Applied Microeconometrics
Fundamentals

Contents

Disclaimer: Sections and lines in brown are 'under construction'.

Motivation

Research questions related to the goal of sustainable development bring together social and natural systems, and are therefore particularly conducive to interdisciplinary work. The social system part demands some training in the social sciences, and in effect interdisciplinary researchers may have an economics background.

Applied microeconomics work in recent years has largely concerned the identification of causal relationships between variables, such that the current dominant methods and terminology are largely fitted to that goal. In applied work from other disciplines, one is likely to encounter alternative types of models, estimation methods, terminology, and even ultimate goals of the statistical analysis (e.g., predictive inference vs causal inference). If nothing else, an applied interdisciplinary researcher should be able to communicate with these different academic disciplines. This means notably understanding what a given method does in statistical terms, in other words: where it fits in the 'family tree' of statistical approaches. This will enable them to both: choose the most appropriate method given the problem at hand (when understanding what the method is doing, the empowered researcher need not resort only to the most common method in a given discipline), and justify that choice in front of the different disciplinary communities.

The purpose of this document is therefore twofold:

1. To detail the typical methods of applied microeconomics, which are our reference base. This includes defining and distinguishing common notions that may be conflated (*a model, an equation, a regression, a specification, an estimation method...*);
2. To put those into context, i.e., place them in the greater 'family tree' or space of statistical methods, and delineate a few other branches of that tree that may be relevant for empirical interdisciplinary research.

Let us start by defining microeconometrics:

Econometrics = (originally) the application of statistical methods to economic data, in order to measure the relationships of economic theory, i.e., obtain estimates that can be given a structural interpretation.

Microeconometrics = the use of these statistical methods to study microdata pertaining to individuals, households, and firms.

Ultimately, applied economics is a specific area of applied statistics. A distinguishing feature is the emphasis placed on causal modeling.

1 Statistical models

A model is a formal representation of a theory about a system, to ultimately describe that system.

A statistical model is a mathematical model of the data generating process (DGP)[a] of the sample $\{y_i, \mathrm{x}_i\}_{i=1}^n$.

- What distinguishes it from other mathematical models is that it is non-deterministic: some variables are stochastic or "random", they have probability distributions.[b]
- It is written as relationships between these *random* variables and some non-random variables, to study the **variation** of random variables. Specifically, it can serve 3 purposes: description (summarizing a sample); extraction of information; prediction.

[a] Formally, it combines the set of possible observations or "sample space" $\mathcal{S}$ and a collection of joint probability distributions on $\mathcal{S}$ (which ideally would include the "true" probability distribution induced by the DGP; but it doesn't need to, we accept that are models are false).

[b] Indeed, the task of statistics can be described as quantifying evidence and reasoning under *uncertainty*.

1.1 ⊃ Microeconometrics models

All empirical investigations in *microeconometrics* aim to uncover important relationships to understand microeconomic behavior. They can broadly be separated into two types of approaches, depending on the extent to which they rely on microeconomic theory:

- **Structural analysis** heavily depends on economic theory. Model specifications are derived from specifications of the economic behavior. The goal is to analyze structural relationships for interdependent microeconomic variables (e.g., to estimate structural parameters that characterize individual preferences or technological relationships).

$$g\big(y, \mathrm{x}, e|\theta\big) = 0, \quad \theta = \text{structural parameters}$$

- **Reduced form analysis** makes much less use of economic theory. The goal is to uncover associations among variables, by using regression models.

  The **reduced form** of a system of structural equations is the result of solving the system for the dependent (i.e., nonlagged and endogenous) variables. This **gives the dependent variables as functions of the independent variables (exogenous variables or lags of the dependent)**.

$$y = h\big(\mathrm{x}, e|\pi\big), \quad \pi = \text{reduced form parameters that are functions of } \theta$$

1.2 ⊃ Regression models

A regression model is a statistical model which models a *dependent variable* y as a function of *independent variables* x.

The variables $\{y, x_1, ..., x_k\}$ have an unknown joint distribution and complicated covariance structure. Instead of looking at the full joint distribution, regression models simplify the problem by **focusing on the *conditional distribution*[1] of y, given x.**

One generally uses regression models for three purposes: estimation, hypothesis testing, and prediction.

[1] Different regression models will look at different parts of the distribution, and specify them differently. Ex: classical linear regression model: $\mathbb{E}[y_i|\mathrm{x}_i] = f(\mathrm{x}_i) = \mathrm{x}_i'\beta$; quantile regression model: $\mathbb{Q}[y_i|\mathrm{x}_i] = f(\mathrm{x}_i)$...

Writing a regression model means that we consider that a sample $\{y_i, \mathrm{x}_i\}_{i=1}^n$ is generated by the process described by that model. We can write the model interchangeably:[2]

- in index notation, i.e., as a system of n equations: $y_i = f(\mathrm{x}_i, e_i|\beta),\ \forall i = 1, .., n$
- in matrix notation: $\mathrm{y} = f(\mathrm{X}, \mathrm{e}|\beta)$, where the error term e is a vector of n random variables, with an $n \times n$ symmetric covariance matrix.

Example: The classical linear regression model assumes a linear conditional expectation function and an additive error term:

$$y_i = \mathrm{x}_i'\beta + e_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + e_i, \quad e_i \overset{\text{iid}}{\sim} (0, \sigma^2), \quad i = 1, ..., n$$

$$\mathrm{y} = \mathrm{X}\beta + \mathrm{e}, \quad \mathrm{e} \sim (0, \sigma^2 \mathrm{I}_n)$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

Choosing a model specification To carry out regression analysis, one must first choose a model specification: select which independent variables to include and an appropriate functional form of $f()$.

Specification error occurs when either the functional form or the choice of independent variables poorly represents relevant aspects of the true DGP. Though "correct specification" is, in practice, unrealistic, as we do not observe the true DGP, we try to avoid the three basic types of misspecification:

- using an inappropriate functional form;
- including an x that is theoretically *irrelevant* (it has no partial effect on y) → *overspecified* model;
- excluding an x that is theoretically *relevant* (it may cause y) → *underspecified* model.

1.3 ⊃ Non-/Semi-/Parametric models

The specification of a statistical model can be:

- **parametric or "finite-dimensional":** the model is a family of distributions that has a *finite* number of parameters.[3] We assume that the data come from a population that can be adequately modeled by a probability distribution with a *fixed* set of parameters.
  - For *regression* models, it means that the distribution of the error term is fully characterized.
  - When the parameters uniquely specify the distribution,[4] we say that they are "identifiable".

  Ex: The Poisson family of distributions is parametrized by a single number $\lambda > 0$; the normal family is parametrized by two numbers $\{\mu, \sigma\}$.

[2] Here we have adopted the convention of Bayesian inference, where parameters are considered random variables, therefore the DGP is written conditional on β: $\mathrm{y} = f(\mathrm{X}, \mathrm{e}|\beta)$. In frequentist inference, the parameters are considered fixed, therefore we write $\mathrm{y} = f(\mathrm{X}, \mathrm{e}, \beta)$.

[3] Recall that a statistical model is a collection $\mathcal{P}$ of probability distributions on some sample space $\mathcal{S}$. We can write it as $\mathcal{P} = \{P_\theta | \theta \in \Theta\}$, where Θ is the parameter space. Hence we can write a parametric model as $\mathcal{P} = \{P_\theta | \theta \in \Theta \subseteq \mathbb{R}^k\}$.

[4] I.e., the correspondence of each distribution in $\mathcal{P}$ with a θ is 1-1, s.t. $P_{\theta_1} = P_{\theta_2} \Rightarrow \theta_1 = \theta_2$.

- **non-parametric:** the model makes no assumptions about a parametric distribution, it determines it from data.[5] The model has parameters, but their number and nature aren't fixed in advance.
  - For *regression* models, it means that no parametric form is assumed for the relationship between the dependent and the independent variables. *Ex: Kriging; LOESS.*
- **semi-parametric:** the model combines parametric and nonparametric models.
  Ex: Only a few moments are specified: $\mathbb{E}[\mathrm{e}] = 0_n$ *and* $\mathbb{V}[\mathrm{e}] = \mathbb{E}[\mathrm{ee}'] = \Omega$.

Why care about parametrization? Because what we are interested in is the class of probability distributions (as this will be our postulated model for observed data), and the parameter describes an integral feature of the probability distribution, s.t. knowledge about the parameter translates easily to knowledge about the distribution.

Identification in parametric models

> **Identification of a parameter =** its unique determination, given sufficient observations. *Assuming we had enough observations, could we determine the parameter?*

The model being "well-identified", i.e., the identification of all its parameters, is required for consistent estimation — and thus for meaningful statistical inference. It can be obtained through the functional form (by the parametrization of the error distribution) or from exclusion, inequality and covariance restrictions.

Example of non-identification: in the linear regression $y = \mathrm{x}\beta + e$, perfect collinearity between regressors means we can't identify β.

[5] Nonparametric regression requires larger sample sizes than regression based on parametric models, because the data must supply the model structure in addition to the model estimates. Nonparametric models also usually contain strong assumptions about independencies.

2 Data

Empirical studies can be separated into two classes, based on the type of data collected:

Study	Data collection
Experimental	The researcher records data about subjects while applying treatments and controlling conditions (active participation).
Observational	The researcher records data about subjects without applying a treatment (passive participation). If the goal is to uncover characteristics of a population, they may: • inspect the entire population: perform a **census**; • inspect a subset: take **sample data** S_t from the population probability distribution $F(\mathrm{W}_t \mid \theta_t)$.

2.1 Types of observational data

Observational data can be grouped into 3 categories, based on the dimensions: units (N) and time (T):

- **Cross-sectional [N]**: observations for several units, at one point in time;
- **Time series [T]:** observations for a single unit, at repeated points in time;
- **Longitudinal [N × T]**: observations for several units, at repeated points in time.

  When *the same units* are observed over time, we have **panel data.**[6] The panel can be:

  – balanced: all observed units i have data across all periods t;

  – unbalanced: some units have more observations than others.

Variation *between* units at one point in time is called *between*-variation, while variation *within* one unit across time is called *within*-variation. The total variance of observed variables can be split into within- and between-variation.

One of the strengths of longitudinal data is its potential for supporting causal relationships because of its ability to deal with observable and unobservable effects.

2.2 Sampling procedures

Random sampling ensures the *data* probability distribution is the same as the *population* distribution. If sampling isn't random, it is **biased**: the data distribution differs from the population distribution.

Common random sampling procedures include:

- **Simple random sampling** — the assumption on which statistical inference theory is based.
- **Stratified random sampling:** the population is divided into L subgroups or "strata", of $N_1 \neq N_2, ..., N_L$ units. Simple random samples of sizes $n_1, n_2, ..., n_L$ are drawn independently.

  – **Proportionate stratified random sampling**
  Ex: in a "10% sample, stratified across subgroups", the same fraction is applied on each subgroup.

[6] "Panel data" and "longitudinal data" are often used interchangeably, as most often it is the same units that are observed over time. However keeping the distinction, as delineated in Mertens et al. (2017), can be useful.

3 Statistical inference [under a frequentist approach]

Inferential statistics or **statistical inference** consists in *infering* properties of a population,[a] by calculating statistics from a sample drawn from the population.

It contrasts with descriptive statistics, which is solely concerned with properties of the observed data, not a larger population.

[a] *Population, DGP*, and underlying *probability distribution* could be used interchangeably. The data observed are of random variables, and we want to estimate parameters θ of their joint probability distribution. Making statistical inferences = deducing properties of (conditional) probability distributions.

Statistical inference combines data and (explicit or implicit) prior assumptions,[7] and generally involves:

- **Estimation** — 1. Estimating the value (point estimation) or potential range of values (confidence interval estimation) of an unknown parameter θ that characterizes the probability distribution of some feature of interest in the population; 2. Assessing the uncertainty around that estimate.
- **Hypothesis testing** — Testing for a specific value of the unknown parameter θ.

3.1 Frequentist vs Bayesian inference

There are two main paradigms for inference, whose difference is rooted in their definition of probability. Consider a parameter θ of unknown true value θ_0, and an *event* $\theta = \tilde{\theta}$ (i.e., θ taking this value $\tilde{\theta}$).

Frequentist approach	Bayesian approach
Definition of *probability* $\mathcal{P}$	
$\mathcal{P} \equiv$ the frequency of occurrence of an event; hence only repeatable events have $\mathcal{P}$s (ex: coin flips).	$\mathcal{P} \equiv$ one's belief in an event; hence any event, incl. non-repeatable, can have a $\mathcal{P}$.
Implication regarding θ	
$\implies$ θ is *fixed*. We can't assign $\mathcal{P}$s to events such as $\theta \lessgtr \tilde{\theta}$. We handle our uncertainty in the value of θ by limiting error rates (over imaginary experiments).	$\implies$ θ is a *random variable*. We can assign a $\mathcal{P}$ distribution over possible values of θ, to represent our uncertainty/belief in the value of θ.
Estimating θ using data	
1. Collect sample data, estimate the value (point $\hat{\theta}$) or potential range of values (confidence interval CI[θ]) of θ that is most consistent with the data. Result: a conclusion, summary of data, in the form of: – a "true/false" statement from a significance test, expected to be correct ...% of the time; or – a confidence interval, expected to cover the true value ...% of the time. ("time" = number of possible samples from the pop.)	1. Define a $\mathcal{P}$ distribution over possible values of θ 2. Collect sample data and update this distribution, by applying Bayes' theorem to each possible value: $$P(\tilde{\theta}|\text{data}) = \frac{P(\text{data}|\tilde{\theta}) \times P(\tilde{\theta})}{P(\text{data})}$$ Result: a *posterior* $\mathcal{P}$ distribution for θ. We can compute a 95% credible interval, s.t. "after seeing the data, there is a 95% chance that this CI contains the true θ."
Prediction	
Use the point estimate $\hat{\theta}$ as the most likely value of θ, and its CI.	Use the full posterior $\mathcal{P}$ distribution of $\hat{\theta}$, which allows for taking into account the uncertainty in $\hat{\theta}$.

[7] E.g., in Bayesian inference, an accurate prior (an assumption) will pull our estimates toward the true value. In frequentist inference, assuming a particular error distribution (i.e., parametric inference techniques) lends us power.

The sections below describe the *ABC* of statistical inference in the context of regression analysis, and under a frequentist approach, which is the classical approach in econometrics.

3.2 Estimation

3.2.1 Regression analysis

Regression analysis = a set of statistical processes for **estimating the relationship between a dependent variable y and independent variables x**:[a] $y = f(\mathrm{x}, e|\theta)$.

It is a way of summarizing and drawing inferences from data. It can have two purposes:

- prediction (interest is in $\hat{y}$): the **prediction** of the conditional distribution **of y**, given x;
- comparison (interest is in $\hat{\beta}$): comparing groups (which differ in x) or estimating causal effects.[b]

[a] Recall the definition of a regression model.

[b] Regression coefficient estimates $\hat{\beta}$ should be interpreted as "effects" only in causal inference. Otherwise, the safest interpretation is as a comparison, using the word "differences" rather than the words "effects" or "changes". E.g., *"the average difference in y, comparing two individuals that differ in x by one unit, is $\hat{\beta} = 0.29$"* or *"adding 1 unit to x corresponds to an increase of $\hat{\beta} = 0.29$ in an individual's predicted y"*.

