STATISTICAL MODELS

Focus on REGRESSION MODELS : study the relationship between variables

The role of the variables is asymmetric : there are 2 types of variables :

- RESPONSE / DEPENDENT variable y
- one or more PREDICTORS / COVARIATES / INDEPENDENT variables $x_1, x_2, \dots, x_p$

Goal of regression models: study how the response var. is influenced by the predictors

Examples : - evaluate how the blood pressure (y) is affected by a specific treatment (x_1), while also controlling for the individual characteristics (x_2 = age , x_3 = weight, ...)

- predict the number of claims (y) given the insurer's characteristics (age, past accidents, ...)

$\Rightarrow \quad y = g(x_1, \dots, x_p)$

our goal is to study $g(\cdot)$

As usual, the variables are observed on several individuals / statistical units

n = number of observations

We only consider 1 response variable y

The number of predictors is $p \geq 1$.

The data are organized into a matrix : rows → individuals
columns → variables

statistical unit	response variable	1st predictor	2nd predictor	...	p-th predictor
1	y_1	x_{11}	x_{12}	...	x_{1p}
2	y_2	x_{21}	x_{22}	...	x_{2p}
⋮	⋮	⋮	⋮		⋮
i	y_i	x_{i1}	x_{i2}	...	x_{ip}
⋮	⋮	⋮	⋮		⋮
n	y_n	x_{n1}	x_{n2}	...	x_{np}

the submatrix of the covariates

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix}$$

is called the "model matrix"

When do we need statistics? In the applications we consider, the value of the response variables is not fully determined, given the values of the covariates → there is UNCERTAINTY

The relationship between y and $(x_1, \dots, x_p)$ is stochastic.

STATISTICAL MODELS: we assume that the observations are REALIZATIONS of RANDOM VARIABLES

$\Rightarrow$ we study how the DISTRIBUTION of the RESPONSE VARIABLE depends on the values of the covariates

$\Rightarrow \quad Y \sim f(y \,;\, x_1, \dots, x_p)$

- common assumption : the covariates are non-stochastic and measured without error.
  this is justified in experimental settings (e.g. I fix the dose of the treatment and study the outcome).
  In observational studies this is clearly not possible (e.g. demographic / economic / social studies).
  For simplicity, the hypothesis is maintained, with the interpretation that the analysis is performed CONDITIONALLY on the observed values of the covariates (i.e. $Y_i \mid X_1 = x_1, \dots, X_p = x_p \sim f(y \,;\, x_1, \dots, x_p)$)

How do we actually build a model and perform the analysis?

The FUNDAMENTAL STEPS:

1) MODEL SPECIFICATION
   given the goal of the study and the available data, specify the model (also using past info, theories on the problem, ...)
2) ESTIMATATION
   estimate the model parameters (unknown quantities that define $g(\cdot)$) on the basis of the observed data
3) MODEL CHECKING / DIAGNOSTIC
   are the hypotheses underlying the model coherent with the observed data? Yes: use the model
   no: go back to 1) and repeat

1 MODEL SPECIFICATION

1A. The RANDOM COMPONENT

The type of model that we specify mainly depends on the nature of the response variable (since we are modeling the distribution of Y)

RESPONSE VARIABLE

- QUANTITATIVE
  - continuous (support $\mathbb{R}$) → Gaussian linear model; linear model via OLS (no Gaussian assumption)
  - discrete / counts (support $\mathbb{N}_0$) → Poisson regression (GLM)
- QUALITATIVE (categorical)
  - nominal variables (no order in the levels)
    - binary (only 2 levels, e.g. presence/absence) → Logistic regression (logit model), probit model (GLM)
    - more than 2 categories, not ordered (e.g. hair color) → Logistic regression / multinomial model (GLM)
  - ordinal variables (the categories have an intrinsic ordering, e.g. rankings low/medium/high, rates very unsatisfied / unsatisfied / satisfied / very satisfied) → Cumulative logit/probit model (GLM)

The type of response variable drives the choice of the distribution $f(y \,;\, x_1, \dots, x_p)$.

1B. The RELATIONSHIP between Y and $X_1, \dots, X_p$: $g(\cdot)$

it is deterministic : it is also called the SYSTEMATIC COMPONENT

we will consider the case where $g(\cdot)$ is completely specified by a FINITE set of (unknown) REAL PARAMETERS $\vartheta \in \Theta \subseteq \mathbb{R}^q$, $q \geq 1$ finite.

