Applied Microeconometrics in the Tree of **Statistics**

Fundamentals

Contents

Motivation

Research on/for sustainable development is particularly conducive to interdisciplinary work. At their core, related research questions bring together social systems and natural systems. Interdisciplinary researchers may therefore be heavily trained in social sciences, and in particular economics.

Microeconometrics faces some criticisms from the statistical community, while its strengths for specific types of analyses – notably causal inference – are also recognized. In interdisciplinary research heavily grounded in microeconometrics, one is likely to encounter alternative types of models and estimation methods. To justify one's choice of method to address a specific research question in front of different disciplinary communities, the interdisciplinary researcher must understand where these methods diverge in the 'family tree' of statistical analyzes, and which method is the most relevant, instead of which is the default in this specific discipline.

The purpose of this document is therefore twofold: to detail the methods of applied (micro)econometricians, which are our reference base; and to put them into context (i.e., in the space of statistical inference methods). Specifically:

- to define and organize notions that are omnnipresent in applied microeconomics papers, but whose distinction -and therefore understanding- may be fuzzy (what are the differences between *a model, an equation, a regression, a specification, an estimation method, an identification strategy*[1]...);
- to tie microeconometrics to its greater 'family tree' of statistical analyzes, and delineate a few other branches of that tree that might be relevant for empirical research bringing together economics and other disciplines.

Let us start by defining microeconometrics:

Econometrics = one specific area of applied statistics. The distinguishing feature of econometrics is the emphasis placed on causal modeling.

Microeconometrics = the theory and applications of methods of data analysis developed for micro-data pertaining to individuals, households, and firms.
The original goal: obtaining estimates that can be given a structural interpretation.

*Disclaimer: sections and lines in brown correspond to content which is **very much** 'under construction'.*

[1]The concept of identification strategy, in that it corresponds to a very specific type of research question and goal, is addressed in a separate document, see 'Causal inference in observational studies – Theory, Methods and Presentation'.

1 Models

A model is a formal representation of a theory about a system.

A statistical model

= a mathematical model that specifies relationships between random and non-random variables. ($\rightarrow$ It is non-deterministic: some variables are stochastic, they have probability distributions.)

= a mathematical model of the data generating process (DGP)[a]. (Statistical modelling = considering that each observation in a sample $\{y_i, X_i\}_1^n$ is generated by an underlying process described by the model.)

The goal is to explain the **variation** of random variables.

[a] Formally, it combines the set of possible observations or "sample space" $\mathcal{S}$ and a collection of 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).

1.1 $\supset$ 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.

$$g\big(y, X, \varepsilon|\theta\big) = 0, \quad \theta = \text{structural parameters}$$

- **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 equations is the result of solving the system for the endogenous variables. This **gives the endogenous variables as functions of the exogenous variables**.

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

1.2 $\supset$ 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*[2] of $\boldsymbol{y}$, given $\boldsymbol{X}$.**

[2] Different regression models will look at different parts of the distribution, and specify them differently. Ex: classical linear regression model: $\mathbb{E}[\boldsymbol{y}|\boldsymbol{X}] = \boldsymbol{f}(\boldsymbol{X}) = \boldsymbol{X}\boldsymbol{\beta}$; quantile regression model: $\mathbb{Q}[\boldsymbol{y}|\boldsymbol{X}] = \boldsymbol{f}(\boldsymbol{X})$...

Writing a regression model We consider a sample $\{y_i, X_i\}_1^n$ generated by the underlying process described by that model; we can then write the model interchangeably:

- as a system of n equations: $y_i = f(X_i, \varepsilon_i|\beta),\ \forall\ i = 1, .., n$
- using matrix notation[3]: $y = f(X, \varepsilon|\beta)$

Ex: the classical linear regression model:

$$y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \varepsilon_i, \quad \varepsilon_i \overset{\text{iid}}{\sim} (0, \sigma^2),\ i = 1, ..., n$$

$$y = X\beta + \varepsilon, \quad \varepsilon \sim (0, \sigma^2 I_n)$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Choosing a model specification After choosing a general regression model comes model specification: the process of selecting which independent variables to include and choosing an appropriate functional form for the model.

Specification error occurs when the functional form or the choice of independent variables poorly represent relevant aspects of the true data-generating process. In particular, bias (the expected value of the difference of an estimated parameter and the true underlying value) occurs if an independent variable is correlated with the errors inherent in the underlying process. There are several different possible causes of specification error; some are listed below.

"Correct specification" is, in practice, unrealistic, as we do not observe the true DGP. In practice we try to avoid the three basic types of misspecification:

- using an inappropriate functional form.
- including an x that is theoretically irrelevant (has no partial effect on y) $\rightarrow$ *overspecified* model, will be less precise.
  - (For causal inference studies) On the other hand, adjusting for a relevant x, even if it is unrelated with D_i, can be useful, as it increases the precision of our estimate (it reduces the residual variation, therefore the standard error $SE(\hat{\beta})$)[4].
- excluding an x that is theoretically relevant $\rightarrow$ *underspecified* model, may be biased.
  - ⚠ It can get tricky. Ex: adjusting for an additional variable x_3 without adjusting for another x_4 that should also be included can actually take us further away from the effect β_2!