⚠ Regressions calculate the *distribution of values* of the relation between y and x. The output is a conditional *distribution* $f_{y|\mathrm{x}}$. We can then choose to focus on its conditional mean $\mathbb{E}[y|\mathrm{x}]$, its conditional quantiles $\mathbb{Q}_{y|\mathrm{x}}()$...

3.2.2 Estimators

We have a set of observations $x_1, ..., x_n$, i.e., realizations of the sample of random variables $X_1, ..., X_n$.

An estimand $\boldsymbol{\theta}$ is a quantity of interest that we want to estimate, e.g., a parameter or some summary of the data. *Ex: the population mean μ_X.*

An estimator $\hat{\boldsymbol{\theta}}_{\boldsymbol{n}}$ of an estimand is a sample statistic, i.e., a function of the random sample (and therefore a random variable): $T_n = t(X_1, ..., X_n)$. Its values will vary sample to sample.
Ex: the sample mean $\bar{X}_n$ is an estimator for the population mean μ_X.

An estimate is a realization of that r.v. $\hat{\theta}_n$, calculated for our specific sample: $t_n = t(x_1, ..., x_n)$.

The most common estimators in microeconometrics are extremum estimators: they solve a min/max problem.

- **Maximum Likelihood (ML)**[8]

  We want to find the value of θ that makes the observed data most likely. The likelihood function in a

[8] The ML estimator is just a type of statistic, and can be conceptualized under either inference approach. From the vantage point of Bayesian inference, ML is a special case of 'maximum a posteriori' estimation that assumes a uniform prior distribution of the parameters. In frequentist inference, ML is a special case of extremum estimation, where the objective function is the likelihood.

regression model is the probability density of the data given the parameters and predictors:

$$\begin{aligned}\mathcal{L}\big(\mathrm{y} \mid \mathrm{X}, \theta\big) &= f\big(\mathrm{x}_1, ..., \mathrm{x}_n, \theta\big) \\ &= f\big(\mathrm{x}_1, \theta\big)...f\big(\mathrm{x}_n, \theta\big) \\ &= \prod_{i=1}^{n} f\big(\mathrm{x}_i, \theta\big) \\ \log \mathcal{L}\big(\mathrm{y} \mid \mathrm{X}, \theta\big) &= \sum_{i=1}^{n} \log f\big(\mathrm{x}_i{=}x_i, \theta\big)\end{aligned}$$

We compute $\hat{\theta}_{\text{ML}} \equiv \underset{\theta}{\text{argmax}}\ \mathcal{L}\big(\mathrm{y} \mid \mathrm{X}, \theta\big) = \underset{\theta}{\text{argmax}}\ \log \mathcal{L}\big(\mathrm{y} \mid \mathrm{X}, \theta\big)$

- **Least Squares (LS)**

  The fit of a model $y = g(\mathrm{x}, e)$ to each data point i is measured by its residual $r_i \equiv y_i - g(\mathrm{x}_i, \hat{\beta})$. We are interested in the values of the parameters that best fit the data, i.e., that minimize the sum of the squares of (eventually *a function* $k()$ *of*) the residuals.[9]

$$\hat{\theta}_{\text{LS}} \equiv \underset{\theta}{\text{argmin}} \sum_{i=1}^{n} k(r_i)^2$$

  When the model is linear, i.e., a linear combination of the parameters β: $g(\mathrm{x}, \beta) = \sum_j \beta_j h_j(\mathrm{x})$, Least Squares is a **Linear Least Squares (LLS)**.

  * **Ordinary Least Squares (OLS)**
    The OLS estimator has an exact closed-form solution:

$$\hat{\beta}_{\text{OLS}} \equiv \underset{\beta}{\text{argmin}} \sum_{i=1}^{n} r_i^2 = (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{y}$$

    In the simple case of the univariate regression model ($y = \alpha + \beta x + e$), the estimand is $\beta_{\text{OLS}} = \frac{\text{cov}[x,y]}{\mathbb{V}[x]}$ and the estimator (its sample analog) $\hat{\beta}_{\text{OLS}} = \frac{\frac{1}{n}\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n}\sum_i (x_i - \bar{x})^2}$. In a multivariate regression model, the coefficient on each x_k is $\beta^k_{\text{OLS}} = \frac{\text{cov}[\tilde{x}_k, y]}{\mathbb{V}[\tilde{x}_k]}$ where $\tilde{x}_k$ is the residual from the regression of x_k on all the other covariates.

  * **Weighted Least Squares (WLS)**
    When errors are heteroscedastic, i.e., each has variance σ_i, OLS won't be efficient among linear unbiased estimators. For least squares to give us the most *efficient* linear unbiased estimator, we minimize a *weighted* sum of squared residuals, using weights $w_i \propto \frac{1}{\sigma_i}$.

$$\hat{\beta}_{\text{WLS}} \equiv \underset{\beta}{\text{argmin}} \sum_{i=1}^{n} w_i r_i^2$$

  * **Generalized Least Squares (GLS)**
    When errors are heteroscedastic or correlated, i.e., when $x_1, ..., x_n \overset{iid}{\sim} f(x|\theta)$ doesn't hold (the

[9] Indeed, let $e \equiv y - \hat{y}$ be the unobserved error, $L(e)$ the loss. We want to minimize the expected loss $\mathbb{E}[L(e)|\mathrm{x}]$. For a squared error loss function $L(e) = e^2$, that function is the **conditional mean**: $g(X, \beta_{\text{LS}}) = \underset{g()}{\text{argmin}}\ \mathbb{E}\big[(y - g(\mathrm{x}, \beta))^2\big] = ... = \mathbb{E}[y|\mathrm{x}]$. With a given dataset, we look for the fit $g(\mathrm{X}, \beta)$ that minimizes the mean of that function $L()$ of the residuals; for the squared error loss function, it means minimizing the sum of squared residuals $\sum_i r_i^2$.

covariance matrix $\Omega \equiv \text{cov}[\text{e}|\text{X}]$ is not diagonal with values σ^2), OLS will again be inefficient. We minimize instead the squared *Mahalanobis length*[10] of the residuals:

$$\hat{\beta}_{\text{GLS}} \equiv \underset{\beta}{\text{argmin}} \sum_{i=1}^{n} \overrightarrow{d_M}^2(r_i)$$

When the model is a linear combination of the parameters, the GLS estimator has an exact closed-form solution: $\hat{\beta}_{\text{GLS}} = \underset{\beta}{\text{argmin}}(\text{y} - \text{X}\beta)'\Omega^{-1}(\text{y} - \text{X}\beta) = ... = (\text{X}'\Omega^{-1}\text{X})^{-1}\text{X}'\Omega^{-1}y$

* **Two-Stage Least Squares (2SLS)**
When regressors are correlated with the errors, we need a matrix of instruments Z s.t. $\mathbb{E}[z_i e_i] = 0$.
$\hat{\beta}_{\text{2SLS}} = \left(\text{X}'\text{Z}(\text{Z}'\text{Z})^{-1}\text{Z}'\text{X}\right)^{-1}\text{X}'\text{Z}(\text{Z}'\text{Z})^{-1}\text{Z}'\text{y}$

- **Least (symmetric) absolute error**
We are interested in minimizing a different loss function: the absolute error loss, $L(e) = |e|$. The corresponding estimator will be more robust to outliers. The optimal fit, i.e., the least absolute deviations fit, is the **conditional median**: $g(X, \beta_{\text{LSA}}) \equiv \underset{g()}{\text{argmin}}\ \mathbb{E}\left[|y - g(\text{x},\beta)|\right] = ... = \text{med}_{y|\text{x}}$.

- **Least asymmetric absolute error**
We can generalize to an asymmetric loss function: $L_\alpha(e) \equiv \begin{cases} (1-\alpha)|e| & \text{if } e < 0 \\ \alpha|e| & \text{if } e \geq 0 \end{cases} = \left(\alpha - \mathbb{1}\{e<0\}\right) \times e$, which places a different penalty on overprediction and underprediction. The optimal fit is the **conditional quantile** $g(X, \beta_{\text{LAA},\alpha}) \equiv \underset{g()}{\text{argmin}}\ \mathbb{E}\left[L_\alpha(y - g(\text{x},\beta))\right] = ... = \mathbb{Q}_{y|\text{x}}(\alpha)$.

Note: We have phrased all of the above in terms of the objective of finding the best fit. We could have also phrased it with the objective of making predictions about a specific part of the outcome distribution:

Objective: fit	*Objective: prediction*	*Optimal estimator/predictor*
$\min L(e) \equiv e^2$	*predict* $\mathbb{E}[y\|\text{x}]$	$\beta_{\text{LS}} \equiv \underset{\beta}{argmin}\ \sum_i \left(y_i - g(\text{x}_i, \beta)\right)^2$
$\min L(e) \equiv \|e\|$	*predict* $\text{med}_{y\|\text{x}}$	$\beta_{\text{LSA}} \equiv \underset{\beta}{argmin}\ \sum_i \|y_i - g(\text{x}_i, \beta)\|$
$\min L_\alpha(e) \equiv \left(\alpha - \mathbb{1}\{e<0\}\right)e$	*predict* $\mathbb{Q}_{y\|\text{x}}(\alpha)$	$\beta_{\text{LAA},\alpha} \equiv \underset{\beta}{argmin}\ \sum_i L_\alpha\left(y_i - g(\text{x}_i, \beta)\right)$

3.2.3 Estimator properties

See section 2 in https://clairepalandri.github.io/docs/CLRM&estimators.pdf.

3.2.4 Uncertainty in the estimate: computing SEs & CIs

i. The uncertainty in any sample statistic can be captured by its SE & $\text{CI}_{95\%}$

Samples are not unique. Many different samples could have been taken from the population. Any sample statistic (sample mean, slope parameter estimates...) will vary from sample to sample, hence it is a random variable, with a *sampling* probability distribution.

[10] The Mahalanobis distance is a measure of the distance between a point P and a distribution D. It is a multi-dimensional generalization of the idea of measuring how many standard deviations away P is from the mean of D. It is unitless and scale-invariant, and takes into account the correlations of the data set.

We are interested in the population parameter θ, and have computed an estimate $\hat{\theta}$ from our sample. As different samples would have lead to different $\hat{\theta}$s, $\hat{\theta}$ has a sampling distribution. If the distribution is rather condensed, i.e., the standard deviation is low *relative to the estimate*, it means we have high certainty about our estimate. We would quantify this certainty by computing $\text{SD}[\hat{\theta}]$ – and then use it to construct confidence intervals and test statistics. As we do not observe the sampling distribution (we haven't taken all the possible samples), we cannot observe $\text{SD}[\hat{\theta}]$. However, we can estimate it, and we'll call that estimate a "standard error" $\text{SE}[\hat{\theta}]$.

For any sample statistic $\hat{\theta}$, estimated with $n-k$ degrees of freedom:

- **Standard Error SE$[\hat{\theta}]$** = an estimate of the standard deviation of its distribution.
- The **95% Confidence Interval CI$_{95\%}[\theta]$** = the range of values s.t. *"I have a 95% confidence level that the true θ is in that range."*

Correctly interpreting the CI This confidence interval is based on the *sampling* distribution; the confidence refers to our uncertainty about the *sampling* method. The CI is therefore correctly interpreted in terms of repeated samples: *"Imagine we drew all possible random samples of size n. This interval would contain the true θ in 95% of the samples."*[11] I.e., we believe the 95% CI contains the true value, with the understanding that we'll be wrong 5% of the time. Another — maybe more adequate — name suggested for such intervals is "compatibility intervals", as they give a range of parameter values that are most compatible with our data and model/assumptions (Gelman and Greenland, 2019).

ii. Traditional approach: asymptotic theory

Consider a parameter of interest θ, and its estimator with $n-k$ degrees of freedom $\hat{\theta}$.

1. Standard Error SE$[\hat{\theta}]$

- Example 1: θ is the population mean $\boldsymbol{\mu_x}$, $\hat{\theta}$ is the sample mean $\bar{\boldsymbol{x}}$.
  - Population: X's mean μ_x and variance σ_x^2 are unobserved.
  - Sample: We measure the sample mean $\bar{x}$. Its variance $\mathbb{V}[\bar{x}] = \frac{\sigma_x^2}{n}$ is unobserved, as the population variance σ_x^2 is unobserved. A reasonable estimate for σ_x^2 that we do observe is the *sample* variance s_x^2.[12] We can thereby estimate $\mathbb{V}[\bar{x}]$ by $\hat{\mathbb{V}}[\bar{x}] \equiv \frac{s_x^2}{n}$, and $\text{SD}[\bar{x}]$ by $\text{SE}[\bar{x}] \equiv \frac{s_x}{\sqrt{n}}$.
- Example 2: θ is a regression slope $\boldsymbol{\beta_{\text{OLS}}}$ in the multivariate linear regression model.
  - Population: parameter β and error variance σ^2 are unobserved.
  - Sample: We measure the parameter estimate $\hat{\beta} \sim \left(\beta, \mathbb{V}[\hat{\beta}]\right)$. The formula of $\mathbb{V}[\hat{\beta}]$ is known but unobserved — as it is notably a function of σ.
    For simplicity, consider the simple case of normal errors: $e|X \sim \mathcal{N}(0, \sigma^2 I)$. Then $\mathbb{V}[\hat{\beta}] = \sigma^2(\text{X}'\text{X})^{-1}$. We can consistently estimate the population variance σ^2 by the bias-adjusted *sample* variance $s^2 \equiv \frac{1}{n-k}\sum_i r_i^2$.[13] We can thereby estimate $\mathbb{V}[\hat{\beta}]$ by $\widehat{\mathbb{V}}[\hat{\beta}] \equiv s^2(\text{X}'\text{X})^{-1} =$

[11] This is a probability statement about the interval, not the population parameter. It says $\text{P}(\beta \in \text{CI} \mid \beta) = 95\%$. This is different from saying *"there is a 95% probability that the true β lies within this range"*, i.e., $\text{P}(\beta \in \text{CI} \mid \text{CI}) = 95\%$. CIs are a frequentist concept, and this second erroneous interpretation contradicts the frequentist interpretation of probability. In the strict frequentist paradigm, the parameter is unobserved but it is set, so a probability statement on its value does not make sense. The probability applies to the interval, not to the true parameter value.

[12] ⚠ The standard deviation of the sample s has nothing to do with the standard error of the estimate $\text{SE}[\hat{\theta}]$. The first converges to the standard deviation of the population σ as $n \to \infty$, the second to 0.

[13] The bias here refers to that from the reduced degrees of freedom stemming from estimating the sample means.

$\frac{1}{n-k}\sum_i r_i^2\ (\mathrm{X}'\mathrm{X})^{-1}$, and $\mathrm{SD}[\hat{\beta}]$ by $\mathrm{SE}[\hat{\beta}] \equiv \sqrt{\frac{1}{n-k}\sum_i r_i^2 (\mathrm{X}'\mathrm{X})^{-1}}$.

