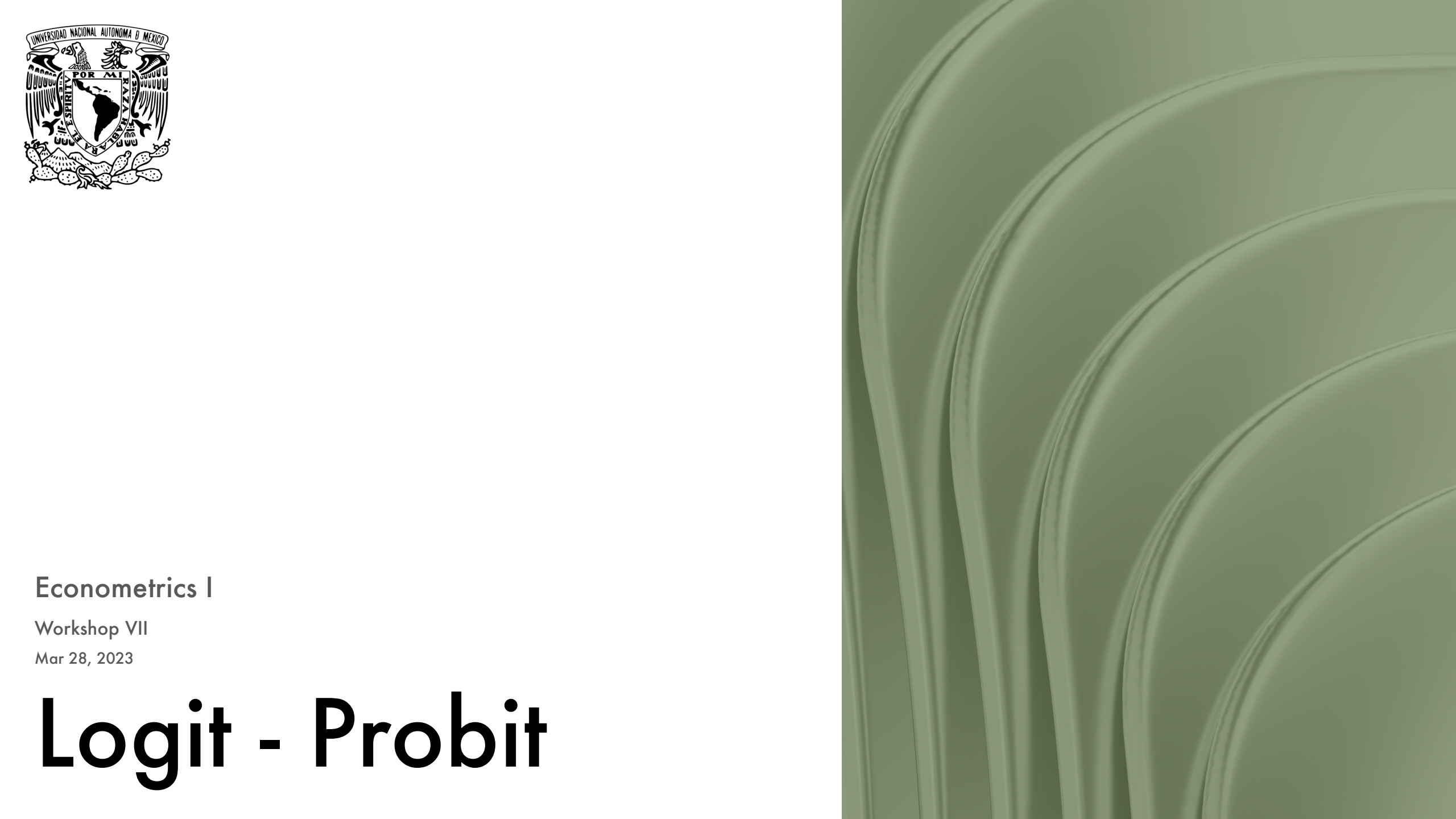


Econometrics I

Workshop VII

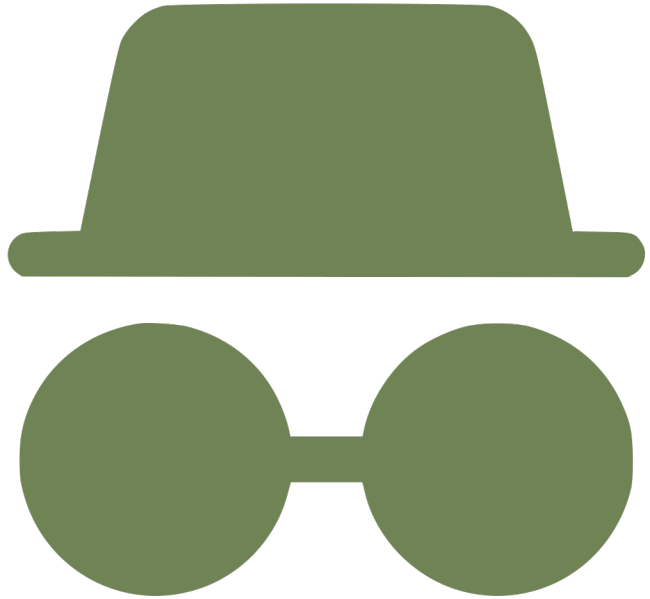
Mar 28, 2023

Logit - Probit

Up to now, we have studied the regression model, which means we have assumed a unidirectional relationship exists.

This unidirectional relationships in turn assumes there is one dependent variables and a set of independent that explicate the latest.

Implicitly, there is one more assumption:



Have you noticed the data type for dependent variable?