[3]The error term ε is a vector of n random variables. It has a symmetric variance-covariance matrix of dimensions $n \times n$.

[4]Do not control for x if there is multi-colinearity (x correlated with D). The SE[$\hat{\beta}$] might increase, and the coefficient itself might change: including a variable that responds to the treatment introduces some selection bias. In particular, do not control for an intermediate outcome.

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[5]. We assume that the data comes 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[6], 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 $\theta = (\mu, \sigma)$.
- **non-parametric:** the model makes no assumptions about a parametric distribution, it determines it from data[7]. 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 predictors and dependent variable. *Ex: Kriging; LOESS.*
- **semi-parametric:** the model combines parametric and nonparametric models.
  Ex: only a few moments are specified: $\mathbb{E}[\varepsilon] = 0$ *and* $\mathbb{V}[\varepsilon] = \mathbb{E}[\varepsilon\varepsilon'] = \Omega$.

Why care about parametrization? Because what we are really 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 it is associated with, so that knowledge about the parameter translates easily to knowledge about the distribution.

Identification in parametric models

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

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 by functional form (by parameterization of the error distribution) or from exclusion, inequality and covariance restrictions.

Ex of nonidentification: in the linear regression $y = X\beta + e$, perfect colinearity between regressors means we can't identify β.

[5] 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 $\mathcal{P} = \{P_\theta | \theta \in \Theta \subseteq \mathbb{R}^k\}$.

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

[7] 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

We can separate studies into two classes, leading to different types of data:

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 $F(W_t \mid \theta_t)$.

2.1 Types of observational data

We can group observational data into 3 categories, based on the dimensions: units (N) and time (T):

- **Cross-sectional [N]**: observations for several units, at one t
- **Time series [T]:** observations for a single unit, at repeated points in time
- **Longitudinal [N $\times$ T]**: observations for several units, at repeated points in time.

  When *the same units* are observed over time, we have **panel data**[8]. 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-variation and between-variation.

One of the strengths of longitudinal data (particularly balanced panel 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**: assumption on which is based statistical inference theory.
- **Stratified random sampling**
  The population is divided into L subgroups or "strata", containing $N_1 \neq N_2, ..., N_L$ units. Simple random samples of sizes $n_1, n_2, ..., n_L$ are drawn independently from the strata.

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

[8] "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 or DGP, based on statistics calculated from a sample drawn from the population.
It uses a mathematical model and follows 2 steps:
1. estimating the model
2. assessing uncertainty in the fit.

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

Statistical inference combines data with (explicit or implicit) prior assumptions, and involves:

- **estimation** – 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.
  It is rooted in the probabilistic approach, by which:

  > The data observed is of a random variable X. We want to estimate parameters θ of its conditional probability distribution $f(X \mid \theta)$.

- **hypothesis testing** – testing for a specific value of the unknown parameter θ.

3.1 Frequentist vs Bayesian inference

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

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	
$\implies$ the parameter 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 LT error rates.	$\implies$ the parameter θ is a *random variable*. We can assign a $\mathcal{P}$ distribution over possible values of θ, to represent our uncertainty 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 $\widehat{CI}_\theta$) of θ that is most consistent with the data. Result: a conclusion, 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
Prediction	
Use the point estimate (as the most likely value) of θ, and its CI.	Use the full posterior $\mathcal{P}$ distribution of θ, which allows us to take into account the uncertainty in θ.

The sections below describe the *ABC* of statistical inference in the context of regression, 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**.

Regressions can have 2 purposes:

- prediction: the **conditional prediction** of the distribution **of y**, given some X
- comparison: comparing these predictions for different values of X, to make simple comparisons between groups or estimate causal effects[a].

[a] Refering to regression coefficient estimates $\hat{\beta}$ as "effects" is misleading – unless we are doing causal inference. The safest interpretation of a regression estimate is as a comparison, using "differences" rather than "effects" and "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 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 *distribution* for each coefficient. We can then choose to focus on its conditional mean $\mathbb{E}[y|X]$, its conditional quantile $\mathbb{Q}[y|X]$...

3.2.2 Estimators

We have a set of observations $x_1, ..., x_n$, i.e., a realization of the random sample $X_1, ..., X_n$ (each X_i is a random variable).

An estimation method or **estimator** T_n of the population parameter θ 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., calculated on 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 Estimator (MLE)**[9]
  We want to find the value of θ that makes the observed data most likely. The likelihood function in a

[9] MLE is just a type of statistic, so it is conceptualized under either inference approach: from the vantage point of Bayesian inference, MLE is a special case of maximum a posteriori estimation (MAP) that assumes a uniform prior distribution of the parameters. In frequentist inference, MLE is a special case of an extremum estimator, with the objective function being the likelihood.