2. Confidence Intervals CI[θ]

We want to give a range of estimates for the unknown parameter θ. Consider *the estimate of* the centered and standardized estimate $\frac{\hat{\theta}-\theta}{\mathrm{SE}[\hat{\theta}]}$.[14]

- Example 1: $\hat{\theta} \equiv \bar{x}$
  $\frac{\bar{x}-\mu}{\sqrt{s^2/n}}$ has a $\mathcal{T}$ distribution with $n-1$ degrees of freedom. Hence, by definition:

$$\begin{aligned} & \mathrm{P}\Big(q_{\mathcal{T}_{n-1}}(0.025) \ \leq\ \tfrac{\bar{x}-\mu}{\sqrt{s^2/n}}\ \leq\ q_{\mathcal{T}_{n-1}}(0.975)\Big) = 0.95 \\ \iff\ & \mathrm{P}\Big(\bar{x} - q_{\mathcal{T}_{n-1}}(0.975)\tfrac{s}{\sqrt{n}} \ \leq\ \mu\ \leq\ \bar{x} - q_{\mathcal{T}_{n-1}}(0.025)\tfrac{s}{\sqrt{n}}\Big) = 0.95 \end{aligned}$$

  where $q_{\mathcal{T}_{n-1}}()$ is the quantile function of the $\mathcal{T}_{n-1}$ distribution. We can thus define the 95% CI:

$$\mathrm{CI}_{95\%}[\mu] \equiv \Big[\bar{x} - q_{\mathcal{T}_{n-1}}(0.975)\ \tfrac{s}{\sqrt{n}}\ ;\ \bar{x} - q_{\mathcal{T}_{n-1}}(0.025)\ \tfrac{s}{\sqrt{n}}\Big] = \Big[\bar{x} \pm q_{\mathcal{T}_{n-1}}(0.975)\ \tfrac{s}{\sqrt{n}}\Big]$$

- Example 2: $\hat{\theta} \equiv \hat{\beta}_{\mathrm{OLS}}$
  If errors are normally distributed, then the sampling distribution of $\frac{\hat{\beta}-\beta}{\mathrm{SE}[\hat{\beta}]}$ is a Student's $\mathcal{T}$ distribution with $n-k$ degrees of freedom. Hence, by definition:

$$\begin{aligned} & \mathrm{P}\Big(q_{\mathcal{T}_{n-k}}(0.025) \ \leq\ \tfrac{\hat{\beta}-\beta}{\mathrm{SE}[\hat{\beta}]}\ \leq\ q_{\mathcal{T}_{n-k}}(0.975)\Big) = 0.95 \\ \iff\ & \mathrm{P}\Big(\hat{\beta} - q_{\mathcal{T}_{n-k}}(0.975)\ \mathrm{SE}[\hat{\beta}] \ \leq\ \beta\ \leq\ \hat{\beta} - q_{\mathcal{T}_{n-k}}(0.025)\ \mathrm{SE}[\hat{\beta}]\Big) = 0.95 \end{aligned}$$

  We can thus define the 95% CI:

$$\mathrm{CI}_{95\%}\big[\beta\big] \equiv \big[\hat{\beta} - q_{\mathcal{T}_{n-k}}(0.975)\ \mathrm{SE}[\hat{\beta}]\ ;\ \hat{\beta} - q_{\mathcal{T}_{n-k}}(0.025)\ \mathrm{SE}[\hat{\beta}]\big] = \big[\hat{\beta} \pm q_{\mathcal{T}_{n-k}}(0.975)\ \mathrm{SE}[\hat{\beta}]\big]$$

The larger its degrees of freedom, the closer a $\mathcal{T}$ distribution gets to the standard normal distribution. Therefore, in both examples, when $n-k$ is sufficiently large, we can simply use the normal distribution:[15]

$$\mathrm{CI}_{95\%}\big[\theta\big] \simeq \big[\hat{\theta} \pm q_{\mathcal{N}}(0.975)\ \mathrm{SE}[\hat{\theta}]\big] = \big[\hat{\theta} \pm 1.96\ \mathrm{SE}[\hat{\theta}]\big]$$

iii. Simulation approach: Bootstrap

The traditional approach relies on the assumed *asymptotic* sampling distribution of the statistic. This distribution rests on asymptotic theory (that usually leads to limit normal and χ_2 sampling distributions). When our sample size is small (making this asymptotic approximation incorrect), or when analytical expressions for the uncertainty of the particular statistic are complicated, i.e., when conventional analytic approximations fail, we can create an alternative sampling approximation of the finite-sample distribution of interest by **"Bootstrap".**

The Bootstrap procedure is a way to estimate the sampling distribution of the sample statistic, by resampling with replacement from the current sample to generate multiple "resamples".[16] Supposing 100 Bootstrap resamples, we can obtain 100 estimates and estimate $\mathrm{SE}[\hat{\theta}]$ by their standard deviation.

Advantages and limits:

[14] The standardized estimate is $\frac{\hat{\theta}-\theta}{\mathrm{SD}[\hat{\theta}]}$. The scaling term $\mathrm{SD}[\hat{\theta}]$ is unknown, therefore we replace it by its estimate $\mathrm{SE}[\hat{\theta}]$.

[15] $1.96 \simeq 2$, therefore it is common to read that statistically significant estimates are at least two standard errors from zero.

[16] Of course, we sample with replacement, to get samples of the same size n.

+ It does not assume any underlying distribution of the data.
+ It can be applied to any sample statistic.
+ Bootstrap CIs are asymptotically consistent (though we can't know the true CI) and more accurate than the traditional intervals.
− Inference still relies on an appropriately drawn sample; and assumes independent resamples. Therefore with structured models, one must think carefully about the design of the resampling procedure (e.g. with clusters: should we sample within or across clusters?).
− Simple but time-consuming.

3.3 Hypothesis testing

3.3.1 Statistical tests

A statistical test is a method of verifying a statistical hypothesis.

A statistical hypothesis is a hypothesis on the probability distribution of T, where T is a **test-statistic** computed from the data, whose probability distribution is connected to our research question.

The general approach to conducting a statistical test consists of the following steps:

1. Write the null hypothesis H_0 — the hypothesis to nullify.
2. Design a test statistic T that summarizes the data's deviation from what would be expected under H_0, and that has a specific distribution under H_0. Ex:
   - an F-test is any test in which the test statistic has an $\mathcal{F}$ distribution under H_0.
   - a t-test is any test in which the test statistic has a Student's $\mathcal{T}$ distribution under H_0;[17]
   - a Wald Chi-squared test is a test in which the test statistic has an *asymptotic* χ^2 distribution under H_0;
   - a z-test is any test in which the test statistic has an *approximately* normal distribution under H_0.
3. Compute the realized value of T for our data: T_{obs}.
4. Look whether it falls in the tails of the distribution. That would mean it is very unlikely given H_0. Therefore we can reasonably reject H_0.

3.3.2 Null Hypothesis Significance Testing (NHST) paradigm

Our goal is to statistically test the **hypothesis of a relationship between y and x_j**, i.e., that $\beta_j \neq 0$. Null hypothesis testing proceeds by *reductio ad absurdum*: a hypothesis is assumed valid if its counterclaim is highly implausible. We'll test whether $\beta_j = 0$ is highly implausible.

For simplicity, let's assume we are estimating a linear model by OLS, and are using the regular t-test.

[17] t-tests are commonly applied for test statistics that would follow a normal distribution if the value of their scaling term (in our case of interest: the standard deviation of the coefficient estimate) were known. When the scaling term is unknown and is replaced by an estimate based on the data (in our case of interest: the standard error of the coefficient estimate), these test statistics follow a Student's $\mathcal{T}$ distribution — under certain conditions.

1. Write H_0 — we define the null hypothesis $H_0 : \beta = 0$
2. Design T — we define the t-statistic $T \equiv \dfrac{\hat{\beta} - \beta_0}{\widehat{\text{SD}}[\hat{\beta}]} = \dfrac{\hat{\beta} - 0}{\text{SE}[\hat{\beta}]}$. *If errors are normal,* $T \underset{H_0}{\sim} t_{n-k}$.[18]
3. Compute T_{obs} — $T_{\text{obs}} \equiv T(\text{observed data})$
4. Interpret — we define the *2-tailed*[19] *p-value* $\equiv \text{P}\left(\text{observing a } |T| \geq |T_{\text{obs}}| \mid H_0\right)$, i.e., the probability of observing data as extreme as that actually observed, assuming H_0.[20]

p-value small $\iff$ T_{obs} falls in the tail of the Student's t-distribution
$\iff$ observing our T_{obs} under H_0 is highly unlikely
$\implies$ reject H_0
$\implies$ there is a relationship between y and x.

In econometrics, the standard approach is to dichotomize the evidence using a p-value threshold, usually the *significance level* $\alpha = 5\%$. $\hat{\beta}$ is "statistically significant" iff $p \leq 0.05$, i.e., there is less than a 5% chance of observing the effect size that was observed if there was in fact no effect.

3.3.3 Type I/II errors, size and power

A test can lead to two types of mistakes:

- **Type 1 error** or *false positive*: $\{- \mid H_0\}$ reject H_0 when shouldn't... *(overconfident)*
- **Type 2 error** or *false negative*: $\{+ \mid \cancel{H_0}\}$ don't reject H_0 when should *(overcautious)*

We define a test's:

- **size** α_T = probability of erroneously rejecting H_0 $\equiv \text{P}\big(\text{type 1 error}\big) = \text{P}(- \mid H_0)$
- **power** κ_T = probability of correctly rejecting H_0 $\equiv 1 - \text{P}\big(\text{type 2 error}\big) = \text{P}(- \mid \cancel{H_0})$

Intuitively, we would like to minimize the size and maximize the power of our test. To guarantee $\alpha_T \leq 0.05$, we simply set the significance level $\alpha = 0.05$. To guarantee $\kappa_T \geq 0.80$, we need a sufficiently large sample zize N, or the "Minimum Detectable Effect" will be very high.[21]

[18] This is a very strong assumption! And it means that if errors are far from normal, the result of the t-test has no interpretation...

[19] We can actually use the test statistic T to carry out two different tests:
– A two-tailed test: if we want to test for the possibility of the relationship in both directions. $H_0 : \beta_j = 0, H_a : \beta_j \neq 0$. Both tails of T's distribution constitute therefore the "critical region", each containing $\frac{\alpha}{2}$ of the values. By default, statistical packages report the two-tailed p-values.
– A one-tailed test: to test for the possibility of the relationship only in one direction. E.g.: $H_0 : \beta_j = 0, H_a : \beta_j > 0$. Only one tail of T's distribution makes the critical region, containing α of the values. Only z- and t-tests can accommodate one-tailed tests. F-tests, χ_2-tests... cannot as their distributions are not symmetric.

[20] ⚠ The p-value is often misinterpreted to be the probability of the null hypothesis, whereas it is the probability of the data, given the null. *p-value* $= \text{P}(\text{obs} \mid \text{hyp}) \neq \text{P}(\text{hyp} \mid \text{obs})$. I.e., if we want to make inferences about the actual values of parameters, p-values and frequentist regression fail us: p-values make inferences about the probability of the data, not parameter values. Only Bayesian methods allow us to make inferences about the actual values of the parameters, e.g., to assess the probability of the null. To summarize: a frequentist's conclusion (in the form of a p-value) is a statement considering that data is random and model parameters are fixed, whereas a Bayesian's conclusion (in the form of a credible interval) is a distribution of parameter values that would generate the fixed data.

[21] In hypothesis testing in econometrics, we typically want at least 80% power and a maximum size of 5%. I.e., we accept to incorrectly reject the null a maximum of 5% of the time, and to correctly reject it at least 80% of the time (i.e., 80% of

Power calculations Having adequate power means that if there really is an effect, the empirical strategy and data will enable the test to detect it. Low powered studies will instead "miss" the effect.[22] Post-estimation, it is useful to perform a retrospective design analysis and ask: *"Was my study sufficiently powered?"*, especially if we found a statistically significant non-null effect. But it must be done correctly:
⚠ To estimate the power one must first postulate a 'true' effect size, which can be thought of as that observed in an infinitely large sample. That effect size should be determined from a literature review, not the effect size observed in one's study! The latter is noisy, and generally overestimated (publication bias), and would therefore lead to overestimates of power.

3.3.4 Criticisms of the NHST and 'statistical significance'

The 2-way binary approach to statistical hypothesis testing, based on the NHST falsificationist paradigm (where the underlying truth is H_0 "no effect" or H_a "effect") and the measured outcome is a binary statement of 'statistical significance' from a p-value threshold, is heavily criticized. It is argued that:

- The underlying reality is not a simple Yes or No: in social sciences, the null hypothesis of zero effect (i.e., conditional independence of y and T given x) is generally implausible — there are virtually no true zeros — and thus uninteresting. The null model is very false, so we are very likely to reject it with enough data.
  - ↪ Instead, we need to find alternatives to thinking in terms of conditional independence in order to study causality. The idea would be to estimate these dependences directly, rather than modeling the world in terms of conditional independence and estimating this structure through the testing of null effects.
- Interpreting p-values dichotomously loses a lot of information.
  - ↪ Instead, one could interpret p-values continuously, the strength of evidence for H_0 being a continuous function of the p-value.
- Interpreting p-values dichotomously may induce selection bias: to be publishable, estimates must be 'significant', i.e., more than two standard errors away from 0; which selects for overestimates.

studies conducted with a given sample size will correctly reject the null). 95% > 80%: econometrics is more focused on avoiding overconfidence than worried about being overcautious. Note also that having a high sample size n is not sufficient to have higher statistical power — empirical studies have actually found zero or weak correlations between the two. The power of a study depends indeed on the sample size, the true size of the effect, measurement variance, and the number of comparisons performed. Note finally that studies of small effects, although potentially important, are unlikely to be statistically significant because they have insufficient power to detect the magnitudes of effects.

[22] Lacasse et al. (2020) is a good example of this. Rephrasing their specific independent and dependent variables as generic x and y: *"Because enrollment in the trial was stopped before we had reached our proposed sample size, the trial was underpowered, with the consequence of a wide confidence interval around the point estimate. [..] The data that were accrued could not rule out benefit or harm from x."* As summarized in the abstract: *"Our underpowered trial provides no indication that x has a positive or negative effect on y"*.

4 Statistical inference [under a Bayesian approach]

As aforementioned, one of the main distinguishing features of Bayesian inference is the expression of all information, including uncertainty, using probability. From this paradigm stem new possibilities at every step of inference.

4.1 Steps of Bayesian inference

We start with the same situation: consider a population parameter of unknown true value θ, and an *event* $\theta = \tilde{\theta}$ (i.e., θ taking this specific value). We are interested in estimating θ using data.

(0) Definition of θ: whereas in frequentist inference θ is considered fixed, now it is considered a *random variable*. This means that at all times it has a probability distribution $\mathcal{P}$ which represents our state of belief about its actual value.

(1) Prior to observing data, this $\mathcal{P}$ is θ's "prior distribution" and represents our prior belief about θ.

