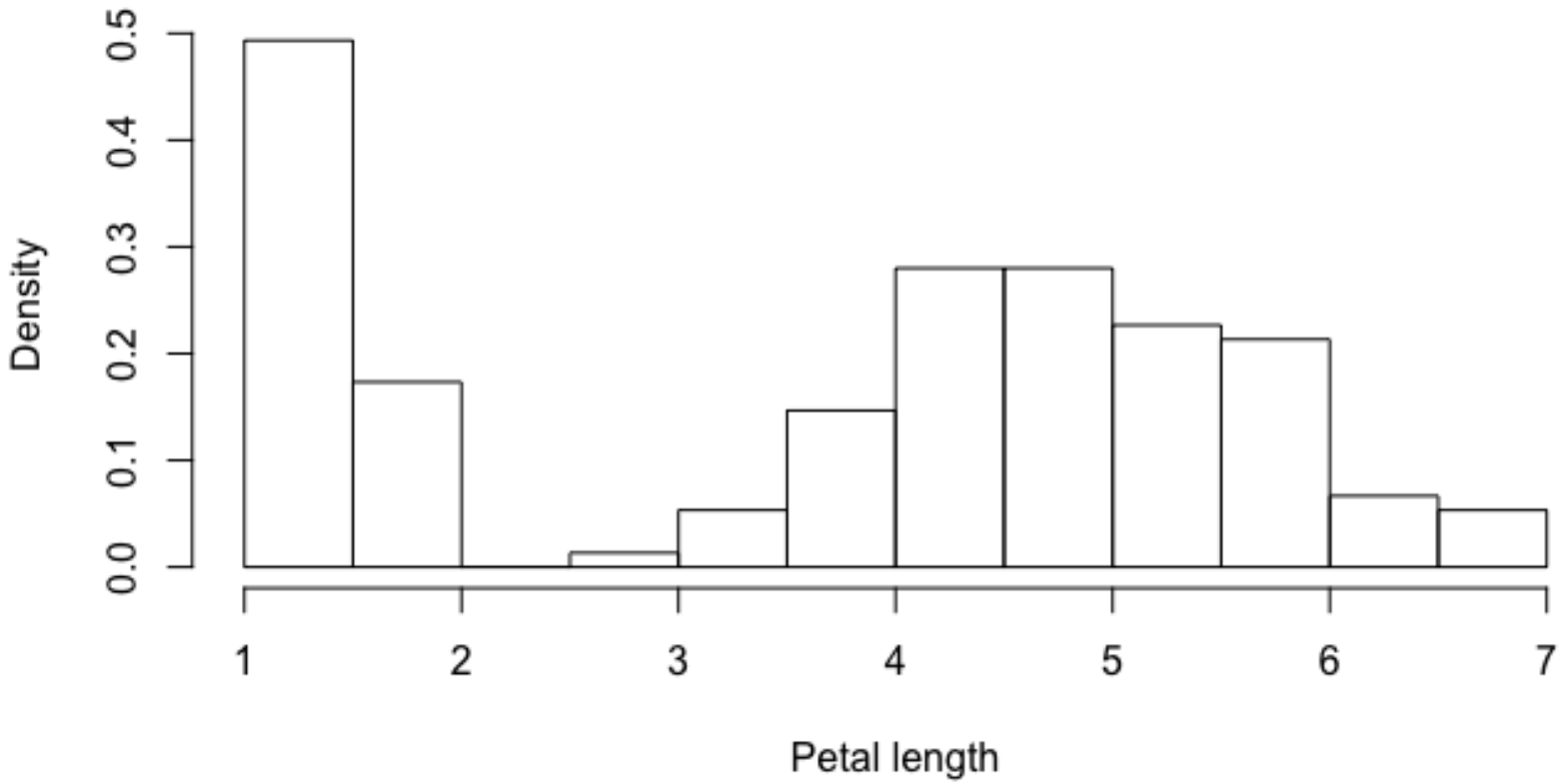
Mixture models: overview

A mixture model is a probabilistic model for representing the presence of subpopulations within an overall population.