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

$$\begin{aligned} L\big(y \mid X, \theta\big) &= f\big(X_1, ..., X_n, \theta\big) \\ &= f\big(X_1, \theta\big) ... f\big(X_n, \theta\big) \\ &= \prod_{i=1}^{n} f\big(X_i, \theta\big) \\ \log L\big(y \mid X, \theta\big) &= \sum_{i=1}^{n} \log f\big(X_i{=}x_i, \theta\big) \end{aligned}$$

We calculate $\hat{\theta}_{\text{MLE}} \equiv \underset{\theta}{\text{argmax}}\ L\big(y \mid X, \theta\big) = \underset{\theta}{\text{argmax}}\ \log L\big(y \mid X, \theta\big)$

- **Least Squares (LS)**
  The fit of a model $y = g(X, \varepsilon)$ to each data point is measured by its residual $r_i \equiv y_i - g(x_i, \beta)$. We compute the values of the parameters that minimize the sum of the squares of (eventually *a function* $k()$ *of*) the residuals; they are those that best fit the data.[10]

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

  When the model is a linear combination of the parameters $g(X, \beta) = \sum_j \beta_j h_j(X)$, Least Squares is a **Linear Least Squares (LLS)**.

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

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

    In the special case of the univariate or "simple" regression model ($y = \alpha + \beta x + \varepsilon$), $\hat{\beta}_{\text{OLS}} = \frac{\text{cov}[x,y]}{\mathbb{V}[x]} = \frac{\frac{1}{n}\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n}\sum_i (x_i - \bar{x})^2}$

  - **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 covariance matrix $\text{Cov}(\varepsilon|X) = \Omega$ is not diagonal with values σ^2), OLS will again be inefficient.

[10] Indeed, let $\varepsilon \equiv y - \hat{y}$ be the unobserved prediction error, $L(\varepsilon)$ the loss. We want to minimize the expected loss $\mathbb{E}\left[L(\varepsilon)|X\right]$, so we want the predictor that minimizes the sum of that function $L()$ of the residuals. For a squared error loss function $L(\varepsilon) = \varepsilon^2$, it means minimizing the sum of squared residuals $\sum_i r_i^2$. That predictor is the **conditional mean**: $\underset{g()}{\text{argmin}}\ \mathbb{E}\left[(y - g(X))^2\right] = \mathbb{E}[y|X]$. I.e., by computing $\hat{\beta}_{\text{LS}}$, our subsequent predictions will be of the conditional mean of y.

We minimize instead the squared *Mahalanobis length*[11] of the residuals:

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

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

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

- **Least (asymmetric) absolute error**
We are interested in minimizing a different loss function, in order to place a penalty of $(1-\alpha)|\varepsilon|$ on overprediction and a different penalty $\alpha|\varepsilon|$ on underprediction.
$L(\varepsilon) = ~~\varepsilon^2~~ = \begin{cases} (1-\alpha)|\varepsilon| & \text{if} \quad \varepsilon < 0 \\ \alpha|\varepsilon| & \text{if} \quad \varepsilon \geq 0 \end{cases}$: asymmetric absolute error loss
The optimal predictor is the **conditional quantile**: $\underset{g()}{\text{argmin}}\, \mathbb{E}\left[L(y - g(X))\right] = \mathbb{Q}[y|X]$. I.e., by computing $\hat{\beta}_{\text{LAA}}$, our subsequent predictions will be of the conditional quantile of y.

In the specific case of $L(\varepsilon) = |\varepsilon|$, i.e., *symmetric* absolute error loss, the least absolute deviations predictor is the **conditional median** $\text{med}[y|X]$. I.e., by computing $\hat{\beta}_{\text{LSA}}$, our subsequent predictions will be of the conditional median of y.

The least absolute errors estimators are more robust than LS in the presence of outliers.

3.2.3 Estimator properties

See section 2.1 in `https://clairepalandri.github.io/CLRM&estimators.pdf`. Much shorter version to be added here eventually.

To add:

1. Estimation methods:
   - LOESS algorithm
   - MCMC
2. Why/When to use LS vs MLE...

3.2.4 Uncertainty in the estimate: computing SEs & CIs

i. The uncertainty in any sample statistic can be captured by its SE & $\widehat{\text{CI}}_{0.95}$

[11] 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 thus unitless and scale-invariant, and takes into account the correlations of the data set.

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.

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 could 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 *sampling* distribution.
- The **95% Confidence Interval $\widehat{\text{CI}}_{0.95}(\hat{\theta})$** = 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."*[12]
Another - maybe more adequate - name for such intervals may be "compatibility intervals", in that they give a range of parameter values that are most compatible with the observed data (Gelman and Greenland, 2019).
All these statements are of course based on the assumption that the model is correct.

ii. Traditional approach: asymptotic theory

Considering a parameter of interest θ and its estimation with $n-k$ degrees of freedom $\hat{\theta}$. The sampling distribution of the standardized estimate $\frac{\hat{\theta}-\theta}{\text{SD}[\hat{\theta}]}$ is a Student's t distribution with $n-k$ degrees of freedom. Hence, we could describe our uncertainty in $\hat{\theta}$ by the interval that covers 95% of the distribution mass:

$$\text{CI}_{95\%}(\hat{\theta}) = \Big[\hat{\theta} + q_{t,n-k}(0.025) \times \text{SD}[\hat{\theta}] \;;\; \hat{\theta} + q_{t,n-k}(0.975) \times \text{SD}[\hat{\theta}]\Big]$$

We do not know $\text{SD}[\hat{\theta}]$, but we can estimate it by $\text{SE}[\hat{\theta}]$, thus we compute the interval *estimate*:

$$\widehat{\text{CI}}_{95\%}(\hat{\theta}) = \Big[\hat{\theta} + q_{t,n-k}(0.025) \times \text{SE}[\hat{\theta}] \;;\; \hat{\theta} + q_{t,n-k}(0.975) \times \text{SE}[\hat{\theta}]\Big]$$