Which kind of type is it?

Dependent variable has always been quantitative!

If dependent variable is **quantitative**…

Objective: to estimate expected value given the regressors

Examples:

- Simple Lineal Regression
- Multiple Linear Regression
- MC2E

If dependent variable is **qualitative**…

Objective: to find probability that an event happens

Examples:

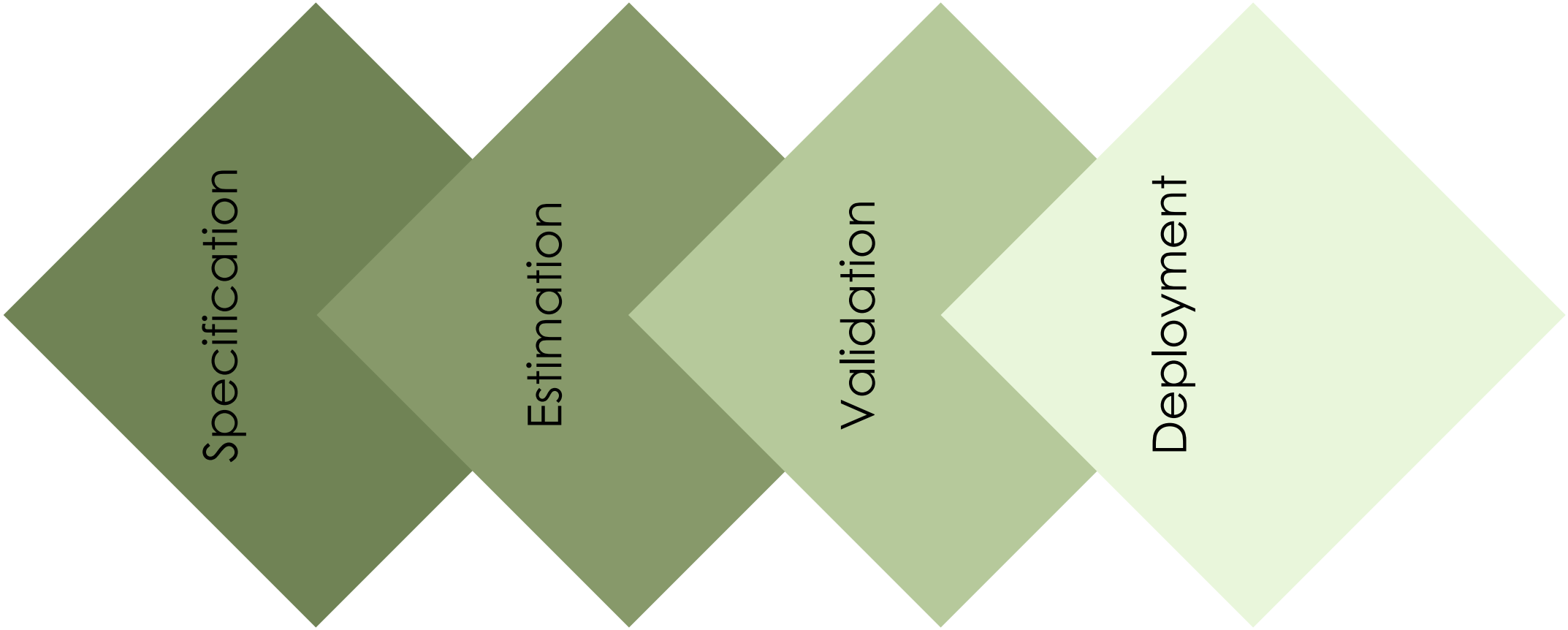
- Linear Probability Model (LPM)
- LOGIT
- PROBIT
- TOBIT