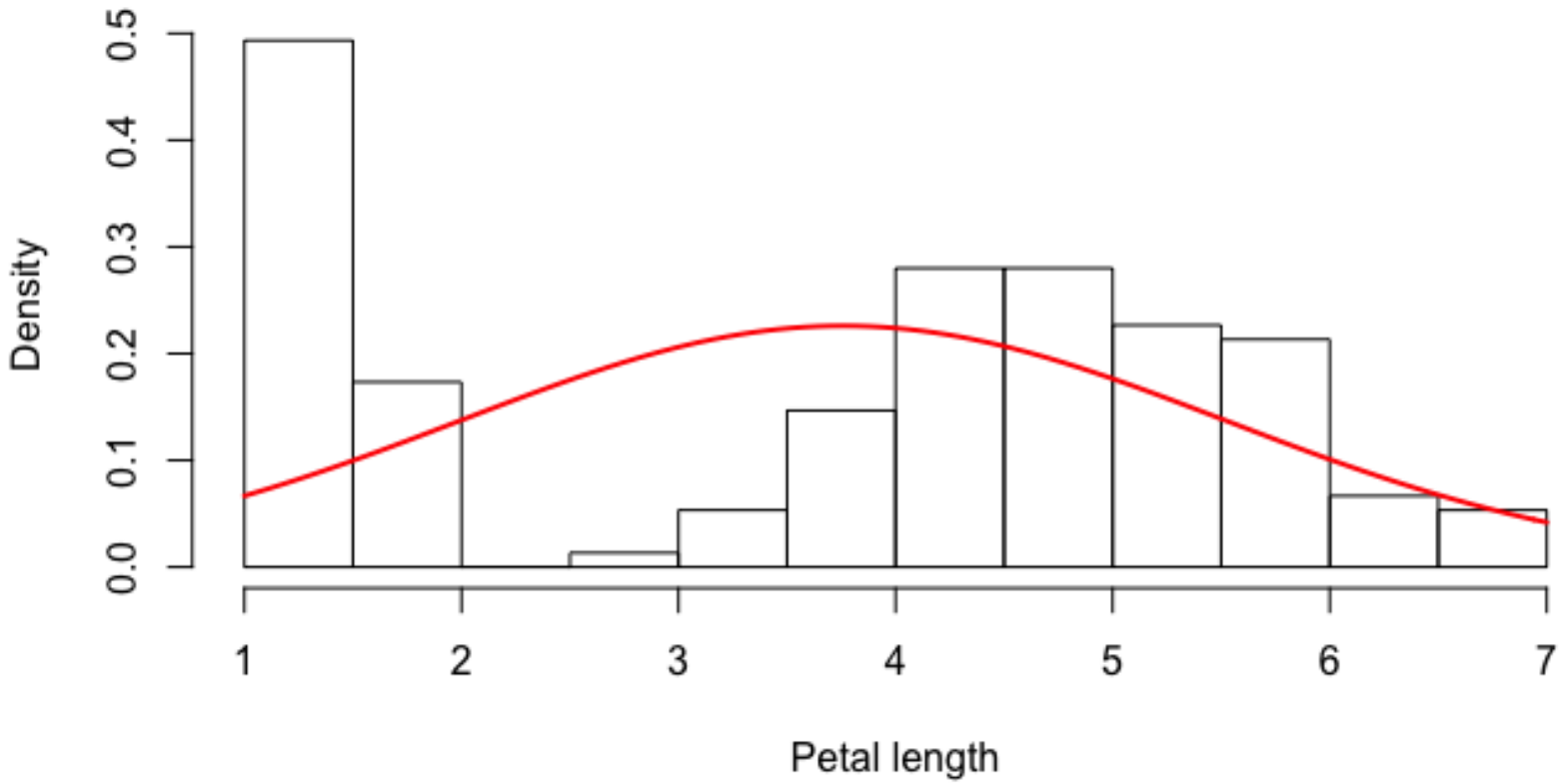
It does not require that an observed data set should identify the sub-population to which an individual observation belongs, i.e:

- we observe a sample of n individuals;
- we assume that in our sample there are K subpopulations;
- we do not specify the subpopulation to which each individual belongs.