Finally, the larger the degrees of freedom $n-k$, the closer a t distribution gets to the standard normal distribution. Therefore, when $n-k$ is sufficiently large, we can simply use the ǂ normal distribution:

$$\widehat{\text{CI}}_{95\%}(\hat{\theta}) = \Big[\hat{\theta} + q_{\mathcal{N}}(0.025) \times \text{SE}[\hat{\theta}] \;;\; \hat{\theta} + q_{\mathcal{N}}(0.975) \times \text{SE}[\hat{\theta}]\Big] = \Big[\hat{\theta} \pm 1.96 \times \text{SE}[\hat{\theta}]\Big]$$

[12] 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 frequentists 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.

Example 1: sample statistic is the sample mean $\bar{x}$

- Population: mean μ_x and standard deviation σ_x are unobserved
- Samples: suppose we take many random samples of size n, for each sample s compute its mean $\bar{x}_s$; and finally plot the distribution of the $\{\bar{x}_s\}_s$, i.e., the sampling distribution of $\bar{x}$. This distribution is a t distribution with $n-1$ degrees of freedom, centered around μ_x and with standard deviation $\frac{\sigma_x}{\sqrt{n}}$. The larger n, the closer this t_{n-1} distribution will be to a normal distribution. σ_x is unobserved; a reasonable estimate for it is the *sample* standard deviation[13] s_x, which we observe. Therefore we compute $\text{SE}[\bar{x}] = \frac{s_x}{\sqrt{n}}$

Example 2: sample statistic is a regression slope $\hat{\beta}$ Let us consider the multivariate linear regression model, and assume homoscedastic errors s.t. $\mathbb{E}[\varepsilon\varepsilon'] = \sigma^2 I$.

- Population: parameter β, and variance σ^2 are unobserved
- Sample: parameter estimate $\hat{\beta} \sim \big(\beta, \text{SD}[\hat{\beta}]\big)$ is observed;
  its variance $\mathbb{V}[\hat{\beta}] = \sigma^2(X'X)^{-1}$ and hence standard deviation $\text{SD}[\hat{\beta}]$ are not, as σ is not. We can consistently estimate the population variance σ^2 by the *sample* variance $s^2 = \frac{1}{n-k}\sum_i r_i^2$.
  We therefore estimate $\mathbb{V}[\hat{\beta}]$ by $\widehat{\mathbb{V}}[\hat{\beta}] = \hat{\sigma}^2(X'X)^{-1} = s^2(X'X)^{-1} = \frac{1}{n-k}\sum_i r_i^2\ (X'X)^{-1}$, and thus $\text{SD}[\hat{\beta}]$ by $\text{SE}[\hat{\beta}] = \frac{1}{\sqrt{n-k}}\sqrt{\sum_i r_i^2}(X'X)^{-\frac{1}{2}}$

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 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 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"[14]. Supposing 100 bootstrap resamples, we can obtain 100 estimates and estimate $\text{SE}[\hat{\theta}]$ by their standard deviation.

Advantages and limits:

+ 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

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

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

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 numerical **test-statistic**, computed from the data, whose probability distribution is connected to our question of interest.

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

1. write H0
2. design a test statistic T that summarizes the deviation of the data from what would be expected under H0, and has a specific distribution under H0

   Ex: – a t-test is a test in which the test statistic has a Student's $\mathcal{T}$-distribution under H0
   – an F-test is a test in which the test statistic has an $\mathcal{F}$-distribution under H0
3. compute the realized value of T for our data
4. look whether it falls in the tails of the distribution. That would mean it is very unlikely given H0. Therefore we can reasonably reject H0.

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.

1. write H0 — we define the null hypothesis – the hypothesis to nullify – H0: $\beta_j = 0$
2. design T — we define the t-statistic $t_\beta = \dfrac{\beta - h_0}{\mathrm{SD}(\beta)} \underset{h_0}{\sim} \mathcal{T}_{n-2}$
3. compute T — $t_{\hat{\beta}} = t_\beta(\text{observed data}) = \dfrac{\hat{\beta} - h_0}{\mathrm{SE}(\hat{\beta})}$
4. interpret — we define the *2-sided*[15] *p-value* $= \Pr\left[\text{observing a } T > |t_{\hat{\beta}}|\right]$ under H0, i.e., the probability of observing data as extreme as that actually observed, assuming H0.[16]

   p-value small $\iff$ $t_{\hat{\beta}}$ falls in the tail of the Student's $\mathcal{T}$-distribution
   $\implies$ observing our $t_{\hat{\beta}}$ is highly unlikely under H0
   $\implies$ reject H0
   $\implies$ there is a relationship between y and x.

   In econometrics, the standard approach is to dichotomize the evidence using a p-value threshold. The *significance level* $\alpha = 5\%$ is selected. If $p < 0.05$ then $\hat{\beta}$ is "statistically significant", if $p > 0.05$ it is not.