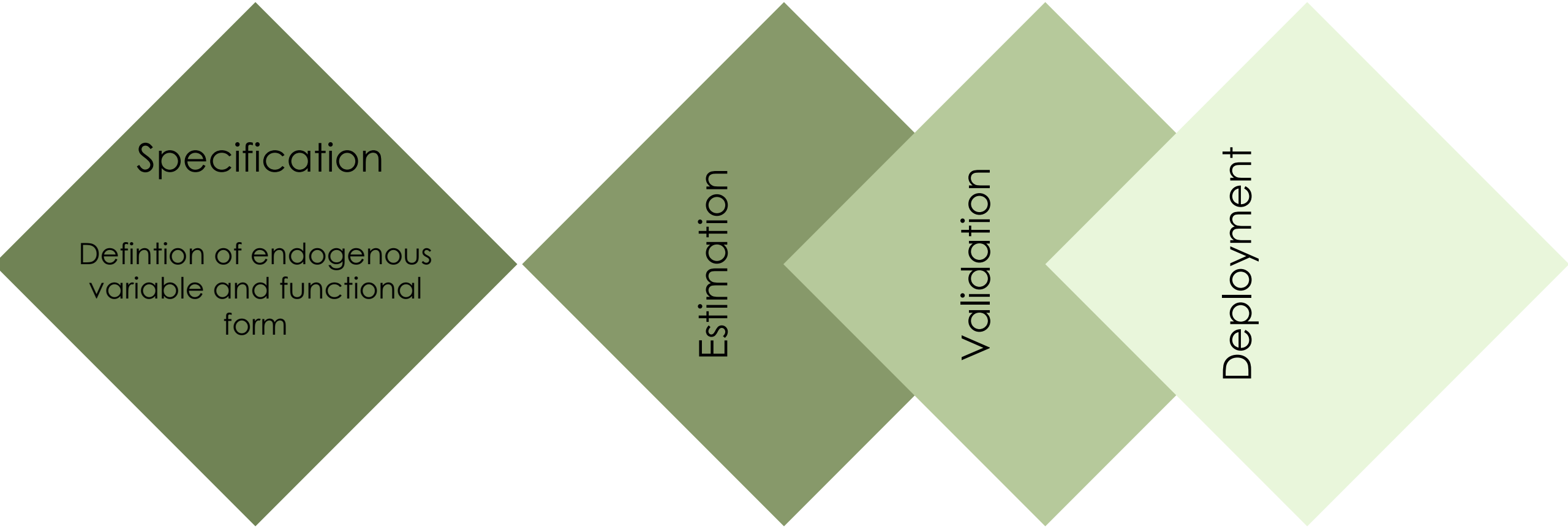
Daniel L. McFadden, American economist and cowinner (with James J. Heckman) of the 2000 Nobel Prize in Economic Sciences for his development of theory and methods used in the analysis of individual or household behaviour, such as understanding how people choose where to work, where to live, or when to marry