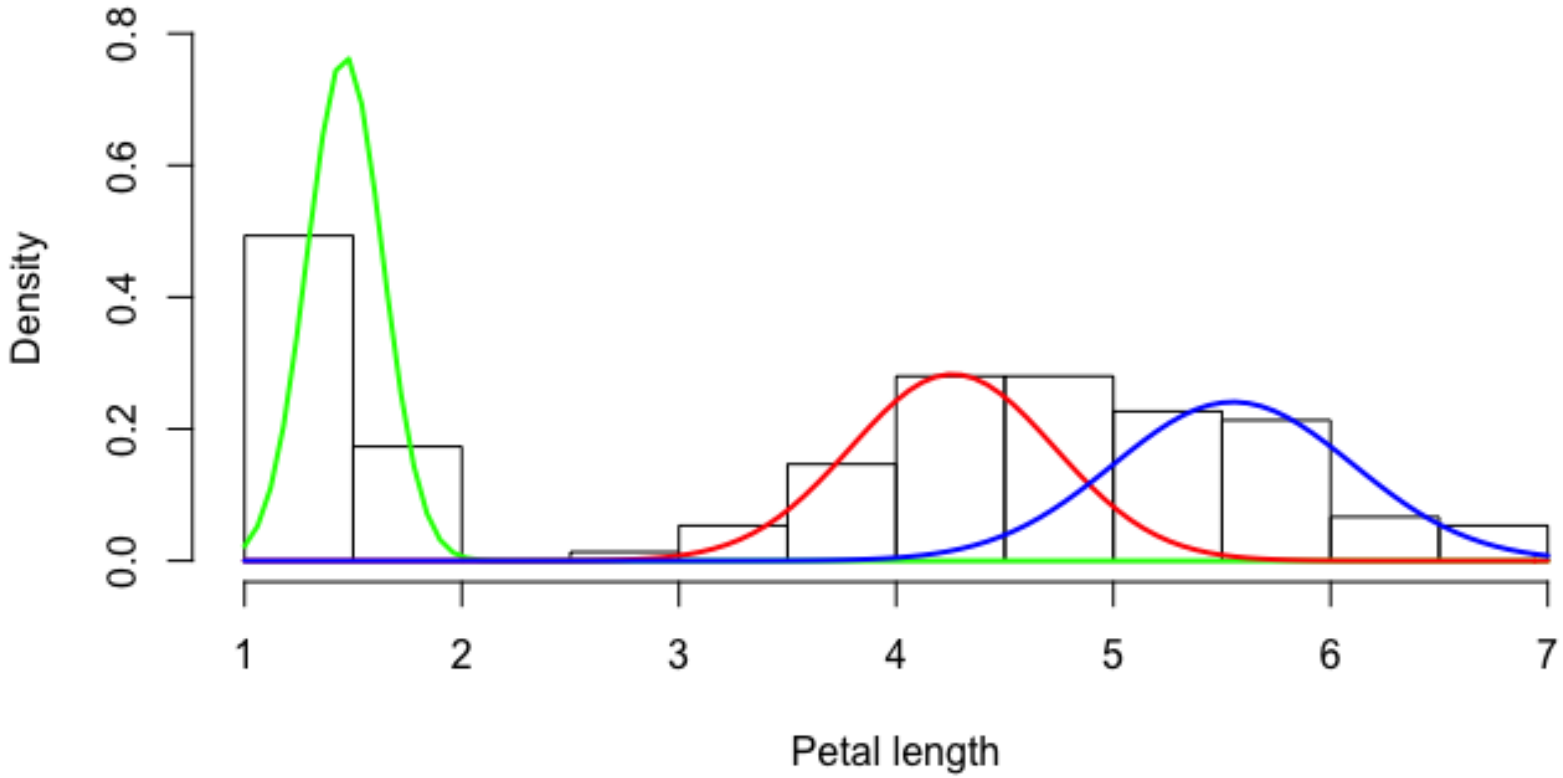
Mixture models: Iris data example



Mixture models: Iris data example



Mixture models: Iris data example



Finite mixture models

We define a mixture distribution $f_\theta \in \Omega$, with $\theta \in \Theta$:

$$f_\theta(y) = \sum_{k=1}^{K} w_k f_{\theta_k}(y)$$

where:

- $f_{\theta_k} \in \Omega \quad \forall k \in K$;
- $\sum_{k=1}^{K} w_k = 1$.

Note that in a sample of n observed individuals $K \leq n$.

Finite mixtures in regression analysis

Let y be a dependent variable and let x represent a covariate. We define a mixture of regression models as a distribution $f_\theta \in \Omega$, with $\theta = (\alpha, \beta) \in \Theta$:

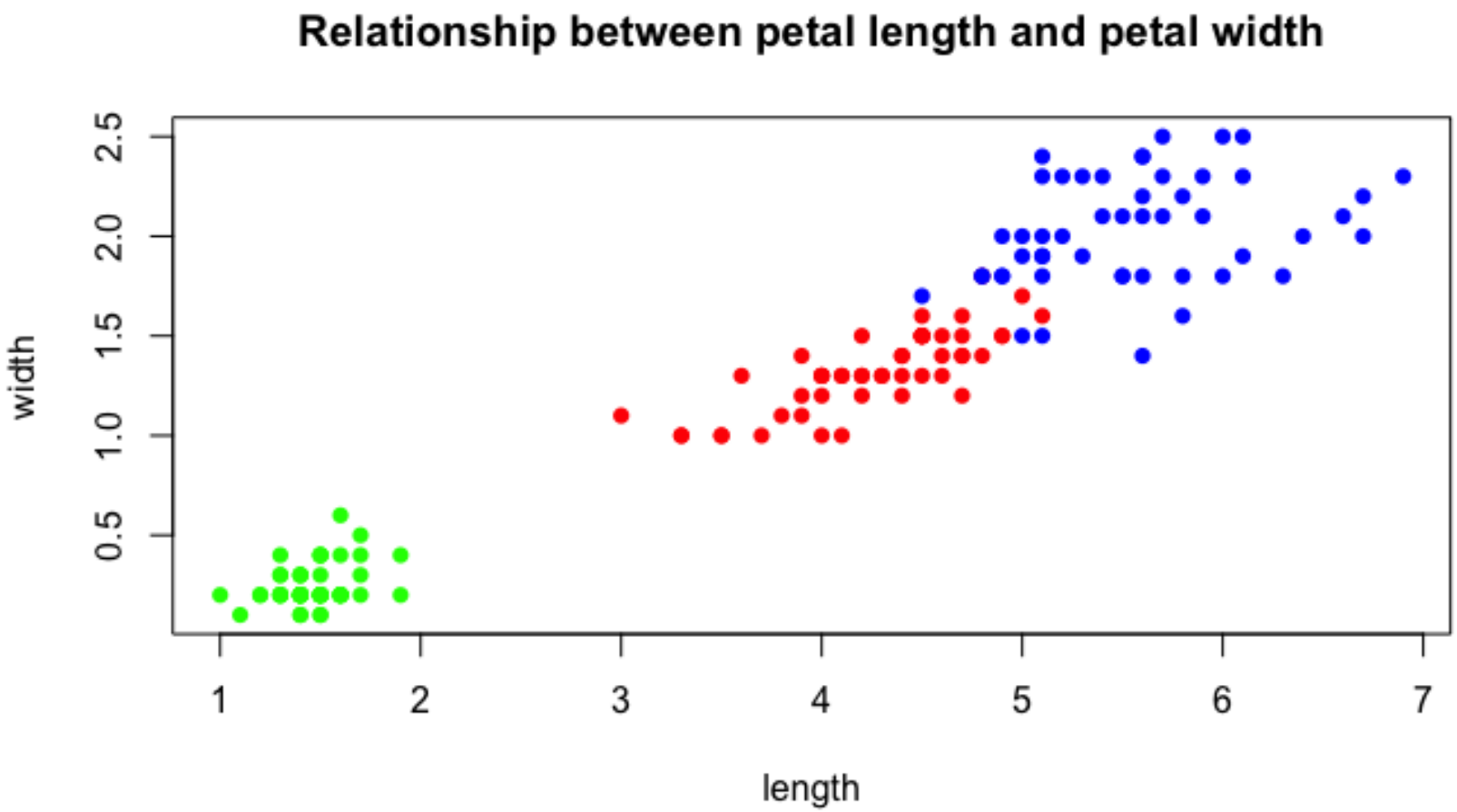
$$f_\theta(y|x) = \sum_{k=1}^{K} w_k f_{\theta_k}(y|x)$$