(2) Estimation: as more evidence (data) becomes available, we use Bayes' Theorem to update probability statements about θ, which results in a *posterior* distribution:

$$\underbrace{f(\theta \mid \text{data})}_{\text{posterior density}} = \frac{\overbrace{f(\theta)}^{\text{prior density}} \; \overbrace{f(\text{data} \mid \theta)}^{\text{likelihood}[23]}}{\underbrace{f(\text{data})}_{\substack{\text{scaling factor or} \\ \text{"evidence"}}}}$$

In the canonical setup of a regression of y on x, the estimation step precisely consists of combining the model, data, and prior through Bayes' theorem, which is is applied to each possible value of θ to compute a posterior distribution of θ:

- $y = \{y_1, ...y_n\}$ the observed sampled values of the outcome variable of interest Y
- θ a parameter of y's distribution
- α a hyperparameter of θ's distribution

$$\underbrace{\mathrm{P}(\tilde{\theta} \mid y, \alpha)}_{\text{posterior}} = \frac{\mathrm{P}(\tilde{\theta}, y \mid \alpha)}{\mathrm{P}(y \mid \alpha)} = \frac{\mathrm{P}(y \mid \tilde{\theta}, \alpha)\, \mathrm{P}(\tilde{\theta} \mid \alpha)}{\mathrm{P}(y \mid \alpha)} \propto \overbrace{\mathrm{P}(y \mid \tilde{\theta}, \alpha)}^{\text{likelihood}} \; \underbrace{\mathrm{P}(\tilde{\theta} \mid \alpha)}_{\text{prior}}$$

(3) Inference: we can summarize our updated belief about θ from the posterior distribution, by reporting:
- a measurement of central tendency (e.g., the mean or median);
- a 95% credible interval, s.t. "after seeing the data, there is a 95% chance that this CI contains the true θ."

If the goal were prediction rather than inference, we would proceed similarly up to the last step; then:

(3') Prediction: we propagate the uncertainty in θ into the predictions of new data points $\{\tilde{y}\}$ by using simulations: we repeatedly draw from the posterior a value of θ and compute a new data point $\tilde{y}$, thereby creating a posterior predictive distribution.

Predictive distributions of $\tilde{Y}$:

- Before observing data: *prior* predictive distribution: $\mathrm{P}(\tilde{Y} = \tilde{y}) = \int p(\tilde{y}|\theta)\; p(\theta)\; d\theta$

[23] Note that the likelihood is now $f(D|\theta)$ instead of $f(D)$, as θ is no longer fixed. Or, equivalently, the (conditional) probability model is $f(y|\theta,\mathrm{x})$ instead of $f(y|\mathrm{x})$.

- After observing data: *posterior* predictive distribution: $\mathrm{P}(\tilde{Y} = \tilde{y} \mid Y = y) = \int p(\tilde{y}|\theta, y)\ p(\theta|y)\ d\theta = \int p(\tilde{y}|\theta)\ p(\theta|y)\ d\theta$

Advantages/Distinctions of Bayesian inference w.r.t. frequentist inference

- *Intuitive interpretation of findings*
  Because uncertainty is encoded probabilistically, our uncertainty about θ after observing the data is represented by a distribution of values. We can effortlessly compute a 95% credible interval from the posterior, with an intuitive interpretation: "There is a 95% chance that this CI contains the true θ."
  In comparison, the frequentist 95% confidence interval refers to our uncertainty about the *sampling method* — not θ — and is thereby interpreted in terms of repeated samples: *"Imagine we drew all possible random samples of size n. This interval would contain the true θ in 95% of the samples."*
  More generally, the Bayesian framework enables us to actually answer questions like "What is the probability that $\theta = 4$?" Whereas the Frequentist framework produces convoluted estimates: the probability of the data assuming that $\theta = 4$.
- *Including prior information*
  In Bayesian inference emerges a compromise between prior information and data. More generally, it is a way to include multiple sources of information.
- *Making predictions is facilitated by the computation being ~~optimization-~~ simulation-based*
  A Bayesian will argue that one should do predictions based on the whole posterior distribution of possible coefficient values, while prediction based on point estimates disregards all infomation about how imprecise the point estimate is. Because the uncertainty about θ is encoded in a probability distribution, we can simply propagate this uncertainty into predictions of a new data point $\tilde{y}$, by using simulations. We draw a value from θ's posterior distribution and make a *probabilistic* prediction of a new data point $\tilde{y}$ for this value of θ, and repeat this simulation S times. The S resulting values make the *posterior predictive distribution* $f(\tilde{y}|Y,\alpha) = \int_S f(\tilde{y}|\theta)\ f(\theta|Y,\alpha)\ d\theta$.

Bayesian inference is the discipline of updating our belief about the world based on further observation of the world. Whereas frequentist inference is focused on summarizing the information in the data. These summaries of data have known statistical properties but have limited value as predictions.

4.2 Choosing θ's prior distribution

We include additional information using a prior distribution

- Using an uninformative or "flat" prior (the uniform distribution) results in the posterior distribution being equal to the product of the likelihood and a mere constant, s.t. the mode of the posterior distribution is the ML or LS estimator.
- Weakly Informative Priors: "What you should be doing when you think you want to use noninformative priors." `https://statmodeling.stat.columbia.edu/2009/05/24/handy_statistic/`. Ex: The R function `rstanarm::stan_glm()` adjusts the default priors based on the scale of the variables in the model.
- "Conjugate" prior probability distributions (for the ... distribution): the posterior distributions $f(\theta|x)$ are in the same family as the prior probability distribution $f(\theta)$.
- Bayesian inference is a compromise between prior and data, where each has a weight proportional to the inverse square of its s.e. $\rightarrow \mathrm{SE}_{\mathrm{Bayes}} <$ both $\mathrm{SE}_{\mathrm{prior}}$ and $\mathrm{SE}_{\mathrm{data}}$

4.3 Estimating θ's posterior distribution

- When the likelihood has a known analytical form, we can combine it with the prior to derive the posterior analytically.
- Most of the time, there is no such analytical form. To estimate the posterior distribution, we can use Markov Chain Monte-Carlo (MCMC) algorithms: a family of iterative sampling algorithms[24] that sample simulation draws to form an empirical distribution which approximates the posterior:
  - *"Monte-Carlo"* refers to the practice of estimating the properties of a distribution by examining random samples from the distribution. Ex: instead of finding the mean of a normal distribution by directly calculating it from the distribution's equations, we would draw random samples from the normal distribution and calculate the sample mean.
  - *"chain"* means that the random samples are generated by a special sequential process: each random sample is used as a stepping stone to generate the next random sample.
    Note that this means that the draws are not independent.
  - The *"Markov"* property of the chain is that, while each new sample depends on the one before it, new samples do not depend on any samples before the previous one.

$\longrightarrow$ Most of the time, **fitting a Bayesian model = generating a set of posterior simulations** (representing different possible values of the parameter vector θ), which we typically summarize using its median, its median absolute deviation (a more robust estimator of scale than the standard deviation), and uncertainty intervals.

Examples: We start by taking S independent samples from $p(\theta|y)$, from which we compute a sequence which will tend to the desired distribution as $S \to \infty$:

- To approximate the distribution of a function $g(\theta)$:

$$\begin{cases} \theta^1 \sim p(\theta|y) & \longrightarrow \text{compute } g\left(\theta^1\right) \\ \dots & \\ \theta^S \sim p(\theta|y) & \longrightarrow \text{compute } g\left(\theta^S\right) \end{cases} \quad \text{The sequence} \quad \begin{Bmatrix} g\left(\theta^1\right) \\ \dots \\ g\left(\theta^S\right) \end{Bmatrix} \xrightarrow[S\to\infty]{} p(g(\theta) \mid y)$$

- To approximate $\mathrm{P}(\theta_1 > \theta_2|y)$:

$$\begin{cases} \theta_1^1 \sim p(\theta_1|y) \quad \text{and} \quad \theta_2^1 \sim p(\theta_2|y) \\ \dots \\ \theta_1^S \sim p(\theta_1|y) \quad \text{and} \quad \theta_2^S \sim p(\theta_2|y) \end{cases} \quad \text{Then } \frac{1}{S}\sum_s \mathbb{1}\{\theta_1^s > \theta_2^s\}$$

- To approximate the posterior predictive density $\mathrm{P}(\tilde{Y}|Y = y)$:

$$\begin{cases} \theta^1 \sim p(\theta_1|y) & \longrightarrow \text{compute } \tilde{y}^1 \sim p\left(\tilde{y}|\theta_1\right) \\ \dots & \\ \theta^S \sim p(\theta_1|y) & \longrightarrow \text{compute } \tilde{y}^S \sim p\left(\tilde{y}|\theta_S\right) \end{cases} \quad \text{The sequence} \quad \begin{Bmatrix} \tilde{y}^1 \\ \dots \\ \tilde{y}^S \end{Bmatrix} \xrightarrow[S\to\infty]{} p\left(\tilde{y}|y\right)$$

[24] For example, Stan uses as inference algorithms two MCMC algorithms: the Hamiltonian Monte Carlo algorithm and its adaptive variant the "no-U-turn sampler".

5 Prediction

Prediction isn't part of statistical inference, but it can be the ultimate research goal, motivating the initial statistical inference step. Whether the ultimate goal is inference or prediction,[25] both first require finding a model that describes the relationship between the independent variables and the outcome in our data. The use of the resulting model then differs:

- Inference: Use the model to learn about the data generating process.
- Prediction: Use the model to predict the outcomes for new data points.

[25] Note: In machine learning, the term inference is sometimes used instead to mean "making a prediction, by evaluating an already trained model". In this context, inferring properties of the model is referred to as training or learning (rather than inference), and using a model for prediction is referred to as inference (instead of prediction).

6 Model comparison

Learning from data has generally one of two ultimate objectives: inference or prediction. Model comparison should proceed in line with the objective. After a brief paragraph on *nested* model discrimination, this section focuses on model comparison for prediction, our objective will therefore be predictive performance.[26] Much of this section is taken from Gelman et al. (2014).

6.1 Comparing nested models: F tests

If two models are *nested*, i.e., one represents a special case of the other, we can easily discriminate between them using a standard hypothesis test of the parametric restrictions on the nested one.

The key questions are: (1) is the improvement in fit large enough to justify the additional difficulty in fitting, and in a Bayesian context (2) is the prior distribution on the additional parameters reasonable?

6.2 Comparing non-nested models: IC, CV

We want to know which model gives the best predictions of new data generated from the true DGP. Ideally, we would measure the model's out-of-sample predictive accuracy or error, for such new data produced from the true DGP. After describing exactly what the quantity we would like to measure is, we will describe methods for estimating an *approximation* of it, given the data we have.

There are different ways of defining a model's predictive accuracy or error:

- If one is predicting a *point*, predictive accuracy can be defined using an error measure, such as the absolute error or the squared error. Individual errors are aggregated and averaged to obtain a summary measure of predictive accuracy, such as the Mean Absolute Error (MAE) or the Root Mean Squared Error (RMSE):[27]

$$MAE \equiv \frac{1}{N}\sum_i |\,\hat{y}_i - y_i\,|\,, \qquad RMSE = \sqrt{MSE} \equiv \sqrt{\frac{1}{N}\sum_i (\hat{y}_i - y_i)^2}$$

- A more general[28] summary is the *log likelihood* or *log predictive density* (LPD).
  For any data $\mathrm{y} = y_1, ..., y_m$ produced from the true DGP, i.e., taken from the *unknown* data distribution f, $\mathrm{LPD(y)} \equiv \ln \mathrm{P(y}|\theta) = \ln \prod_i \mathrm{P}(y_i|\theta)$.

  Therefore for *out-of-sample* data:

If inference for θ is summarized by a point estimate $\hat{\theta}(y)$	If inference for θ is summarized by a posterior distribution $p_{post,\theta}()$
▷ For a new data point $\tilde{y}_i \sim f$:	
$\mathrm{LPD}(\tilde{y}_i) = \ln \mathrm{P}(\tilde{y}_i\|\hat{\theta})$	$\mathrm{LPD}(\tilde{y}_i) = \ln p_{post,y}(\tilde{y}_i) \equiv \int \mathrm{P}(\tilde{y}_i\|\theta) p_{post,\theta}(\theta) d\theta$
▷ As new data points are themselves unknown, the expectation:	
$\mathrm{ELPD} \equiv \mathbb{E}_f\big[\mathrm{LPD}(\tilde{y}_i)\big] = \mathbb{E}_f\big[\ln \mathrm{P}\big(\tilde{y}_i\|\hat{\theta}\big)\big]$	$\mathrm{ELPD} \equiv \mathbb{E}_f\big[\mathrm{LPD}(\tilde{y}_i)\big] = \mathbb{E}_f\big[\ln p_{post,y}\,(\tilde{y}_i)\,\big]$

[26] In classical econometrics focused on inference, especially when the goal is causal inference, the research design drives the model specification such that there isn't so much need for model comparison and selection.

[27] The RMSE is the standard deviation of the residuals, i.e., of the unexplained variation. It is an absolute measure of fit of the model to the data. (Whereas R^2 is a relative measure of fit. Note that one should absolutely not select a model based on R^2, as this would favor overfitting.) Note that: (1) RMSE is scale-dependent (it has the same unit as y), therefore it can only be compared across models in the same units; (2) compared to the MAE, the RMSE penalizes large errors more.

[28] It is proportional to the MSE if the model is normal.

In practice, f and θ are unknown, so we can't compute ELPD. We will try to approximate it, using existing data (hence knowing that any method will be correct at best only in expectation...).

- **Adjusted within-sample predictive accuracy:** a natural estimate of the expected log predictive density for *new* data is the log predictive density for *existing* data. **Information criteria** such as AIC and WAIC give approximately unbiased estimates of ELPD by correcting for how much the fitting of k parameters increases predictive accuracy, by chance alone. These are scoring methods from information theory.
- **Cross-validation:** the model is fit to a training set, then the fit evaluated on a holdout set.

Both methods are based on adjusting the log predictive density of the observed data by subtracting an approximate bias correction. The measures differ in their starting points (how they measure the log predictive density) and their adjustments. Asymptotically, AIC is equal to LOO-CV computed using ML estimation, and Bayesian LOO-CV is equal to WAIC.

Information Criteria (IC)

Goal: we want the best model fit (maximized likelihood), but we penalize model complexity (to not overfit the data). Most IC are expressed on the deviance scale; the model with smallest IC is preferred.[29]

Let k be the number of parameters, n the sample size.

- **Akaike information criterion (AIC)**
  - starting point: the log predictive density, conditional on a point estimate: $\ln \hat{\mathcal{L}} \equiv \ln \mathrm{P}(y|\hat{\theta})$;
  - adjustment for overfitting: uses the simplest bias correction, based on the asymptotic normal posterior distribution, for which[30] simply subtracting k corrects for the number of parameters:

$$AIC \equiv -2\left(\widehat{\mathrm{ELPD}}_{\mathrm{AIC}}\right) = -2\left(\ln \hat{\mathcal{L}} - k\right) = -2\ln \hat{\mathcal{L}} + 2k$$

  AIC_c is the AIC corrected for small samples: $AIC_c = -2\ln \hat{\mathcal{L}} + 2k\dfrac{n}{n-k-1} \xrightarrow[n\to+\infty]{} AIC$

  Limit: when we go beyond linear models with flat priors, e.g., models with hierarchical structures or informative priors, the number of effective parameters isn't k so we can't simply substract k.

- **Watanabe-Akaike information criterion (WAIC)**
  - starting point: the log predictive density, averaging over the posterior distribution $p_{\mathrm{post}}(\theta) = \mathrm{P}(\theta|y)$ (i.e., a fully Bayesian approach);
  - adjustment for overfitting: corrects for the *effective* number of parameters.

Cross-validation (CV)

Cross-validation consists in partitioning the data into a training set y_t and a validation set y_v, fitting the model to the training set, and evaluating this predictive accuracy (fit) using the validation set. It is based on the log predictive density, but can use any starting point (i.e., either averaging over the posterior distribution $p_{\mathrm{post}}(\theta)$ or conditioning on a point estimate $\hat{\theta}$).