The specific way each covariate enters the model depends on the type of variable (more details later ...)

2 ESTIMATE

The estimate procedure consists in estimating the unknown parameters on the basis of the observed data.

Once we estimate $\hat{\vartheta}$, the relationship between Y and $x_1, \dots, x_p$ is completely known.

3 MODEL CHECKING

Having uniquely defined the model, we need to check :

- goodness of fit : does the model fit the observed data well?
- do we need all the considered covariates or a more parsimonious model can be defined (without loss of fit)?
- are the distributive assumptions satisfied?

If the model checking highlights some kind of problem, we have to go back to the model specification (and change, for example, the way the variables enter the model, the number of covariates, the assumptions on the law f) and repeat the procedure until step 3 gives good results.

Then, the model can be used for:

- inference on the parameters : understand the effect of each covariate
- prediction: given specific values of the covariates, what is the value of Y? (careful with prediction at values of the x_j outside of the observed range, ie extrapolation)

So far, we have denoted the relationship between Y and $(x_1, \dots, x_p)$ simply as $Y \sim f(y \,;\, x_1, \dots, x_p)$ meaning that the distribution of Y depends on the covariates.

ADDITIVE ERROR TERM

The simplest way to introduce the stochastic component is to consider

$$\underset{\text{stochastic}}{Y} = \underbrace{g(x_1, \dots, x_p)}_{\substack{\text{REGRESSION FUNCTION} \\ \text{deterministic}}} + \underset{\substack{\text{ERROR TERM} \\ \text{stochastic}}}{\varepsilon}$$

(notice : GLMs do not fall into this kind of specification)

Regression models can be classified based on:

1. the number of variables involved
2. the type of function linking Y to the x_j, $j = 1, \dots, p$

1 NUMBER OF VARIABLES

1a. number of INDEPENDENT variables :

- "SIMPLE" regression : only 1 covariate $Y = g(x_1) + \varepsilon$
- "MULTIPLE" regression: $p > 1$ covariates $Y = g(x_1, \dots, x_p) + \varepsilon$

1b. number of DEPENDENT variables :

- univariate : only 1 response Y
- multivariate : the response is a vector $\underline{Y} = (Y_1, \dots, Y_m)$

2 TYPE OF FUNCTION $g(\cdot)$

2a. PARAMETRIC : g can be expressed using a FINITE number of parameters $\underline{\vartheta} = (\vartheta_1, \dots, \vartheta_q) \in \Theta^q \subseteq \mathbb{R}^q$, q finite

→ LINEAR : $g(\cdot)$ is a parametric function and it is LINEAR in the parameters

We denote the parameters with β.

Examples :
$g(x) = \beta_1 x$
$g(x) = \beta_1 x + \beta_2 x^2 + \beta_3 x^3$
$g(x) = \beta_1 \log x - \beta_2 \sqrt{x}$
$g(x_1, x_2, x_3) = \beta_1 x_1 + \beta_2 \log x_2 + \beta_3 e^{x_1 + x_3}$

the parameters $\underline{\beta} = (\beta_1, \beta_2, \beta_3)$ enter linearly.

Notice that the variables x_j need not be linear! We can transform them to better fit the data.

→ "LINEARIZABLE" : the relation is not linear, but there is a transformation to make it so :

Example : the model $Y = \beta_1 \cdot x^{\beta_2} \cdot \varepsilon$ is not linear.

But if we take the logarithm : $\underbrace{\log Y}_{\tilde{Y}} = \underbrace{\log \beta_1}_{\tilde{\beta}_1} + \beta_2 \underbrace{\log x}_{\tilde{x}} + \underbrace{\log \varepsilon}_{\tilde{\varepsilon}}$

$\Rightarrow \tilde{Y} = \tilde{\beta}_1 + \tilde{\beta}_2 \cdot \tilde{x} + \tilde{\varepsilon}$ linear

→ NON-LINEAR : it is parametric but it is not linear nor linearizable

Example : $Y = \dfrac{\beta_1 x}{\beta_2 + x} + \varepsilon$

2b. NONPARAMETRIC : the parameter space Θ is not a subset of $\mathbb{R}^q$ (e.g. kernel regression, trees, RF, splines, nearest neighbors, ... GP regression)