[15] We can actually use the test-statistic T to carry out two different tests:

– a two-sided test: if we want to test for the possibility of the relationship in both directions. H0: $\beta_j = 0$, H1: $\beta_j \neq 0$. Both tails of the test-statistic's distribution constitute therefore the "critical region", each containing $\frac{\alpha}{2}$ of the values.
By default, statistical packages report the two-sided p-values.

– a one-sided test: to test for the possibility of the relationship only in one direction. E.g.: H0: $\beta_j = 0$, H1: $\beta_j > 0$. Only the right tail of T's distribution makes the critical region, containing α of the values.
Only z- and t-tests can accomodate one-sided tests. F-tests, χ_2-tests... cannot as their distributions are not symmetric.

[16] ⚠ The p-value is often misinterpreted to be the probability that H0 is true, when it is the probability of observing data as extreme or more extreme than that actually observed, assuming H0. *p-value* = P(obs | hyp) ≠ P(hyp | obs).

3.3.3 Type I/II errors, size and power

A test can lead to two types of mistakes:

- a **type 1 error** or "false positive": $\{- \mid h_0\}$ reject h_0 when it is actually true ...
- a **type 2 error** or "false negative": $\{+ \mid h_1\}$ accept h_0 when it is actually false ...

We define a test's:

- **size** = probability of erroneously rejecting h_0 $\equiv P\big[\text{type I error}\big] = P\big[-|h_0\big]$
- **power** = probability of correctly rejecting h_0 $\equiv 1 - P\big[\text{type II error}\big] = P\big[-|h_1\big]$

Intuitively, we would like to minimize the size and maximize the power of our test:

- To guarantee a test size inferior to 0.05, we simply need to set the significance level α to 0.05.
- To guarantee a power superior to 0.80, we need a sufficiently large sample zize N or the "Minimum Detectable Effect" will be very high.

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[17]. 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 appropriately: ⚠ 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 bad, and often overestimates of power.

3.3.4 Criticisms of the NHST and 'statistical significance'

The NHST paradigm and the binary concept of 'statistical significance' based on p-value thresholds are heavily criticized, as the sharp point null hypothesis of zero effect is generally implausible and thus uninteresting, and interpreting p-values dichotomously loses a lot of information. In addition, as for any statistical estimate to be 'significant', it has to be at least 2 s.e. from 0, the larger the s.e., the higher the estimate must be, to be publishable. This induces selection bias, statistically significant estimates tend to be overestimates.

Instead, notably according to Andrew Gelman, researchers should interpret p-values continuously, viewing the strength of evidence for H0 as a continuous function of the magnitude of the p-value. Statistical power as a measure of the strength of a study is also flawed, as the narrow emphasis of statistical significance is placed as the primary focus of study design.

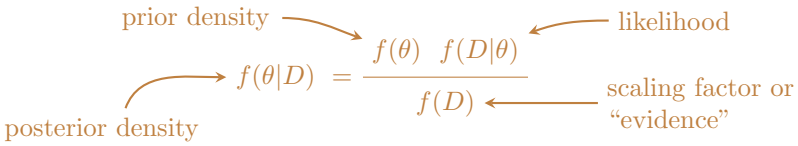
A. Gelman makes further comments not explored here, namely that:

- whereas we think that studies with a high sample size N have higher statistical power, empirical studies have found zero or weak correlations between the two.
- the $t_{\hat{\beta}}$ is not as simple as $\propto \dfrac{\text{signal}}{\text{noise}}$.

[17] Lacasse et al. (2020) is a good–and published!–example of this. Rephrasing their specific x and y 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 Conclusions: *"Our underpowered trial provides no indication that x has a positive or negative effect on y"*.

4 Statistical inference [under a Bayesian approach]

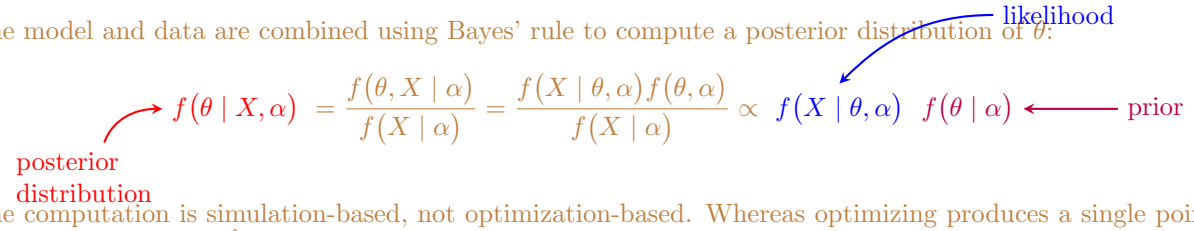
Core idea: θ **is an r.v., and we use Bayes' Theorem to update probability statements (which represent states of beliefs) about** θ **as more evidence (data** D**) becomes available.**



We consider

- x a data point from the sample $X = \{x_1, ...x_n\}$
- θ a parameter of the data point's distribution, $x \sim f(x \mid \theta)$
- α a hyperparameter of θ's distribution, $\theta \sim g(\theta \mid \alpha)$

The model and data are combined using Bayes' rule to compute a posterior distribution of θ:



The computation is simulation-based, not optimization-based. Whereas optimizing produces a single point estimate (the best fit $\hat{\theta}$), since we have uncertainty about the parameters, we describe the entire posterior distribution by producing posterior simulation draws $s = 1, ..., S(\mu_s, \sigma_s)$.