[29] For models with different fixed effects, residual likelihoods are not comparable. Therefore if such models were fitted using restricted maximum likelihood (REML), IC cannot be used to select between them. To use IC, the models should be fitted using maximum likelihood.

[30] This is also true in the special case of a normal linear model with a uniform prior distribution.

In Bayesian CV, fitting the model to y_t yields a posterior distribution for θ: $p_{\text{post}}(\theta) \equiv \mathrm{P}(\theta|y_t)$. We assume we can summarize it by S simulation draws $\theta^1, ..., \theta^S$. We can then compute the log predictive density for y_v as: $\mathrm{LPD}(y_v) \equiv \ln \mathrm{P}(y_v|\theta^{\text{“post”}}) \equiv \frac{1}{S}\sum_{s=1}^{S} \mathrm{P}(y_v|\theta^s)$

The CV process is repeated using different partitions, and the resulting log predictive densities are averaged into a single estimate of out-of-sample predictive accuracy.

- **K-fold CV**
  The data are randomly partitioned into K equal-sized sets. $K = 10$ is commonly used. The CV process is repeated K times, each time using one subsample for validation — such that each observation is used for validation exactly once — and the K results are averaged into one estimate:

$$\mathrm{LPD}_{K\text{-CV}} = \sum_{k=1}^{K} \ln\left(\frac{1}{S}\sum_{s=1}^{S} \mathrm{P}\big(y_k|\theta^s\big)\right)$$

- ***‘Leave-one-out’* CV = n-fold CV**
  In the extreme case of n partitions, each validation set represents a single data point:

$$\mathrm{LPD}_{\text{LOO-CV}} = \sum_{i=1}^{n} \ln\left(\frac{1}{S}\sum_{s=1}^{S} \mathrm{P}\big(y_i|\theta^s\big)\right)$$

In any CV process, each prediction is conditioned on $n - v$ data points instead of n, which causes underestimation of the predictive fit. We can correct for this bias by estimating how much better predictions would be obtained if conditioning on n data points (Gelman et al., 2014).

Conclusion Neither cross-validation nor information criteria are perfect. AIC does not work in settings with strong prior information, WAIC relies on a data partition unamenable to structured models such as for spatial or network data, cross-validation is computationally expensive as getting a stable estimate requires many data partitions and fits. Gelman et al. (2014)’s preferred choice is *“cross-validation, with WAIC as a fast and computationally convenient alternative. WAIC is fully Bayesian (using the posterior distribution rather than a point estimate) [...]. A useful goal of future research would be a bridge between WAIC and cross-validation with much of the speed of the former and robustness of the latter.”*

TO ADD: Model Shrinkage Methods, and other methods to deal with highly correlated predictors

- LASSO (Least Absolute Shrinkage and Selection Operator)
- PCA

7 Other branches of statistical modeling

7.1 Statistical Inference Using Agent-based models (ABMs)

Agent-Based Models are computational models[a] that simulate the actions and interactions of autonomous agents within a system, to assess their effects on the system as a whole. The goal is to re-create and predict the emergence[b] of higher-level system properties from simple agent-level behaviors, taking a "bottom-up" approach.

ABMs are generally composed of 3 elements:

1. many **agents** with assigned attributes;
2. simple **rules** about: their individual decision-making process, how they interact, how they learn and adapt—these rules can be deterministic or probabilistic;
3. an **environment**.

[a] Computational models are mathematical models that study the behavior of a system by computer simulation. The system studied is often a complex nonlinear system for which simple analytical solutions are not available. Experimentation is therefore done by modifying the model's parameters, and comparing outcomes. Examples include weather forecasting models, flight simulator models, neural network models, and ABMs.

[b] The process of *emergence* can be expressed as "the whole is greater than the sum of its parts".

Goal of ABMs ABMs allow us to observe how the behaviors of individual agents affect the system as a whole and if any emergent structure develops within the system. They show how small-scale changes can affect large-scale outcomes within the system.

At a formal level, an ABM is just a statistical model. But agent-based modeling differs from other types of statistical modeling because it describes only the behavior of the agents in a system, rather than global properties of the system.

Use in different fields

- In economics: ABMs can describe the microeconomic actions of adaptive agents, which give rise to emergent behavior in the form of macroeconomic structures; which, in turn, influence agent decisions. Ex: we can represent the economy as a complex system, with crashes and booms that emerge from non-linear responses to small changes.
- In ecology: ABMs are often called individual-based models (IBMs), and are used to study population dynamics, plant-animal interactions...
- In epidemiology: epidemiological ABMs now complement traditional compartmental models (such as the deterministic SIR — Susceptible/Infectious/Recovered — model) which they have tended to surpass in terms of prediction accuracy to model the spread of epidemics.

Statistical inference

1. Model validation and selection, uncertainty quantification, and fitting ABMs to data: There does not seem to be (yet) formal guidelines and procedures from the statistical literature, for: fitting ABMs to data, for making quantified statements of uncertainty about the outputs, e.g., calculating confidence intervals on predictions, nor for testing whether a specific parameter (rule) is needed in an ABM. See Banks and Hooten (2021); Heard et al. (2015).
2. Statistical inference
   Because of the variety of input rules and the complexity of outputs, the likelihood function of an ABM

is generally intractable. One must hence perform likelihood-free inference. Heard et al. (2015) suggest that two main tools allow that: emulators and approximate Bayesian computation (ABC).

Key ideas

Statistics is about reasoning under uncertainty, and therefore **probability distributions**. Inferential statistics proceeds by learning from data, it asks: *given sample data, what are we able to infer about the population?*

In microeconometrics, inference is usually conducted under a frequentist approach:

Steps	Options
1. Choose & write a model, the one we think is closest to the true and unobserved DGP.	*(linear regression model w. normal errors, logistic regression model, SEM...)*
⋆ **Bring in data** ⋆	
2. Estimate the model, i.e., estimate the conditional distribution. When the specification is parametric, it means estimating parameters.	*(OLS, 2SLS, ML,...)*
a. estimation $\Longrightarrow$ "$\hat{\beta}$ = ..."	
b. hypothesis testing $\Longrightarrow$ "$\hat{\beta}$ is/isn't statistically significant"	
3. Validate & compare the model.	

In frequentist statistics, we trust that the results given by these statistical tools (estimators, tests...) give us relevant indications about the population, because of the tools' asymptotic properties (which stem from laws of large numbers (LLNs) and central limit theorems (CLTs)).

References

Banks, D. L. and Hooten, M. B. (2021). Statistical Challenges in Agent-Based Modeling. *Am. Stat.*, pages 1–8, DOI: 10.1080/00031305.2021.1900914.

Bell, A., Fairbrother, M., and Jones, K. (2019). Fixed and random effects models: making an informed choice. *Quality & quantity*, 53(2):1051–1074, ISSN: 0033-5177, DOI: 10.1007/s11135-018-0802-x.

Fezzi, C. and Bateman, I. (2015). The Impact of Climate Change on Agriculture: Nonlinear Effects and Aggregation Bias in Ricardian Models of Farmland Values. *Journal of the Association of Environmental and Resource Economists*, 2(1):57–92, ISSN: 2333-5955, DOI: 10.1086/680257.

Gelman, A. and Greenland, S. (2019). Are confidence intervals better termed "uncertainty intervals"? *BMJ*, 366, DOI: 10.1136/bmj.l5381.

Gelman, A. and Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, ISBN: 9781139460934.

Gelman, A., Hwang, J., and Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Stat. Comput.*, 24(6):997–1016, DOI: 10.1007/s11222-013-9416-2.

Heard, D., Dent, G., Schifeling, T., and Banks, D. (2015). Agent-based models and microsimulation. *Annual Review of Statistics and Its Application*, 2(1):259–272, DOI: 10.1146/annurev-statistics-010814-020218.

Horrace, W. C. and Oaxaca, R. L. (2006). Results on the bias and inconsistency of ordinary least squares for the linear probability model. *Econ. Lett.*, 90(3):321–327, DOI: 10.1016/j.econlet.2005.08.024.

Lacasse, Y., Sériès, F., Corbeil, F., Baltzan, M., Paradis, B., Simão, P., Abad Fernández, A., Esteban, C., Guimarães, M., Bourbeau, J., Aaron, S. D., Bernard, S., and Maltais, F. (2020). Randomized trial of nocturnal oxygen in chronic obstructive pulmonary disease. *New England Journal of Medicine*, 383(12):1129–1138, DOI: 10.1056/NEJMoa2013219.

Mertens, W., Pugliese, A., and Recker, J. (2017). Analyzing Longitudinal and Panel Data. In *Quantitative Data Analysis: A Companion for Accounting and Information Systems Research*, pages 73–98. Springer International Publishing, Cham, ISBN: 978-3-319-42700-3.

Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R, Second Edition*. CRC Press/Taylor & Francis Group, ISBN: 9781498728331.

Wooldridge, J. M. (2013). *Introductory Econometrics: A Modern Approach*. Cengage Learning, fifth edition, ISBN: 9781111531041.

A A small library of regression models

The textbook Classical Linear Regression Model (CLRM) can be generalized in various dimentions, such as:
– the power of the independent variables (→ polynomial regression);
– the link function relating the linear predictor $\mathrm{x}\beta$ to the outcome $\mathbb{E}[y|\mathrm{x}]$ (→ generalized linear model);
– the number of levels in the data (→ multilevel model)...
This section presents some of the models resulting from these generalizations.

Notation Recall that a statistical model is the combination of a sample space and a collection of *joint probability distributions* on that space; the goal being to represent the specific distribution induced by the DGP. Rather than look at the full joint distribution, regression models simplify the problem and focus on the *conditional distribution* of $y|\mathrm{x}$.[31] All regression models are therefore first *conditional distributions*, and can be written as such. Based on the properties of each distribution, we can then also write them in a *conditional mean +/× error* form.

A.1 Expanding from the CLRM

- **Classical Linear Regression Model**

$$y_i|\mathrm{x}_i \sim \mathcal{F}(\mathrm{x}_i'\beta, \sigma^2)$$

$$\iff y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik} + e_i, \quad e_i \overset{\text{iid}}{\sim} \mathcal{F}(0, \sigma^2)$$

$$\iff y_i = \mathbb{E}[y_i|\mathrm{x}_i] + e_i, \quad \mathbb{E}[y_i|\mathrm{x}_i] = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik}, \quad e_i \overset{\text{iid}}{\sim} \mathcal{F}(0, \sigma^2)$$

- **Polynomial Regression**
  – Ex: LOESS (locally estimated scatterplot smoothing) is a nonparametric regression algorithm, in which $\mathbb{E}[y_i|\mathrm{x}_i]$ at each data point i is estimated using a weighted low-degree polynomial regression model that gives higher weights to the neighboring points (along x).

$$\mathbb{E}[y_i|\mathrm{x}_i] = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + ... + \beta_p x_i^p, \quad e_i \overset{\text{iid}}{\sim} \mathcal{F}(0, \sigma^2)$$

- **Generalized linear model (GLM)**

  GLMs are often used to predict outcomes of bounded or discrete form (outcomes that cannot be fit well with normally distributed additive errors). A GLM consists of three elements: a probability distribution $\mathcal{F}()$ from the exponential family we assume the outcome to be generated from,[32] a linear predictor $\mathrm{x}_i'\beta$, and an invertible link function $g()$ that relates $\mathbb{E}[y_i|\mathrm{x}_i]$ to $\mathrm{x}_i'\beta$.

  – Ex: the linear regression model is a GLM with normal outcome data and identity link.
  – Ex: the logistic regression model is a GLM with Bernouilli outcome data and logit link.
  – Ex: the Poisson regression model is a GLM with Poisson outcome data and log link.

$$y_i|\mathrm{x}_i \sim \mathcal{F}_{\text{ExpFamily}(...)}, \quad g\big(\mathbb{E}[y_i|\mathrm{x}_i]\big) = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik}$$

[31] For simplicity, we assume we adopt a frequentist approach, therefore we need not write distributions of y as conditional on θ, as θ is fixed. If we adopted a Bayesian approach, we would make it explicit that distributions of y are conditional on θ.

[32] Distributions in the exponential family have a probability density (or mass) function $f_\theta(y)$ whose form make them highly tractable mathematically. In particular, we can obtain general expressions for their mean and variance in terms of their canonical parameters θ using differentiation. See Wood (2017), section 3.1, on how this allows to develop a general method for fitting a GLM by maximum likelihood, as θ is ultimately determined by the regression parameters.

- **Generalized additive model (GAM)**

  GAMs generalize further to allow for $g\big(\mathbb{E}[y_i|\mathrm{x}_i]\big)$ to depend linearly on smooth (nonlinear) functions of some predictors. Relationships with $x_{i1}, ..., x_{ik}$ are represented via smooth functions $h_1(.), ..., h_k(.), l_j(.,.), ...$, which allow to estimate flexible non-linear relationships and interaction effects. GAMs can thereby be considered semi-parametric models. They have an additional aspect: they are fitted using penalized estimation. A penalty for the size of the coefficients for the smooth functions is added to the objective function (whether the likelihood or a loss function), to prevent overfitting the data.

$$y_i|\mathrm{x}_i \sim \mathcal{F}_{\text{ExpFamily}(...)}, \qquad g\big(\mathbb{E}[y_i|\mathrm{x}_i]\big) = \beta_0 + h_1(x_{i1}) + ... + h_k(x_{ik}) + l(x_{i1}, x_{i4})$$

- **Multilevel or "hierarchical" models**

  The lowest-level model is a regression, higher-level models model coefficients of the model immediately below them. These higher-level models can be regressions or distributions. All models are fitted simultaneously.

  – Ex: 2-level, varying-intercept model; the group-level model is a regression:

$$\begin{cases} y_i \sim \mathcal{F}\left(\alpha_{j[i]} + \beta \mathrm{x}_i,\ \sigma_y^2\right) & \forall i = 1, ..., n, \quad j = 1, ..., J \\ \alpha_j \sim \mathcal{F}\left(\gamma_0 + \gamma_1 \mathrm{w}_j,\ \sigma_\alpha^2\right) & \forall j = 1, ..., J \end{cases}$$

  – Ex: 2-level, varying-intercept & slope model; the group-level models are distributions:

$$\begin{cases} y_i \sim \mathcal{F}\left(\alpha_{j[i]} + \beta_{j[i]} \mathrm{x}_i,\ \sigma_y^2\right) & \forall i = 1, ..., n, \quad j = 1, ..., J \\ \alpha_j \sim \mathcal{N}\left(\mu_\alpha, \sigma_\alpha^2\right) & \forall j = 1, ..., J \\ \beta_j \sim \Gamma\left(\gamma_\beta, \delta_\beta\right) & \forall j = 1, ..., J \end{cases}$$

- **Incomplete data models**

  – Missing data. For some problems, we can set up a model specifically to handle the missingness mechanism. Ex censored data: extensions of ML / Bayesian regression include the censoring into the likelihood.

  – Measurement error in the predictors x:[33] we observe $x^* = x + \eta$. If we can estimate the variance of the measurement errors, we can either just apply a bias correction on the raw estimate from the regression of y on x^*, or directly fit the full "simultaneous-equation model" using a marginal likelihood or Bayesian approach. Same maths as in IV.