where:

- $f_{\theta_k} \in \Omega \quad \forall k \in K$;
- $\sum_{k=1}^{K} w_k = 1$;
- $g(E(y|x)) = \alpha_k + \beta_k x$.

Mixture in regression analysis: Iris data exam

Rosario Barone



Finite mixture models

Let assume to have a sample of N individuals. We define a mixture model as a *hiearchical model* composed by:

- N random variables that are observed, each distributed according to a mixture of K components belonging to the same parametric family of distributions, but with different parameters;
- N random **latent** variables specifying the identity of the mixture component of each observation, each distributed according to a *K-dimensional categorical distribution*;
- A set of K mixture weights w, which are probabilities that sum to 1.
- A set of K parameters, each specifying the parameter of the corresponding mixture component.

Finite mixture models: inference

Most of the approaches for finite mixture estimation that have been proposed focus on maximum likelihood methods.

Two scenarios to consider:

- if K is assumed to be known: expectation maximization (EM) is the most popular technique used to determine the parameters of a mixture with an a priori given number of components;
- if K is assumed to be uknown: (in general) methods to determine the number and functional form of the mixture components are distinguished from methods to estimate the corresponding parameter values.

Control for unobserved heterogeneity

Three solutions for dealing with unobserved heterogeneity:

- Instrumental variables
- Mixture models
- **Panel data models**

Panel data

Panel data (also longitudinal data or repeated measures) are repeated observations on the same cross section observed for several time periods.

Notation: let suppose to observe N individuals T times. We indicate the observation of the individual i at time t as $y_{i,t}$ with $i = 1, \ldots, N$ and $t = 1, \ldots, T$.

Short panel: large cross section of individuals observed for a few time periods, rather than a long panel such as a small cross section of countries observed for many time periods.

Introduction to panel data

First advantage: increased precision in estimation.

- this is the result of an increase in the number of observations ;
- for a valid inference there is the needing to control for likely correlation of regression model errors over time for a given individual;
- the usual OLS estimators in a pooled OLS regression typically underestimated standard errors and t-statistics that can be greatly inflated.

Second advantage: possibility of control for the consistent unobserved individual-specific effects:

- Fixed effects model: allows for unobserved individual heterogeneity data from a short panel, if heterogeneity is assumed to be additive and time-invariant.
- Random effects model: treat any unobserved individual heterogeneity as a realization of a random variable distributed independently of the regressors.

Third advantage: possibility of learning more about the dynamics of individual behavior than is possible from a single cross section.

Example: across section may yield a poverty rate of 20% but we need panel data to determine whether the same 20% are in poverty each year.

Panel data models and estimators: overview

- Panel data provide information on individual behavior in two-dimensions: across time and across individuals.
- Standard panel data analysis uses a much wider range of models and estimators than is the case with cross-section data.
- Obtaining correct standard errors of estimators is also more complicated than in the cross-section case.
- One needs to control for correlation over time in errors for a given individual, in addition to possible heteroskedasticity.