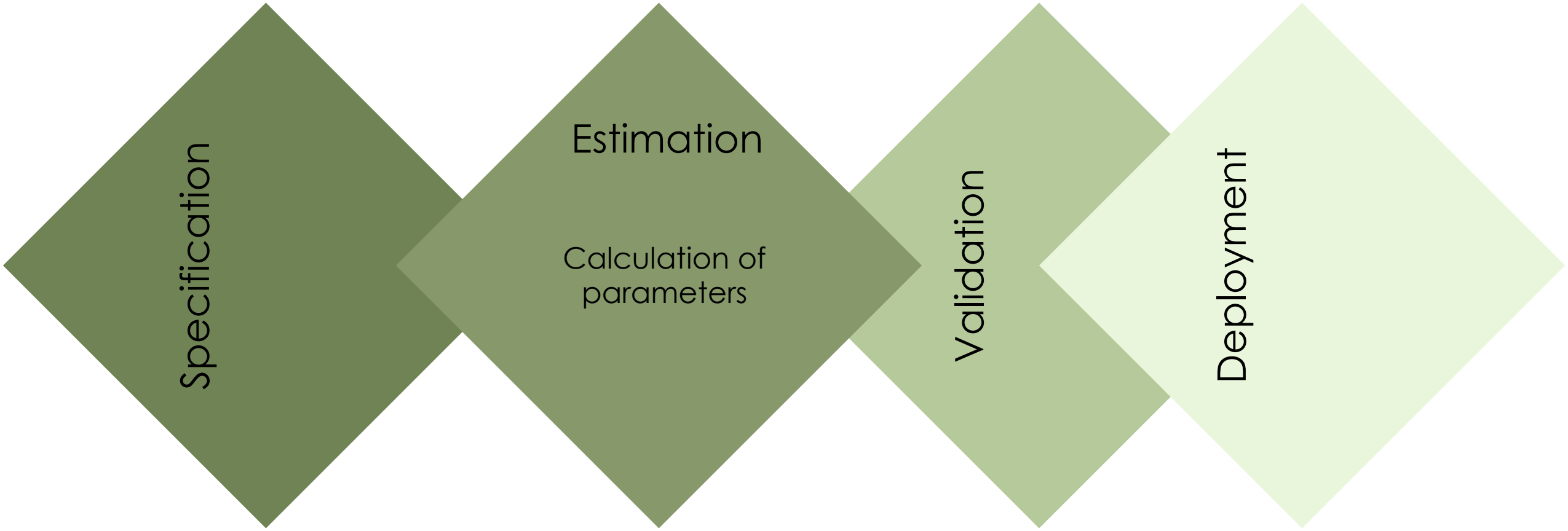
This models (qualitative dependent variable) allow to explain decisions taken by an individual from a set of independent variables

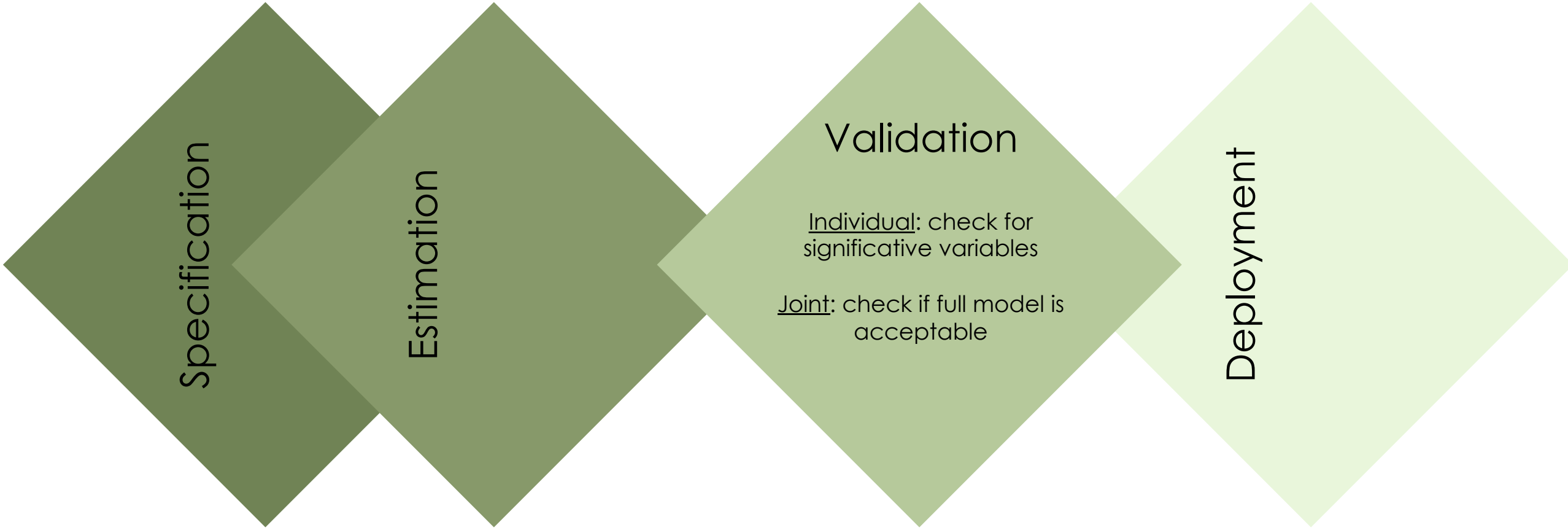
Given the number of alternative decisions that an individual can take, we differentaite between two types of models:

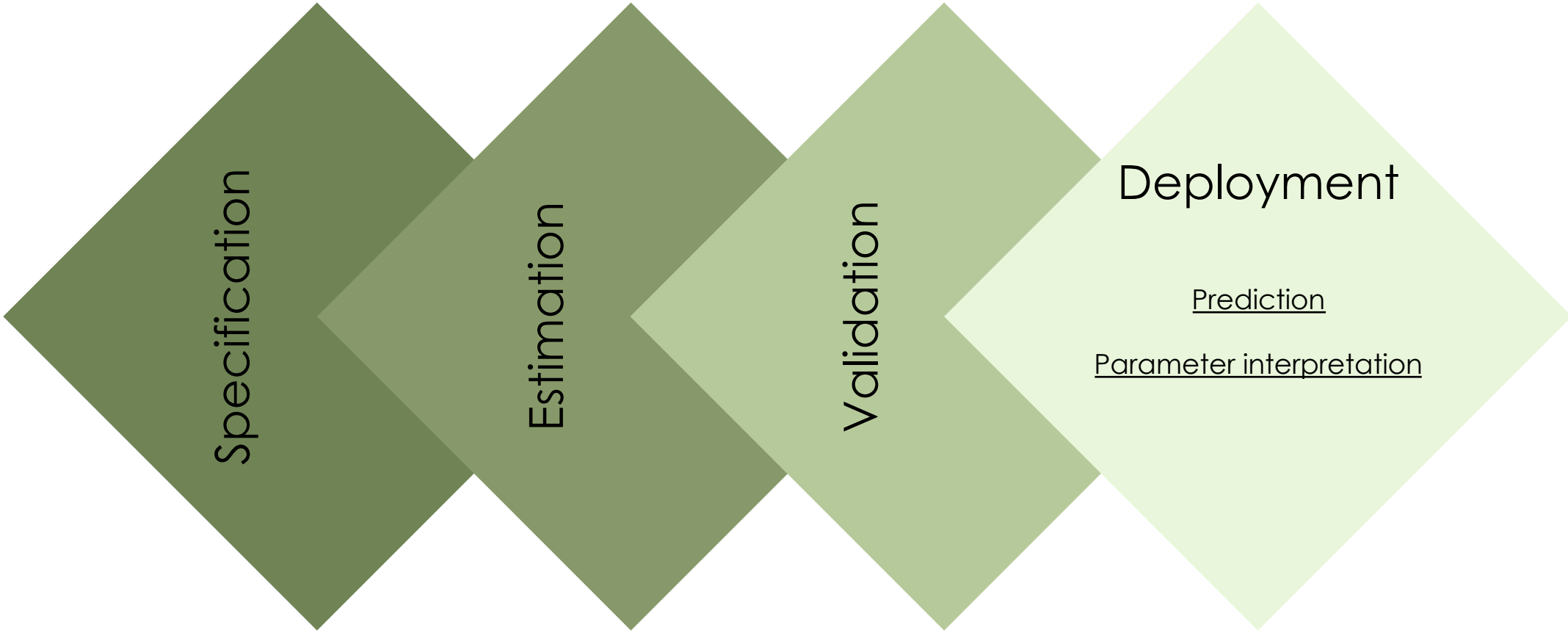
- Binary election
- Multiple election


Specification
Estimation
Validation
Deployment


Specification
Defintion of endogenous variable and functional form
Estimation
Validation
Deployment


Specification
Estimation
Calculation of parameters
Validation
Deployment


Specification
Estimation
Validation
Individual: check for significative variables
Joint: check if full model is acceptable
Deployment


Specification
Estimation
Validation
Deployment
Prediction
Parameter interpretation

Binary election models are those which dependent variable is dichotomous: it can take just two possible outcomes

Individuals can choose between two possible and alternative options.

These options are mutually exclusive!

$$Y_i = \begin{cases} 1\ yes \\ 0\ no \end{cases}$$

Multiple election models are those in which dependent variable can refer to more than categories

Individuals can choose between two or more possible options.