A.2 Generalized linear models (GLMs)

Generalized linear models (GLMs) are often used to predict outcomes of limited form, i.e., that are categorical or constrained to fall in a certain range. With such data, linear regression estimation is not appropriate as it does not take into account the constraint on values of the dependent variable. The strategy is to transform the limited y into a continuous, real-valued variable $y' \equiv g(y) \in (-\infty, \infty)$, that we can then model as $y' = \mathrm{x}\beta + \varepsilon$, using a link function $g()$.

A GLM consists of three elements: a probability distribution $\mathcal{F}()$ from the exponential family for the outcome, a linear predictor $\mathrm{x}_i'\beta$, and an invertible link function $g()$ that relates $\mathbb{E}[y_i|\mathrm{x}_i]$ to $\mathrm{x}_i'\beta$:

$$y_i|\mathrm{x}_i \sim \mathcal{F}_{\text{ExpFamily}(...)}, \qquad g\big(\mathbb{E}[y_i|\mathrm{x}_i]\big) = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik}$$

[33] Measurement error in y does not pose a problem besides imprecision, as it just goes into the error term. It is measurement error in x that poses a problem: estimated regression coefficients can be attenuated (i.e., it doesn't just increase standard errors, but can drive the coefficient down).

Limited y	Appropriate regression models
binary: $y \in \{0,1\}$	probit regression, logit regression
count: $y \in \{0,1,2,3,...\}$	Poisson regression, negative binomial regression
interval: $y \in [0,1]$	fractional response
censored	censored regression, e.g., Tobit

A.2.1 Binary outcome models

The outcome variable is binary, i.e., it follows a Bernoulli distribution:

$$y_i|\mathrm{x}_i \sim Ber(\pi) \equiv \begin{cases} 1 & \text{with probability } \pi \\ 0 & \text{with probability } 1-\pi \end{cases}$$

The conditional mean $\mathbb{E}[y_i|\mathrm{x}_i]$ is equal to the conditional probability $\pi \equiv \mathrm{P}(y_i{=}1|\mathrm{x}_i)$.[34] A regression model is therefore formed by expressing π as a function of x_i and β;[35] and we look for a link function $g()$ that maps the [0,1] interval to the real line.

$$y_i|\mathrm{x}_i \sim Ber(\pi_i), \quad \pi_i = \mathbb{E}[y_i|\mathrm{x}_i] = g^{-1}(\mathrm{x}_i, \beta)$$

- **Linear probability model**

  $$y_i|\mathrm{x}_i \sim \mathcal{F}(\pi_i), \quad \pi_i = \mathrm{x}_i'\beta$$

  This model is probably the first one that comes to mind. It is not appropriate, as the identity link is not a CDF, it will not constrain the predicted values to be in [0,1], since the predictor $\mathrm{x}_i'\beta$ can take any real value. Yet, it is still frequently preferred to Logit or Probit, on grounds that it is computationally simpler, the estimated marginal effects are easier to interpret, and are usually very similar anyway, especially with a large sample size.
  However, Horrace and Oaxaca (2006) show that in almost all circumstances, the LPM yields biased and, most importantly, *inconsistent* estimates. I.e., the LPM gives the wrong answer, with almost certainty, even with an infinitely large sample: *"consistency seems to be an exceedingly rare occurrence as one would have to accept extraordinary restrictions on the joint distribution of the regressors. Therefore, OLS is frequently a biased estimator and almost always an inconsistent estimator of the LPM."*

- **Logit model = Logistic regression model**

  $$y_i|\mathrm{x}_i \sim Ber(\pi_i)$$

  $$\pi_i = \mathrm{logit}^{-1}(\mathrm{x}_i'\beta) \equiv \frac{e^{\mathrm{x}_i'\beta}}{1+e^{\mathrm{x}_i'\beta}} \iff \mathrm{logit}(\pi_i) = \mathrm{x}_i'\beta$$

  We choose as link function $g()$ the logit $\mathrm{logit}(.) \equiv \ln\left(\frac{\cdot}{1-\cdot}\right)$ (i.e., we choose as $g^{-1}()$ the CDF of the logistic distribution: $\mathrm{logit}^{-1}()$), which maps $[0,1]$ to $[-\infty,\infty]$. We transform the probability outcome using this logit or "log-odds" transformation. As this new outcome need not be in [0,1], we can model it as a *linear* function of the covariates.

[34] As $\mathbb{E}[y|\mathrm{x}] = 1 \times \mathrm{P}(y{=}1|\mathrm{x}) + 0 \times \mathrm{P}(y{=}0|\mathrm{x}) = \mathrm{P}(y{=}1|\mathrm{x})$.

[35] The function $g^{-1}()$ should be a *cumulative distribution function*, to ensure that $0 \le \pi_i \le 1$.

Interpretation of each coefficient $\hat{\beta}_k$ (keeping all the other predictors fixed):

– logit scale $[-\infty, \infty]$ *"a 1-unit difference in x corresponds to a $\hat{\beta}_k$-unit difference in* log-odds(y=1)*"*

– odds[36] scale $[0, \infty]$ *"a 1-unit difference in x corresponds to a $e^{\hat{\beta}_k}$ multiplicative difference in* odds(y=1)*"*

– probability scale [0,1] *"a 1-unit difference in x corresponds to a $\frac{\hat{\beta}_k}{4}$-unit maximum[37] difference in* P(y=1)*"*

- **Probit model**

$$y_i|\mathrm{x}_i \sim Ber(\pi_i)$$

$$\pi_i = \mathrm{probit}^{-1}(\mathrm{x}_i'\beta) \equiv \int_{-\infty}^{X_i'\beta} \phi(t)dt \iff \mathrm{probit}(\pi_i) = \mathrm{x}_i'\beta$$

We choose as link function $g()$ the probit, which is the quantile function of the standard normal distribution (i.e., we choose as $g^{-1}()$ the CDF of the normal distribution: $\mathrm{probit}^{-1}()$), which maps $[0, 1]$ to $[-\infty, \infty]$.

We cannot interpret each coefficient $\hat{\beta}_k$ directly, we need to *compute* the marginal effects.

Note: As a rule of thumb, probit regression coefficients are roughly equal to logistic regression coefficients divided by 1.6.

Estimation by Maximum Likelihood, as the distribution of the data $y|X$ must be the Bernouilli. The conditional density of each observation is: $f(y_i|\mathrm{x}_i) = \pi_i^{y_i}(1-\pi_i)^{(1-y_i)}$. Given independence over i, the (log-)likelihood of the data is then the (log-)likelihood for n independent Bernoulli observations:

$$\begin{aligned}
\hat{\theta}_{\mathrm{ML}} = \underset{\theta}{\mathrm{argmax}}\ \log \mathcal{L}\big(\mathrm{y}|\mathrm{X},\theta\big) &= \underset{\theta}{\mathrm{argmax}}\ \log \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i} \\
&= \underset{\theta}{\mathrm{argmax}}\ \sum_{i=1}^{n} y_i \ln(\pi_i) + (1-y_i)\ln\big(1-\pi_i\big) \\
&= \underset{\beta}{\mathrm{argmax}}\ \sum_{i=1}^{n} y_i \ln\Big(g^{-1}(\mathrm{x}_i'\beta)\Big) + (1-y_i)\ln\Big(1-g^{-1}(\mathrm{x}_i'\beta)\Big)
\end{aligned}$$

⚠ *Don't fit logistic models for binary outcomes when the underlying continuous variable is available. For inference or prediction, it is much more efficient to model the underlying continuous variable and then map it back to the probability of the discrete outcome. Ex:*
– *basketball game: model the expected score differential, and then map it to* P(winning).
– *elections: predict vote differential and then map that to* P(winning).
– *health: model change in blood pressure, and then convert it to the binary disease state* P(hypertension).

A.2.2 Count data models

$y_i \in \{0, 1, 2, ...\}$: number of occurrences of an event. *Ex: number of children in a household, number of doctor visits per year, number of new cases of an infectious disease per day...*

[36] The odds of success are defined as the ratio of the probability of success π over the probability of failure. Here, where "success" is y=1, the odds of y=1 are $\frac{\pi}{1-\pi}$ to 1.

[37] ⚠ The logistic function $\mathrm{logit}^{-1}()$ is nonlinear, so the expected difference in $\mathrm{P}(y=1)$ from a given difference in x is not a constant along x. We must choose where to evaluate changes, if we want to interpret them on the probability scale. The slope of the logistic regression curve is steepest at its halfway point ($\mathrm{logit}^{-1}() = 0.5$), where it equals $\beta/4$. I.e., the largest change in π from a 1-unit change in x is $\beta/4$.

- **Poisson regression model**

  The Poisson distribution $Pois(\lambda)$ models the number of events occurring in a fixed interval (of time or space), when these events occur at random, independently in time, with the constant mean rate λ. Its probability mass function is therefore $\mathrm{P}(y|\lambda) = \frac{e^{-\lambda}\lambda^y}{y!}$, which further implies $\mathbb{E}[y] = \mathbb{V}[y] = \lambda$. The general Poisson regression model is:

$$y_i|\mathrm{x}_i \sim Pois\big(\lambda_i\big), \quad \lambda_i = g^{-1}(\mathrm{x}_i'\beta)$$

  A common choice of link function $g()$ is $\ln()$. The Poisson regression model is therefore fitted as a log-linear regression with Poisson error distribution: $y_i|\mathrm{x}_i \sim Pois\big(e^{\mathrm{x}_i'\beta}\big)$.

  Estimation by Maximum Likelihood:

$$\begin{aligned}
\hat{\beta}_{\text{ML}} = \underset{\beta}{\text{argmax}}\ \log\mathcal{L}\big(\mathrm{y}|\mathrm{X},\beta\big) &= \underset{\beta}{\text{argmax}}\ \log\prod_{i=1}^{n}\mathrm{P}(y_i|\mathrm{x}_i,\beta) \\
&= \underset{\beta}{\text{argmax}}\ \sum_{i=1}^{n}\log\frac{e^{-e^{\mathrm{x}_i'\beta}}\big(e^{\mathrm{x}_i'\beta}\big)^{y_i}}{y_i!} \\
&= \underset{\beta}{\text{argmax}}\ \sum_{i=1}^{n}\Big[-e^{\mathrm{x}_i'\beta} + y_i(\mathrm{x}_i'\beta) - \ln(y_i!)\Big]
\end{aligned}$$

  A limitation of the Poisson model is that it implies equi-dispersion, i.e., that the variance is equal to the mean: $\mathbb{V}[y_i|\mathrm{x}_i] = \mathbb{E}[y_i|\mathrm{x}_i]$, whereas we often see overdispersion in the data (ex: a few traders will do many trades, many traders will do a few). To accomodate overdispersion, some softwares (e.g., R) have packages that permit an "adjusted" Poisson regression, or we can turn to the negative binomial distribution.

- **Negative binomial model**

  The negative binomial distribution $\mathrm{NB}(p,r)$ models the number of successes in a sequence of iid Bernoulli(p) trials before r failures occur. Its probability mass function is therefore $\mathrm{P}(y \mid p,r) = \binom{y+r-1}{y}\ p^y\ (1-p)^r = \frac{\Gamma(y+r)}{y!\ \Gamma(r)}\ p^y\ (1-p)^r$.

  The negative binomial distribution allows the variance to be larger than the mean, which makes it a useful overdispersed alternative to the Poisson.

  In a regression framework, it is more intuitive to specify the distribution in terms of its mean $\mu = \frac{pr}{1-p}$ and r. r is called the precision parameter or reciprocal overdispersion parameter. The distribution converges to Poisson as $r \to \infty$, i.e., as the overdispersion $\frac{1}{r} \to 0$. We rewrite the probability mass function as $\mathrm{P}(y \mid \mu,r) = \frac{\Gamma(y+r)}{y!\ \Gamma(r)}\ \big(\frac{\mu}{r+\mu}\big)^y\ \big(\frac{r}{r+\mu}\big)^r$.

  Using as link function $g()$ the usual logarithmic transformation $\ln()$, the NB regression model is:

$$y_i|\mathrm{x}_i \sim \mathrm{NB}\left(\mu_i, r\right), \quad \mu_i = e^{\mathrm{x}_i'\beta}$$

  Estimation by Maximum Likelihood:

$$\begin{aligned}
\hat{\theta}_{\text{ML}} \equiv \underset{\theta}{\text{argmax}}\ \log\mathcal{L}\big(\mathrm{y}|\mathrm{X},\theta\big) &= \underset{\theta}{\text{argmax}}\ \log\prod_{i=1}^{n}\mathrm{P}(y_i|\mathrm{x}_i,\theta) \\
&= \underset{\theta}{\text{argmax}}\ \sum_{i=1}^{n}\log\left(\frac{\Gamma(y_i+r)}{y_i!\ \Gamma(r)}\left(\frac{\mu_i}{r+\mu_i}\right)^{y_i}\left(\frac{r}{r+\mu_i}\right)^{r}\right) \\
&= \underset{\beta,r}{\text{argmax}}\ \sum_{i=1}^{n}\log\left(\frac{\Gamma(y_i+r)}{y_i!\ \Gamma(r)}\left(\frac{e^{\mathrm{x}_i'\beta}}{r+e^{\mathrm{x}_i'\beta}}\right)^{y_i}\left(\frac{r}{r+e^{\mathrm{x}_i'\beta}}\right)^{r}\right)
\end{aligned}$$

Interpretation of each coefficient $\hat{\beta}_k$ (keeping all the other predictors fixed):

– log scale $[-\infty, \infty]$ *"a 1-unit difference in x corresponds to a $\hat{\beta}_k$-unit difference in $log(\mu)$."*

– incidence rate ratio scale $[0, \infty]$ *"a 1-unit difference in x corresponds to a $e^{\hat{\beta}_k}$ multiplicative difference in μ."*

Example: $y_i \equiv$ the number of new cases of an infectious disease on day t_i. In the early stages of an epidemic, the rate of new cases can increase exponentially, s.t. $\mathbb{E}[y_i|\mathrm{x}_i] = \gamma \exp(\delta.t_i)$ is a suitable model. Using a log link turns the model into a GLM: $\log(\mathbb{E}[y_i|\mathrm{x}_i]) = \log(\gamma) + \delta.t_i$. The complete specification of a reasonable GLM consists therefore the log link, the linear predictor $\beta_0 + \beta_1.t_i$, and a Poisson or negative binomial outcome distribution.

A.3 Generalized additive models (GAMs)

We want to capture relationships between the response and predictors that are potentially non-linear. GAMs let $g\big(\mathbb{E}[y_i|\mathrm{x}_i]\big)$ depend on smooth (nonlinear) functions $h_1(), h_2(), ...$ of the predictors, where each such smooth term is made of a sum of "basis functions", as is described further below. The overall GAM is thereby a linear model in transformed variables, such that fitting can proceed with similar techniques as for linear models. These smooths enable to capture non-linear relationships and interaction effects with a high degree of flexibility; GAMs are a form of semi-parametric models.