If the ultimate objective of our doing inference is:

- the **estimation** of parameters: we can summarize this posterior distribution using a measurement of central tendency (the mean or median), and credible or "uncertainty" intervals (the Bayesian equivalent of frequentist confidence interval).
- the **prediction** of a new data point $\tilde{x}$: we compute predictive uncertainty by producing posterior simulations, describing the *posterior predictive distribution* $f(\tilde{x} \mid X, \alpha) = \int f(\tilde{x} \mid \theta) f(\theta \mid X, \alpha) d\theta$. We can then summarize it numerically or graphically (median, MAD, hist).

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 MLE.
- A weakly informative family of prior distributions is the normal with mean 0 and scale 2.5 times that of the predictor[18].
- "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. $\rightarrow SE_{\text{Bayes}} <$ both SE_{prior} and SE_{data}

[18] This distribution provides moderate regularization and stabilizes computation. It is the default used in the `R` function `rstanarm::stan_glm()`.

In practice The data are used to update the prior belief by examining the likelihood of the data given a certain value of θ. When the likelihood has an analytical expression, we can combine it with the prior to derive the posterior analytically. Most of the time, there is no such analytical expression. We derive the posterior via Markov Chain Monte-Carlo (MCMC) sampling.

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, 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 generation process.
- Prediction: Use the model to predict the outcomes for new data points.

6 Workflows for inference and prediction

Basic workflow for inference

1. Modeling: reason about the DGP and choose the stochastic model that approximates it best.
   - Consider variable transformations
   - Think about what correlation or structure could be in the errors
     - a group-error structure → consider clustering SEs, bootstrapping SEs, multilevel modeling
     - autocorrelation. Ex: spatial correlation → adjust using Conley s.e.
     - heteroskedasticity. Should always assume it, and thus use "White-corrected" or "robust" SEs
   - Consider doing "Simulated-data experimentation": simulate several fake datasets (under different underlying models), apply your statistical procedure to each, and see what happens. This will clearly show what your method is doing.
2. Estimation & Model validation: check that the underlying assumptions of the model are satisfied; and assess uncertainty in the fit
   - Formal tests. Ex: for Heteroskedasticity: the White Test (tests for heteroskedasticity of an unknown form)
   - Residual plots
     - Normal Q-Q plot = are residuals approximately normally distributed?
     - Residuals against fitted values
3. Application: Use the model to explain the DGP.

Basic workflow for prediction

1. Modeling: Consider several different models and different parameter settings.
2. Model selection: Identify the model with the greatest predictive performance using validation/test sets; select the model with the highest performance on the test set.
3. Application: Predict the outcome for new data, with the expectation that the selected model also generalizes to the unobserved data.

6.1 Model comparison and selection

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[19]. Much of this section is taken from Gelman et al. (2013).

[19] 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.

6.1.1 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?

6.1.2 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 Root Mean Squared Error (RMSE)[20]:

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

  These measures are easy to compute and interpret, but aren't appropriate for models that are far from the normal distribution...

- A more general[21] summary is the *log likelihood* or *log predictive density* (LPD).
  For any data $y{=}y_1, ..., y_m$ produced from the true DGP, i.e., taken from the *unknown* data distribution f, $\text{LPD}(y) \equiv \ln p(y|\theta) = \ln \prod_i p(y_i|\theta)$.

  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 $\tilde{y}_i \sim f$:	
$\text{LPD}(\tilde{y}_i) = \ln p(\tilde{y}_i\|\hat{\theta})$	$\text{LPD}(\tilde{y}_i) = \ln p_{post,y}(\tilde{y}_i) \equiv \int p(\tilde{y}_i\|\theta) p_{post,\theta}(\theta) d\theta$
▷ As new data points are themselves unknown, the expectation:	
$\text{ELPD} \equiv \mathbb{E}_f\big[\text{LPD}(\tilde{y}_i)\big] = \mathbb{E}_f\big[\ln p\big(\tilde{y}_i\|\hat{\theta}\big)\big]$	$\text{ELPD} \equiv \mathbb{E}_f\big[\text{LPD}(\tilde{y}_i)\big] = \mathbb{E}_f\big[\ln p_{post,y}\,(\tilde{y}_i)\,\big]$

  In practice, f and θ are unknown, so we cannot 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

[20] The RMSE is the square root of the variance of the residuals, it can be interpreted as the standard deviation 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 R2, and this would favor overfitting.)
⚠ It is scale-dependent (it has the same unit as y), therefore it can only be compared across models in the same units.
⚠ It is sensitive to outliers (as each error is squared, giving larger errors a disproportionately large effect).

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

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.

And asymptotically, AIC is equal to LOO-CV computed using the MLE, 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.

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: $\ln \hat{L} \equiv \ln p(y|\hat{\theta}_{\text{MLE}})$;
  - adjustment for overfitting: uses the simplest bias correction, based on the asymptotic normal posterior distribution, for which[22] simply subtracting k corrects for the number of parameters:

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

  AIC_c is the AIC corrected for small samples: $AIC_c = -2\ln \hat{L} + 2k\frac{n}{n-k-1} \xrightarrow[n\to+\infty]{} AIC$
  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 $p_{\text{post}}(\theta) = p(\theta|y)$ (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 $p_{\text{post}}(\theta)$ or conditioning on a point estimate $\hat{\theta}$).