These options are mutually exclusive too!

$$Y_i = \begin{cases} 0\ if\ car\ A\ is\ bought \\ 1\ if\ car\ B\ is\ bought \\ 2\ if\ car\ C\ is\ bought \\ 3\ if\ no\ car\ is\ bought \end{cases}$$

Linear Probability Model (LMP) is the simplest model we can use. Nevertheless, given its simplicity, it has many disadvantages

It assumes that the relationship among variables is linear

$$Y_i = \beta_1 + \beta_2 X_2 + \cdots + \beta_k X_{ki} + u_i, \qquad i = 1, \ldots, N$$

LOGIT and PROBIT models are the same, but quite unique in distribution

LOGIT

Accumulative Logistic Distribution

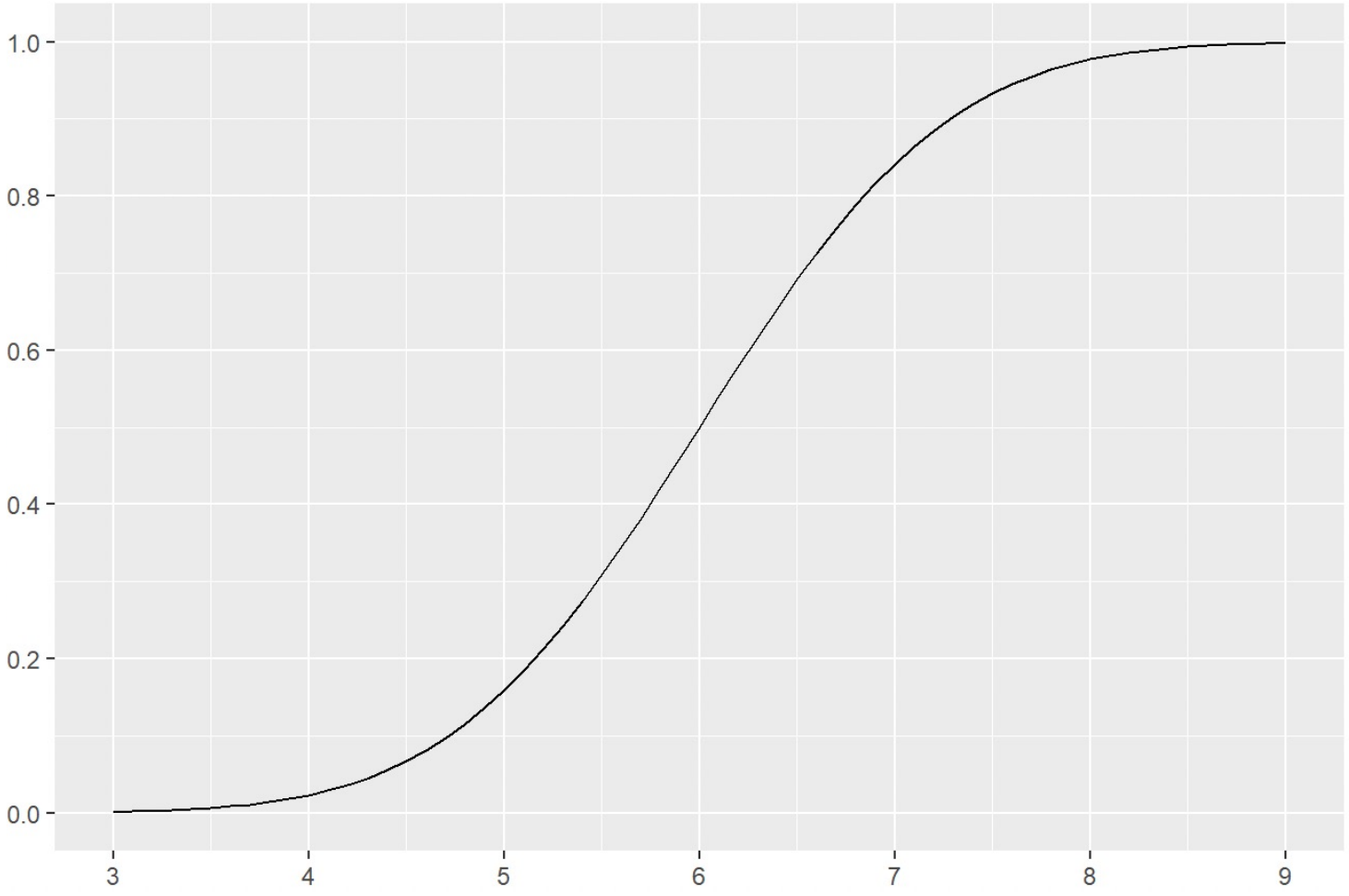
$$Y_i = \int_{-\infty}^{\alpha+\beta X} \frac{1}{(2\pi)^{1/2}} e^{-\frac{s^2}{2}} ds + u_i$$