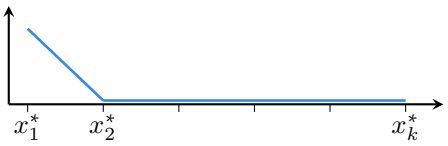
Specification We consider a simple GAM with parametric terms $\mathrm{X}_p\theta$, a univariate smooth term $h(\mathrm{x}_1)$, and a bivariate smooth term $l(\mathrm{x}_2, \mathrm{x}_3)$, and describe the forms that the smooths can take:

$$y_i|\mathrm{x}_i \sim \mathcal{F}_{\mathrm{ExpFamily}(...)}, \qquad g\big(\mathbb{E}[y_i|\mathrm{x}_i]\big) = \mathrm{X}_{ip}\theta + h(x_{i1}) + l(x_{i2}, x_{i3})$$

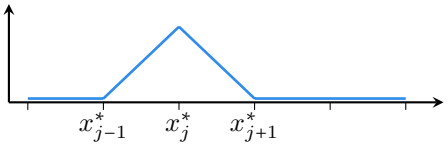
- **Univariate smooth $h()$**
  $h(x)$ is made of a linear basis expansion in x, i.e., of a sum of "basis functions" $b_j()$s, thus named for they belong to a common space of functions or "basis": $h(x) = \sum_{j=1}^{k} b_j(x)\beta_j$. Examples of bases include:
  - **Polynomial basis**
    $b_1(x) = 1,\ b_2(x) = x,\ b_3(x) = x^2,\ ... \implies h(x) = \beta_1 + \beta_2 x + \beta_3 x^2 + ...$
  - **Piecewise bases**
    The range of x is divided into multiple intervals at "knots" $x_1^*, ..., x_k^*$, and piecewise functions are fit on the intervals. The number of knots k, sometimes called the dimension of the piecewise function, controls the degree of flexibility. Examples of basis functions $\{b_j\}_1^k$:
    - **Piecewise constant basis**
      $b_1(x) = \mathbb{1}\{x < x_1^*\},\ b_2(x) = \mathbb{1}\{x_1^* \le x < x_2^*\},\ ...,\ b_k(x) = \mathbb{1}\{x_{k-1}^* \le x\}$. The resulting $h()$ is a step function.
    - **Piecewise linear basis**
      Piecewise linear functions are fit on the intervals.

      $$b_1(x) = \begin{cases} (x_2^* - x)/(x_2^* - x_1^*) & \text{if } x < x_2^* \\ 0 & \text{otherwise} \end{cases}$$

      ...

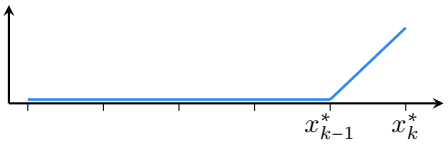


$$b_j(x) = \begin{cases} (x - x^*_{j-1})/(x^*_j - x^*_{j-1}) & \text{if } x^*_{j-1} < x < x^*_j \\ (x^*_{j+1} - x)/(x^*_{j+1} - x^*_j) & \text{if } x^*_j < x < x^*_{j+1} \\ 0 & \text{otherwise} \end{cases}$$



...

$$b_k(x) = \begin{cases} 0 & \text{if } x \le x^*_{k-1} \\ (x - x^*_{k-1})/(x^*_k - x^*_{k-1}) & \text{otherwise} \end{cases}$$



* **Piecewise polynomial basis with continuous derivatives = Spline basis**
  A spline is a piecewise polynomial function of degree m, with the imposed constraint of $m-1$ continuous derivatives, to ensure that the pieces join smoothly. Examples:
  - A linear spline is a piecewise linear polynomial continuous at each knot. It has $2(k+1)-k = k+2$ degrees of freedom.
  - A cubic spline is a piecewise cubic polynomial with continuous 1st and 2nd derivatives at each knot.[38] The set of cubic splines with fixed knots is a vector space with $4(k+1)-3k = k+4$ degrees of freedom. Ex: a cubic spline with a single knot at a point c:

$$y_i = \begin{cases} \beta_{01} + \beta_{11}x_i + \beta_{21}x_i^2 + \beta_{31}x_i^3 + \epsilon_i & \text{if } x_i < c \\ \beta_{02} + \beta_{12}x_i + \beta_{22}x_i^2 + \beta_{32}x_i^3 + \epsilon_i & \text{if } x_i \ge c \end{cases}$$

Note: Polynomials tends to be erratic near the boundaries. To prevent that behavior, one can use a "natural" or "restricted" spline, which further imposes linearity beyond the boundary knots. E.g., a natural cubic spline sets the cubic and quadratic terms there to 0. This frees up 4 degrees of freedom (2 at each end of the curve), moving the total from $k+4$ to k.

Knot selection or penalization Piecewise bases require choosing the number and location of knots. There is a trade-off between a high enough number of knots to capture non-linearities, but not so many that we're overfitting. How to select them? One could place knots either at uniform quantiles of the data, or in regions of the support where more variation in the relationship is expected, and compare different numbers of knots by cross-validation... An alternative method to avoid the knot selection problem entirely is penalization: we set a very high number of knots k, but add to the objective function (whether the likelihood or a loss function) a penalty for the 'wiggliness' or curvature of the smooth, to prevent overfitting the data. The objective function becomes:

* in maximum likelihood estimation, a penalized likelihood: $l_p(\beta) \equiv l(\beta) - \text{penalty}$
* in least squares estimation, a penalized least squares: $\text{Loss} = \sum_{i=1}^{n}(y_i - h(\mathrm{x}_i))^2 + \text{penalty}$

with $\text{penalty} \equiv \lambda \int [h''(x)]^2 dx = \lambda \beta' S \beta$. The smoothing parameter λ establishes a tradeoff between the first usual term of the objective function, which measures closeness to the data, and the second term which penalizes curvature in the function (captured by its second derivative). Increasing λ makes the function smoother, with $\lambda \to \infty$ corresponding to a straight line fit. The penalty can also be expressed in a quadratic form using a penalty matrix S.

[38] The cubic spline $h()$ can be shown to be the smoothest interpolator for any given set of points $\{x_i, y_i : i = 1, ..., n\}$, i.e., $h() = \underset{f()}{\text{argmin}} \int_{x_1}^{x_n} f''(x)^2 dx$, where $f()$ are functions continuous on $[x_1, x_n]$ with continuous 1st derivatives. As a cubic spline can closely approximate any underlying smooth function, there is seldom any good reason to use a higher-order polynomial.

While it may seem that the model is over-parameterized, due to the high number of knots, i.e., of basis functions, the penalty term actually translates to a penalty on the magnitude of the spline coefficients, which are shrunk toward the linear fit, reducing the effective degrees of freedom (EDF).[39] The choice of the basis dimension is therefore not critical, as the smoothing parameters λ_j control the actual model complexity, as long as k is not set to be too small (which will force oversmoothing).[40]

- **Multivariate smooth $l(w, z)$**

  We can also construct smooths of multiple variables.

  - For isotropic smooths, i.e., functions that aim to achieve the same smoothness per unit change in w and z, one can use a thin plate spline — not detailed here.
  - When it is unreasonable to impose the same smoothness per unit change in w and z (e.g., when the covariates are measured in fundamentally different units, such that it is difficult to scale them relative to one another), one can use a **tensor product smooth**. A tensor product is constructed from 'marginal smooths' of single covariates, each with its own basis and associated quadratic penalty, by essentially creating an interaction of each pair of basis functions for each marginal term. The basis dimension of the tensor product smooth is the product of the dimensions of the marginal bases. Ex: A tensor product smooth output with 3 EDF means that the model simplified to $\beta_0 + \beta_1 w_i + \beta_2 z_i + \beta_3 w_i.z_i$.

Bayesian/'Random model' view of smoothing The smoothing penalties on model coefficients can be given a Bayesian interpretation: namely, we have a Gaussian prior on how smooth the function is, and update it after seeing the data. Defining such a prior on model wiggliness actually gives the model the structure of a mixed-effects model. See (Wood, 2017, section 5.8) for details.

- prior: $f()$ is more likely to be smooth than wiggly: $f(\beta) \propto \exp\left(-\lambda \frac{\beta' S \beta}{\sigma^2}\right) \iff \beta \sim \mathcal{N}\left(0, \sigma^2 \frac{S^-}{\lambda}\right)$
- posterior: $\beta|y \sim \mathcal{N}\left(\hat{\beta}, (\mathrm{X}'\mathrm{X} + \lambda S)^{-1}\sigma^2\right)$ and $\hat{\beta}_{\text{MAP}} = (\mathrm{X}'\mathrm{X} + \lambda S)^{-1}\mathrm{X}'y$

Estimation We can make use of this parallel between the smoothing penalty on model coefficients and a random effect modeling of these coefficients in two ways: (i) to incorporate random effects into a GAM by treating them as wiggliness penalties, and estimating them without needing to use random effects methods; (ii) or to fit a GAM as a Generalized Additive *Mixed* Model (GAMM).

- **GAM with explicit penalties**

  Once a basis and a wiggliness penalty have been chosen for each smooth function, the GAM can be estimated. Estimation is by penalized versions of the least squares[41] and maximum-likelihood methods used for linear models. The fitting of a GAM has two components:

  (1) Estimating βs given λs (estimating the model coefficients under the penalty)

  (2) Estimating the degrees of penalization λs. Multiple methods are possible: prediction error methods, such as Generalized Cross Validation (GCV), or marginal likelihood methods based on the Bayesian/random model view of smoothing.

[39] The Effective Degrees of Freedom (EDF) tell you how 'wiggly' the fitted line is. For an EDF of 1, the predictor was estimated to have a linear relationship to the outcome. The EDF are just reported as a summary of the complexity of the estimated smooth function.

[40] In the R package `mgcv::`, choosing the argument k (the basis dimension) in spline functions amounts to setting the maximum possible degrees of freedom allowed for each model term. The actual effective degrees of freedom for each term will usually be estimated from the data, by the chosen smoothness selection criterion, but the upper limit on this estimate is $k - 1$: the basis dimension minus one degree of freedom due to the identifiability constraint on each smooth term.

[41] Least squares can be used only in the case of a ~~generalized~~ additive model.

As aforementioned, this framework also enables to incorporate conventional random effects into the GAM, by representing them like the smooths are represented: as penalized regression terms.[42]

- **GAMM**

  All the smooths are converted into random components, in particular the smoothing parameters become components of the variances of the random effect, and the whole model is estimated as a general mixed model (by ML, REML, or PQL).[43]
  For example, for a univariate spline $s(x) = \sum_j \beta_j b_j(x)$, the coefficients are represented as random effects, i.e., as belonging to a group: $\beta_j \sim \mathcal{N}\left(0, \sigma^2_\beta \mathrm{H}\right)$, where H contains the penalties. At one extreme, an unpenalized likelihood ($\beta_1, ..., \beta_k$ estimated as fixed effects) would overfit and provide "wiggly" estimates, whereas at the other extreme, a linear fit ($\beta_1 = ... = \beta_k = 0$) would smooth excessively and capture the relationship with only two parameters.

  An advantage of this GAMM approach is that it allows to incorporate correlated error structures, by dealing with them via random effects.

  Fezzi and Bateman (2015) uses this method to obtain the optimal (i.e., best linear unbiased predictor) trade-off between excessive smoothing and overfitting of the nonlinear function.

[42] The R function `mgcv::gam` tackles fitting a GAM as a penalized likelihood maximization problem, and uses a separate criterion to estimate the smoothing parameters. The computational strategy is as follows: Basis functions and quadratic penalty matrices are constructed for each smooth term and are combined with a matrix for the strictly parametric part of the model, to form a complete "design" matrix and a set of penalty matrices for the smooth terms. The linear identifiability constraints are also added. The penalized likelihood maximization problem is then solved by Penalized Iteratively Re-weighted Least Squares (PIRLS), and the smoothing parameter estimation problem is solved using the GCV criterion or a Laplace approximation to REML, and is conducted by "outer iteration": The PIRLS scheme is iterated to convergence for each trial set of smoothing parameters, and GCV or REML criteria are only evaluated on convergence (optimization is 'outer' to the PIRLS loop). Random effects can be incorporated by specifying the basis as such. Ex: `s(g,bs='re')` generates an i.i.d. normal random effect; if g is a factor, it produces a random coefficient for each level of g, with the set of coefficients modeled as i.i.d. normal.

[43] In the *normal errors, identity link* case, estimation can be performed using general linear mixed effects modelling software (e.g., in R, as provided by the function `nlme::lme`). This is what the R function `mgcv::gamm` does. It shows it explicitly by returning two items: (i) the fitted model returned by `nlme::lme`, and (ii) an object of class 'gam', including a posterior covariance matrix for the parameters of all the fixed effects and the smooth terms.

A.4 Multilevel models

Context We are interested in the relationship of y_i with x_i. Our data present some hierarchical structure: individuals belong to groups.[44] Hence observations are not independent, we may expect a group effect $a_{j[i]}$ on the outcome, and we may expect also both a *within*-group relationship and a *between*-group relationship of x_i with the outcome.

Structure A multilevel model is a generalization of a classical regression model that not only allows for variation in coefficients but *models* that variation. Let us consider a simple 2-level model. The lower level is a common regression model with individual-level predictors. Now some of its coefficients (the intercept and/or slopes) are allowed to vary by group and are *modeled*, i.e., they are given a probability model, which constitutes the higher- or group-level model.

The generalization proceeds along two dimensions; simple examples of models are given below:[45]

(i) which coefficients are modeled, and whether they are attributed a common multivariate distribution or separate distributions;
(ii) whether group-level predictors are included.

$\forall i = 1, ..., n, \quad j = 1, ..., J$:

	varying intercept	varying intercept and slope	co-varying intercept and slope
w/o group-level predictors	$y_i \sim \mathcal{F}_1(a_{j[i]} + \beta x_i,\ \sigma_y^2)$ $a_j \sim \mathcal{F}_2(\mu_a,\ \sigma_a^2)$	$y_i \sim \mathcal{F}_1(a_{j[i]} + b_{j[i]} x_i,\ \sigma_y^2)$ $a_j \sim \mathcal{F}_2(\mu_a,\ \sigma_a^2)$ $b_j \sim \mathcal{F}_3(\mu_b,\ \sigma_b^2)$	$y_i \sim \mathcal{F}_1(a_{j[i]} + b_{j[i]} x_i,\ \sigma_y^2)$ $\begin{bmatrix} a_j \\ b_j \end{bmatrix} \sim \mathcal{F}_{ab}\left(\begin{bmatrix} \mu_a \\ \mu_b \end{bmatrix}, \begin{bmatrix} \sigma_a^2 & \sigma_{ab} \\ \sigma_{ab} & \sigma_b^2 \end{bmatrix}\right)$
with group-level predictors	$y_i \sim \mathcal{F}_1(a_{j[i]} + \beta x_i,\ \sigma_y^2)$ $a_j \sim \mathcal{F}_2(\alpha_0 + \alpha_1 z_j,\ \sigma_a^2)$	$y_i \sim \mathcal{F}_1(a_{j[i]} + b_{j[i]} x_i,\ \sigma_y^2)$ $a_j \sim \mathcal{F}_2(\alpha_0 + \alpha_1 z_j,\ \sigma_a^2)$ $b_j \sim \mathcal{F}_3(\beta_0 + \beta_1 l_j,\ \sigma_b^2)$	$y_i \sim \mathcal{F}_1(a_{j[i]} + b_{j[i]} x_i,\ \sigma_y^2)$ $\begin{bmatrix} a_j \\ b_j \end{bmatrix} \sim \mathcal{F}_{ab}\left(\begin{bmatrix} \alpha_0 + \alpha_1 z_j \\ \beta_0 + \beta_1 l_j \end{bmatrix}, \begin{bmatrix} \sigma_a^2 & \sigma_{ab} \\ \sigma_{ab} & \sigma_b^2 \end{bmatrix}\right)$

Terminology and relation to classical models

Let us see what assumptions these models embed, and how these compare to those of classical single-level models. For simplicity, we consider a normally-distributed outcome and a single regressor x_i with a homogeneous effect β, i.e., group effects manifest only through the intercept).