In Bayesian CV, fitting the model to y_t yields a posterior distribution for θ: $p_{\text{post}}(\theta) \equiv p(\theta|y_t)$. 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: $\text{LPD}(y_v) \equiv \ln p(y_v|\theta^{\text{"post"}}) \equiv \frac{1}{S}\sum_{s=1}^{S} p(y_v|\theta^s)$

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.

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

- **K-fold CV**
  The data are randomly partitioned into K equal-sized sets. The CV process is repeated K times, each time using one subsample for validation, and the K results are averaged into one estimate:

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

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

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

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., 2013).

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. (2013)'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 modelling

7.1 Agent-based models (ABMs)

Agent-Based Models are computational models[a] that simulate the actions and interactions of autonomous agents, 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. ABMs take a "bottom-up" approach.

ABMs are useful to represent complexities (ex: representing the economy as a complex system with crashes and booms, that emerge from non-linear responses to small changes...)

[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".

Elements of an ABM Most ABMs are composed of: (1) many agents; (2) simple rules about: their individual decision-making process, how they interact, how they learn and adapt; and (3) an environment.

Use in different fields

- In economics: in, agent-based computational economics, ABMs 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
- 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 regard to prediction accuracy.

Statistical inference Because most ABMs have a complicated structure, likelihood functions are generally intractable. Statistical inference must therefore be performed in a likelihood-free context. Two main tools allow that: emulators and approximate Bayesian computation (ABC). Emulators are statistical approximations of computer simulations, that allow one to make predictions of ABM output based on a limited number of simulations – they are therefore particularly interesting for ABMs with substantial run times. ABC is an algorithm that compares simulated data with observed data to estimate posterior distributions for the model parameters. The general approach is to sample some parameter value $\tilde{\theta}$ from a prior distribution, simulate data, and accept the value $\tilde{\theta}$ as a draw from the approximate posterior if the simulated data are sufficiently close to the observed data. This approach requires significantly more ABM simulations than the emulator approach.

Key ideas

Statistics is about uncertainty, and therefore **probability distributions**. Inferential statistics is about learning from data, and 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, MLE,...)*
a. estimation $\implies$ "$\hat{\beta} = ...$"	
b. hypothesis testing $\implies$ "$\hat{\beta}$ 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 & consistency properties (which stem from laws of large numbers (LLNs) and central limit theorems (CLTs)).

A A small library of models

A.1 Common models

All models here are presented in their conditional distributional form. Where the properties of the specific distribution permit, we also write their equivalent form with an error term.
Ex: because of the nice properties of the normal distribution, the linear regression model can be written in terms of the mean and an additive error.

- Classical Linear Regression Model

$$y|X \sim \mathcal{N}(X\beta, \sigma^2 \mathrm{I})$$
$$\iff y = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k + \varepsilon, \quad \varepsilon \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2 \mathrm{I})$$
$$\iff \mu \equiv \mathbb{E}[y] = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k, \quad \varepsilon \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2 \mathrm{I})$$

- Polynomial Regression
  - Ex: LOESS is a nonparametric regression algorithm, in which $\mathbb{E}[y|X]$ at each data point X_i is estimated using a weighted low-degree polynomial regression model that gives higher weights to the neighboring points (in X).

$$\mathbb{E}[y] = \beta_0 + \beta_1 X_1 + \beta_2 X^2 + \varepsilon, \quad \varepsilon \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$$

- Generalized linear model (GLM)
  GLMs are usually used to predict outcomes of bounded or discrete form. A link function $g(\cdot)$ relates $\mathbb{E}[y]$ to the linear predictor vector $X\beta$, and we assume a data distribution $F(y|g^{-1}(X\beta))$.
  - Ex: the linear regression model is a GLM with normal data and 'identity' link.
  - Ex: the logistic regression model is a GLM with binomial data and logit link $\ln\left(\frac{\cdot}{1-\cdot}\right)$.
  - Ex: the Poisson regression model is a GLM with Poisson data and logistic link ln().

$$g\big(\mathbb{E}[y]\big) = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k + \varepsilon, \quad \varepsilon \overset{\text{iid}}{\sim} F(0, ...)$$

- Generalized additive model (GAM)
  GAMs generalize further to allow for $g\big(\mathbb{E}[y]\big)$ to be a *nonlinear* smooth function of each predictor. The space of functions of which h is an element is the "basis".

$$g\big(\mathbb{E}[y]\big) = \beta_0 + h_1(X_1) + ... + h_k(X_k) + \varepsilon, \quad \varepsilon \overset{\text{iid}}{\sim} F_{\text{ExpFamily}}(0, ...)$$

  GAMs penalize the complexity of the model to prevent overfitting the data, by adding a penalty for the size of the coefficients associated with the basis functions.

- Nonparametric models
  Use large numbers of parameters to allow essentially arbitrary curves for the predicted value of $y|X$.

- Multilevel or "hierarchical" models
  Coefficients can vary by group or by situation.

- 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: 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 Limited outcome models

A *limited* dependent variable y, i.e., a y that is categorical or constrained to fall in a certain range, often arises in econometrics. With such data, linear regression is not an appropriate estimation method, as it does not take into account the constraint on possible values of the dependent variable.