PROBIT

Accumulative Standard Normal Distribution

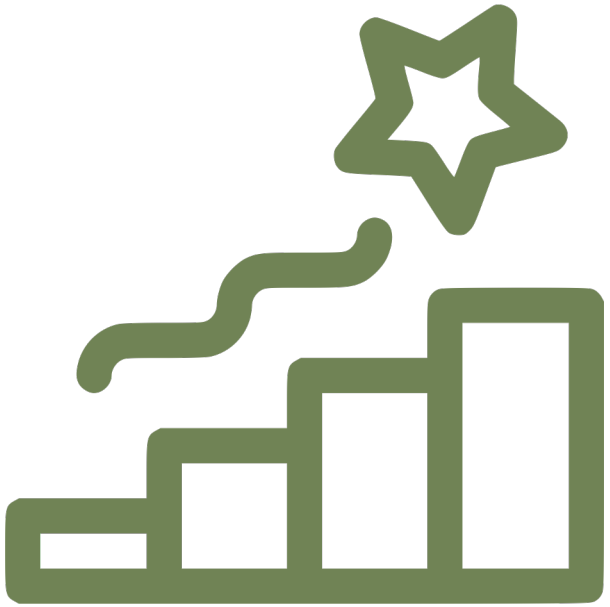
$$Y_i = \frac{1}{1+e^{-\alpha-\beta_k X_{ki}}} + u_i = \frac{e^{\alpha-\beta_k X_{ki}}}{1+e^{\alpha+\beta_k X_{ki}}} + u_i$$

Logit/Probit

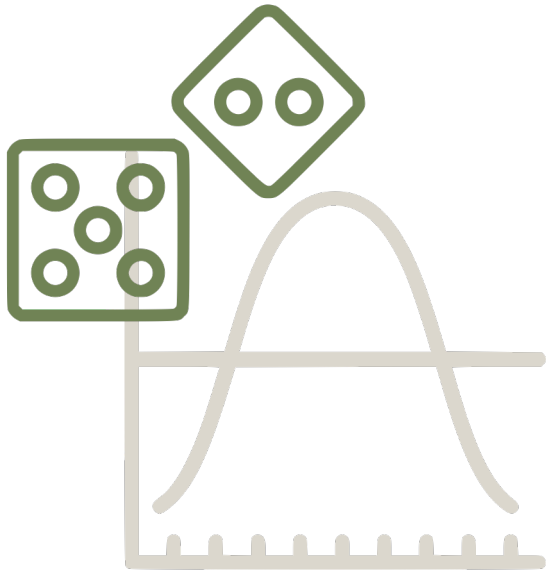


These models take into account conditional probability, (occurrence of Y given X) so output must be between 0 and 1

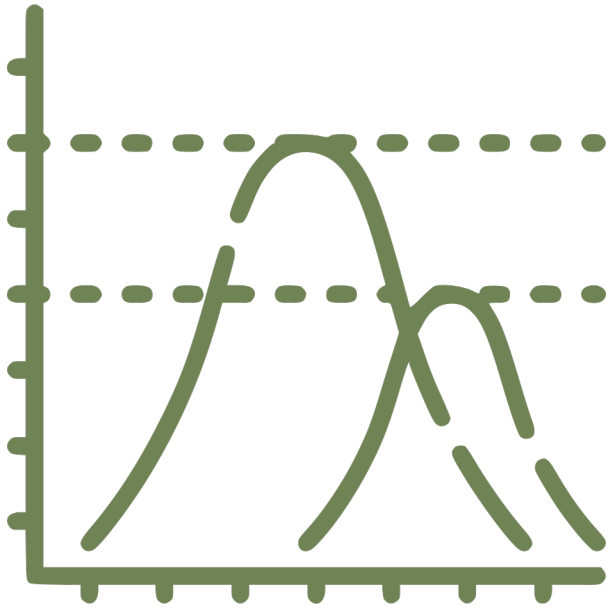
Logit



Logit is the most used and simplest discrete model. Its popularity is based on its equation and its explainability



MCO techniques are not longer suitable (probs must be between 0 and 1, not just 1 as usual)



We use Maximum Likelihood

Binary endogenous variable

Identifies the belonging from an individual into two possible outcomes. If person is 1, then the model will estimate the probability of individual to belong to target class

Exogenous variables

These variables allow to discriminate among groups and determine the belonging from each element to one group or another. They can be in nominal or ordinal scale.