Any model fit embeds some assumption about how these group effects $\{a_{j[i]}\}_j$ are distributed. And notably about their variance σ_a^2, which shows how close they are to each other, i.e., how much they are pooled toward their mean. Classical models correspond to extreme assumptions about this variance, while a multilevel model is a data-driven compromise:

- *Complete pooling*: We don't allow for group effects.

[44] For example, we have observations of n students in J classrooms such that each student i belongs to a group $j[i]$, or we have longitudinal data, where each dated observation $\{i, t\}$ belongs to a unit i. For simplicity, we consider in this section only such hierarchical data with two levels. However, results extend to more levels, that can be nested or not.

[45] For clarity, we use Latin letters for variables and modeled parameters, and Greek letters for fixed parameters (as distinguished in a frequentist framework). The fixed parameters that inform varying parameters are called the "hyperparameters" of the full model.

  - Conceptualization: No group effects.
  - Implementation: We neither model them nor adjust for them. We fit the basic model:

$$\begin{aligned} & y_i = \alpha + \beta x_i + e_i, && e_i \sim \mathcal{N}(0,\ \sigma_y^2) \\ \iff \quad & y_i = a_{j[i]} + \beta x_i + e_i, && e_i \sim \mathcal{N}(0,\ \sigma_y^2),\ a_j \sim \mathcal{F}(\alpha,\ \mathbf{0}) \end{aligned} \tag{m1}$$

- *No pooling*: We allow for group effects but don't model them.
  - Conceptualization: The group effects are "fixed effects", unrelated to each other.
  - Implementation: We fit a separate intercept per group (by including a dummy variable for each j) or de-mean the data by group, which eliminate the group effects. We fit the "fixed effects" model:

$$\begin{aligned} & y_i = \alpha_{j[i]} + \beta x_i + e_i, && e_i \sim \mathcal{N}(0,\ \sigma_y^2) \\ \iff \quad & y_i = a_{j[i]} + \beta x_i + e_i, && e_i \sim \mathcal{N}(0,\ \sigma_y^2),\ a_j \sim \mathcal{F}(\alpha,\ \infty) \\ \iff \quad & (y_i - \bar{y}_{j[i]}) = \beta(x_i - \bar{x}_{j[i]}) + e_i, && e_i \sim \mathcal{N}(0,\ \sigma_y^2) \end{aligned} \tag{m2}$$

- *Partial pooling*: We allow for group effects and model them.
  - Conceptualization: The group effects are "random effects" with a joint probability distribution.
  - Implementation: We model them with error, i.e., we estimate the parameters of that distribution. The $\{a_j\}$s are pooled toward their common mean μ_a, by an amount that depends on the sample size of each group and on σ_a^2,[46] which is also estimated from the data. We fit the multilevel or "random effects" model:[47]

$$\begin{cases} y_i = a_{j[i]} + \beta x_i + e_i & e_i \sim \mathcal{N}(0,\ \sigma_y^2) \\ a_j \sim \mathcal{F}(\mu_a,\ \boldsymbol{\sigma_a^2}) \end{cases} \tag{m3}$$

Partial pooling is thereby a data-driven compromise between *complete pooling* (equivalent to setting $\sigma_a \to 0$) and *no pooling* (setting $\sigma_a \to \infty$). The former assumes away variation between groups, while the latter risks overstating that variation, i.e., overfitting the data, and neither lets us analyze it (Gelman and Hill, 2006, ch. 12).

Why use multilevel models?

In short: to (1) acknowledge and (2) analyze within-group *and between-group* variations.

- *To account for the group-dependence in our data (e.g., with time series, spatial correlation, networks...).* Traditional regression techniques assume independent observations. Any dependence structure in our data that isn't modeled will be left out in the error term, and the corresponding standard errors of regression coefficients will be underestimated (esp. that of higher-level regressors). A common way to deal with group-dependence in econometrics is to fit group dummies and to *cluster standard errors* by group (after fitting the model, a new estimator of the error covariance matrix is computed that adjusts

[46] Partial pooling is proportional to the variance, not the standard deviation.

[47] In the regression framework, multilevel regression models are a particular case of "random effect" models that pool information across groups, or "mixed effect" models when they include both a "random effect" component and a "fixed effect" component (e.g., a varying-intercept, fixed-slope model). The "fixed effect" vs "random effect" terminology is confusing, as different disciplines use these to refer to different things. For instance, in econometrics, a fixed effect is an intercept that varies by group (and that is estimated using dummy variables, i.e., that isn't modeled), while in other branches of statistics, it means a coefficient that does not vary. Gelman and Hill (2006, p. 245) suggests to avoid these terms and instead describe models explicitly: for example, a 'varying intercepts and constant slopes' model.

for the dependence within groups). Instead, a multilevel model models this group-dependence. It is equivalent to a classical regression model with correlated *and modeled* errors:[48]

$$\begin{cases} y_i = \alpha_{j[i]} + \mathrm{x}_i'\beta + e_i, & e_i \sim \mathcal{N}\left(0, \sigma_y^2\right) \\ \alpha_j = \mu_\alpha + \eta_j, & \eta_j \sim \mathcal{N}\left(0, \sigma_\alpha^2\right) \end{cases} \iff y_i = \mu_\alpha + \mathrm{x}_i'\beta + \underbrace{e_i + \eta_{j[i]}}_{e_i^{\text{all}}}, \quad e^{\text{all}} \sim \mathcal{N}\left(0, \Sigma\right)$$

- *To increase efficiency by pooling information.* By treating groups as a "random-effect" within the model, we can pool shared information about the mean across the groups. *Partially pooling* the varying coefficients will produce more efficient (less noisy) estimates of the J regression lines than by including group indicators, especially when the number of observations in some groups is small.[49]
- *To model heterogeneity in the relationship to a covariate.* For example, in causal inference, we may be interested not just in an average treatment effect but also in how the effect varies across the population. With a multilevel model, we can model variation in the expected treatment effect, for example as a function of pre-treatment covariates x.
- *To generalize results to a population not well-represented in the sample.*
  - To do inference for the population of groups when our data are not random samples: one can generalize to a larger population using *multilevel regression and poststratification (MRP).*
  - To estimate $\hat{y}$ for particular groups: notably to get reasonable estimates even for groups with small sample sizes (which is difficult with classical regression).
  - To predict a new observation in a new group: multilevel regression enables to quantify sources of variation, and hence to propagate the uncertainty about the new group into the uncertainty about the new individual in this group; this distinction isn't provided in classical regression.

  Ex: With fixed effects, these group effects are allowed to take any value whatsoever, which amounts to saying that each individual group is completely different to every other group: having results for 6 groups tells us nothing about a 7$^{\text{th}}$. Treating the groups instead as a random sample from the population of groups allows us to estimate the relationship between y_i and x_i and to generalize beyond the 6 groups in the sample. Each new group would indeed have the same distribution as that of the 6 groups in the sample.

Endogeneity concern in causal inference, solved by REWB

Context: We want to estimate the causal effect β of x_i on y_i. But there are unobserved group-level or "between" effects that are correlated with x_i and determine y_i. Hence the slope estimator of the "within effect" in a simple regression of y_i on x_i would suffer from omitted variable bias.[50] How to tackle this bias on level 1 coefficients due to omitted variables at level 2?

We saw that the J unobserved group effects can be considered as either *fixed effects* (i.e., unrelated) or *random effects* (i.e., belonging to a shared distribution). The FE approach provides an unbiased estimator of the *within* effect, with the caveats that (i) the assumption of unrelatedness is very questionable, and (ii) no group-level variable can be identified as all the group-level variance is accounted for. However, the simple RE model presents another important problem: its estimator $\hat{\beta}$ is biased for the within effect. Indeed, a

[48] The error e_i^{all} is the sum of an individual-level noise e_i and a group-level error $\eta_{j[i]}$ which induces correlation in e^{all}. The covariance matrix Σ is parameterized in some way, and these parameters are estimated from the data.

[49] Note however that when the number of groups J is small, it is difficult to estimate the between-group variation σ_a^2 precisely. As this σ_a^2 determines the amount of partial pooling, a bad estimation of its value results in pooling by a somewhat random amount. Hence multilevel modeling adds little to no-pooling models.

[50] This endogeneity concern can arise with any form of nested data. A very common case is that of longitudinal data (it = level 1, i = level 2), where the concern is of time-constant effects that are plausibly correlated with x_{it}.

correlation with $\bar{x}_i$ is left in the error term, as can be seen by rewriting the model as below, where η_j contains these problematic unobserved group-level effects that are correlated with x_i and determine y_i:

$$\begin{cases} y_i = a_{j[i]} + \mathrm{x}_i'\beta + e_i, & e_i \sim \mathcal{N}\left(0, \sigma_y^2\right) \\ a_j = \mu_\alpha + \eta_j, & \eta_j \sim \mathcal{N}\left(0, \sigma_a^2\right) \end{cases} \iff y_i = \mu_\alpha + \mathrm{x}_i'\beta + \underbrace{e_i + \eta_{j[i]}}_{e_i^{\text{all}}}, \quad e^{\text{all}} \sim \mathcal{N}\left(0, \Sigma\right)$$

The bias can be described: the estimator $\hat{\beta}_{\text{RE}}$ is actually a weighted average of the within estimator and the between estimator.[51]

A solution is to add $\bar{x}_{j[i]}$ as a regressor, as it will extract the problematic correlation from the composite error term e_i^{all}. We recover an estimator of the within effect that is not biased by group-level confounders.[52] The model is known as the 'Within-Between RE' (REWB) model (Bell et al., 2019), as it decomposes x into a within-group component and a between-group component, or 'correlated RE' (CRE) model (Wooldridge, 2013, 14.3). The three expressions of the model below are equivalent:[53]

$$\begin{cases} y_i = a_j + \beta_1(x_i - \bar{x}_j) + \beta_2\bar{x}_j + e_i \\ a_j \sim \mathcal{N}\left(\mu_a,\ \sigma_a^2\right) \end{cases} \iff \begin{cases} y_i = a_j + \beta_1(x_i - \bar{x}_j) + e_i \\ a_j \sim \mathcal{N}\left(\alpha + \beta_2\bar{x}_j,\ \sigma_a^2\right) \end{cases} \iff \begin{array}{l} y_i = \beta_0 + \beta_1(x_i - \bar{x}_j) + \beta_2\bar{x}_j + \nu_j + e_i, \\ \nu_j \sim \mathcal{N}\left(0,\ \sigma_\nu^2\right) \text{ uncorrelated w. } x_i \end{array}$$

The REWB model also makes it possible to:

- estimate the between effect, as well as any relationship between the outcome and a group-level covariate (though one can't interpret their coefficients as causal);
- test whether the between and the within effects are significantly different (with a Hausman test of whether the difference between the two coefficients in the model is statistically different from 0);
- estimate the level-2 variance and compare it to the level-1 variance.

The REWB model can also naturally be extended to varying intercepts *and slopes* and/or additional group-level predictors.[54]

CCL: do FE only when really don't care about between-group relationships + don't worry about efficiency (have a very large sample size).

Inference

[51] $\beta_1 = \frac{w_{\text{W}}\beta_{\text{W}} + w_{\text{B}}\beta_{\text{B}}}{w_{\text{W}} + w_{\text{B}}}$, where $w_{\text{W}} \equiv 1/\text{SE}[\hat{\beta}_{\text{W}}]^2$ and $w_{\text{B}} \equiv 1/\text{SE}[\hat{\beta}_{\text{B}}]^2$ are the precisions of the within estimate and the between estimate, respectively. As there are more data at level 1 (and therefore higher precision of the within estimate), $\hat{\beta}_1$ will often tend towards the within estimate (Bell et al., 2019).

[52] ⚠ With a *non-identity link function,* unbiasedness is guaranteed only if u_j is a linear function of $\bar{x}_{j[i]}$. I.e., to get an unbiased effect, we are trading one assumption of linearity for another (in the standard model, we assume an identity link function; in the REBW model with non-identity link function, we assume u_j is a linear function of $\bar{x}_j$). However, the available evidence (from simulations) suggests that the bias of the REBW method remains small in most situations (Bell et al., 2019). One can also include functions of $\bar{x}_{j[i]}$ as regressors to characterize more flexible functional forms of the correlation. P. Allison suggests using polynomial functions of the means, i.e., including not only $\bar{x}_i$ but also $\bar{x}_i^2, \bar{x}_i^3$ as regressors, or other cluster-level functions of the x_{it}, such as their standard deviation (see `https://statisticalhorizons.com/problems-with-the-hybrid-method/`). If the estimates of the added coefficients aren't significant and the estimate of β_W doesn't change much, it suggests the linearity assumption is reasonable and bias should not be such an issue.

[53] We center level-1 variables, i.e., we use as regressor $\left(x_{it} - \bar{x}_{j[i]}\right)$ rather than x_i, such that there is no correlation between $\left(x_{it} - \bar{x}_{j[i]}\right)$ and $\bar{x}_{j[i]}$ and β_2 represents the between effect rather than the "contextual effect" (Bell et al., 2019).

[54] For example: $\begin{cases} y_i = a_j + b_j(x_i - \bar{x}_j) + e_i \\ \begin{bmatrix} a_j \\ b_j \end{bmatrix} \sim \text{MVN}\left(\begin{bmatrix} \beta_0 + \beta_2\bar{x}_j \\ \beta_1 \end{bmatrix}, \begin{bmatrix} \sigma_a^2 & \sigma_{ab} \\ \sigma_{ab} & \sigma_b^2 \end{bmatrix}\right) \end{cases} \iff y_i = \beta_0 + \beta_1(x_i - \bar{x}_j) + \beta_2\bar{x}_j + \nu_{i0} + \nu_{i1}(x_i - \bar{x}_j) + e_i$, where the composite error term $\nu_{i0} + e_i$ is uncorrelated with x_i. Note that the group average need not be included as a predictor for the slope.

- **Frequentist (point estimation)**
  First the hyperparameters are estimated via Maximum Likelihood or Restricted Maximum Likelihood (REML), then inference is performed for the coefficients conditional on the estimated hyperparameters.

  ⚠ ML estimators are unbiased for the group-level mean parameters but downward-biased for the group-level variance parameters (especially when the number of groups is small), because the mean parameters are assumed to be known with certainty when estimating the variance parameters. Instead, REML accounts for the number of mean parameters estimated, losing 1 degree of freedom for each, and so produces unbiased estimators of variance parameters. Note however that to compare models with *different fixed effects* with a likelihood ratio test, ML must be used, as LR tests for REML require exactly the same fixed effects specification in both models.

  👍 Computational: fast.

  👎 A. Gelman: *"The usual non-Bayesian procedures are designed to work well asymptotically (in the case of hierarchical models, this is the limit as the number of groups approaches infinity). But as noted Bayesian J. M. Keynes could've said, asymptotically we're all dead."*

- **Bayesian**
  All levels are fitted simultaneously. The hyperparameters are given a prior distribution, and we estimate their whole posterior distribution.

  👍 Accounts for all the uncertainty in the parameter estimates when predicting the varying intercepts and slopes, and their associated uncertainty.

  👎 Computational: slow. Markov chain Monte Carlo simulations are generally much slower than (RE)ML estimation.