Multiple answer

It is used when the number of alternatives for modelling are **more than two**

Not sorted data

It is used when endogenous variables show alternatives that **does not indicate any order**

Multinomial

It is used when regressors **relate to sampling observations**, so they vary among observations but not among alternatives

Conditional

It is used when regressors **relate to alternatives**, so its values vary among alternatives **being able or not** to vary among observations

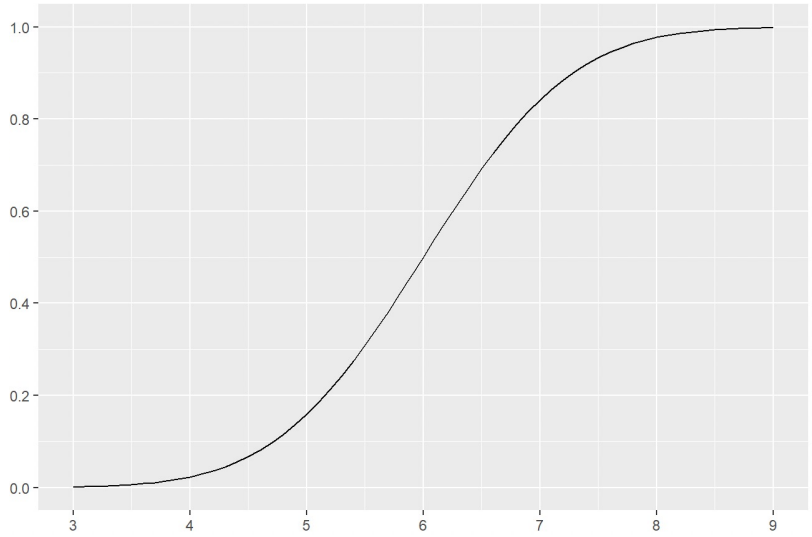
Declaring the model in terms of probability

$$P_i = \alpha + \beta X_i$$

Where P_i is the probability that a household i is an ownership of a house

This relationship generates a following chart

$$P_i = \frac{1}{1 + e^{-(\alpha\beta X_i)}}$$



We define odds ratio as:

$$\frac{P_i}{1 - P_i}$$

In the case of home-owning, the ratio represents the probability that a family owns a house with respect to the probability of not owning a house

For example, if $P_i = 0.8$ it means that probabilities are 4 out of 1 of owning a house (0.8/0.2)

We can interpret as follows

$$P_i = \alpha + \beta X_i$$

β is the slope and it measures a change in Y given a unitary change in X.
It can be interpreted as how the logarithm of probabilities (the condition being 0 or 1) changes as independent variables changes .

α is an autonomous parameter
Maximum likelihood is the estimation method

```
// Dataset: MROZ
// Keep some variables
keep inlf educ exper age kidslt6 kidsge6 repwage

// Run regression
reg inlf educ exper age kidslt6 kidsge6 repwage

// Estimate and run regression
estimates store MPL

logit inlf educ exper age kidslt6 kidsge6 repwage

estimates store LOGIT

estat class
```

```
// Run probit and store
probit inlf educ exper age kidslt6 kidsge6 repwage

estimates store probit fit

// Generate a table to compare
estimates table LOGIT PROBIT, star stat(N R2)
```

When the odds ratio is equal or close to 1, it means that there is NOT an association between `IND` and `DEP`.

Odds ratio thar show an association are above or below 1. All odds ratio less than 1 imply an inverse relationship

```
// Logit
Logit inflf educ exper age kidslt6 kidsge6 repwage
```

```
// Obtain marginal effects
mfx, at(age=30 educ=12)

// Fit probabilities
predict prob, p

// Difference between logit and probit models is the distribution
list prob
```

References

- **Salvatore, D., & Sarmiento, J. C.** (1983). *Econometría* (No. HB141 S39). McGraw-Hill.
- **Gujarati, D. N.** (2009). Basic econometrics. Tata McGraw-Hill Education.
- **Wooldridge, J.M.** (2016). *Introductory Econometrics*, Cengage Learning, 6th edition.
- **CFA Institute** (2020), "Level I, Volume 1, 2020, Ethical and Professional Standards and Quantitative Methods; Reading 7: Statistical Concepts and Market Returns", pp. 422-430