Limited y	Appropriate regression models
binary: $y \in \{0, 1\}$	probit, logit
count: $y \in \{0, 1, 2, 3, ...\}$	Poisson regression model, negative binomial model
censored	censored regression models

A.2.1 Binary outcome models

The data $y|X$ is binary, i.e., it follows a Bernoulli distribution:

$$y|X \sim Ber(\pi) = \begin{cases} 1 & \text{with probability } \pi \\ 0 & \text{with probability } 1 - \pi \end{cases}$$

A regression model is therefore formed by expressing the conditional probability $\pi \equiv \mathrm{P}[y{=}1|X]$ as a function of X and β. I.e.[23], a model for binary outcome is:

$$\begin{gathered} y_i|X_i \sim Ber(\pi_i) \\ \pi_i = g^{-1}(X_i'\beta) \\ \iff y_i|X_i \sim Ber\big(g^{-1}(X_i'\beta)\big) \end{gathered}$$

Models

- **Linear probability model**

  $$\pi_i = X_i'\beta + \varepsilon$$

  This model is probably the first one that comes to mind. However, it is not appropriate, as it will not constrain the predicted values to be in [0,1], since the predictor $X_i'\beta$ can take any real value.

- **Logit model = Logistic regression model**

  $$\pi_i = \text{logit}^{-1}(X_i'\beta) \equiv \frac{e^{X_i'\beta}}{1 + e^{X_i'\beta}} \iff \text{logit}(\pi_i) = X_i'\beta$$

  We transform the probability using the logit or "log-odds" transformation $\text{logit}(.) \equiv \ln\left(\frac{.}{1-.}\right)$ which is the inverse of the logistic function, and maps $[0, 1]$ to $[-\infty, \infty]$. This outcome need not be in [0,1], so we can model it as a *linear* function of the covariates. I.e., we have chosen as $g^{-1}()$ the CDF of the logistic distribution: $\text{logit}^{-1}()$.

[23] Where $g^{-1}()$ is a *cumulative distribution function* (to ensure that $0 \leq p \leq 1$).

Interpretation of the coefficients:

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

– odds[24] scale $[0, \infty]$ *"a 1-unit difference in x corres. to a $e^{\hat{\beta}}$ multiplicative difference in odds[y=1]"*

– probability scale [0,1] *"a 1-unit difference in x corres. to a $\frac{\hat{\beta}}{4}$-unit maximum[25] difference in* P[y=1]*"*

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

$$\begin{aligned}\hat{\theta}_{\text{MLE}} = \underset{\theta}{\text{argmax}}\ \log\Big(L\big(y|X,\theta\big)\Big) &= \underset{\theta}{\text{argmax}}\ \log\left(\prod_{i=1}^{n}\pi_i^{y_i}(1-\pi_i)^{1-y_i}\right)\\ &= \underset{\theta}{\text{argmax}}\ \sum_{i=1}^{n} y_i \ln(\pi_i) + (1-y_i)\ln\big(1-\pi_i\big)\\ &= \underset{\theta}{\text{argmax}}\ \sum_{i=1}^{n} y_i \ln\big(F(X_i'\beta)\big) + (1-y_i)\ln\big(1-F(X_i'\beta)\big)\end{aligned}$$

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.*

- **Poisson regression model**

  We assume $y|X$ follows a Poisson distribution. The Poisson distribution is characterized by a single parameter $\lambda > 0$ (the mean rate of occurrence of the event), s.t. $P[y_i{=}y|\lambda] = \frac{e^{-\lambda}\lambda^y}{y!}$, which further implies $\mathbb{E}[y_i] = \mathbb{V}[y_i] = \lambda$. Thus only λ is to be explained by the predictors[26]. Therefore, the general Poisson regression model is:

$$\begin{aligned} y_i|X_i &\sim Pois\big(\lambda_i\big)\\ \lambda_i &= g^{-1}(X_i'\beta)\\ \iff \mathbb{E}[y_i|X_i] &= \mathbb{V}[y_i|X_i] = g^{-1}(X_i'\beta)\end{aligned}$$

  A common choice of link function $g()$ is ln(), s.t. we fit the regression model:

$$\mathbb{E}[y_i|X_i] = \mathbb{V}[y_i|X_i] = \exp(X_i'\beta)$$

[24] 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.

[25] ⚠ the logistic function $\text{logit}^{-1}()$ is curved, so the expected difference in P[$y = 1$] from a given difference in x is not a constant along x. The slope of the logistic regression curve is steepest at its halfway point ($\text{logit}^{-1}() = 0.5$) and is $\beta/4$. I.e., the largest change in π from a 1-unit change in x is $\beta/4$.

[26] Unlike the normal distribution, there is no σ parameter to be fit; the Poisson distribution has its own scale of variation.

Estimation by Maximum Likelihood:

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

Interpretation of the coefficients:

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

– relative risk[27] scale $[0,\infty]$ *"a 1-unit difference in x corresponds to a $e^{\hat{\beta}}$ multiplicative difference in μ."*

One big limitation of the Poisson model is that it implies equi-dispersion: $\text{Var}[y_i|x_i] = \mathbb{E}[y_i|x_i]$, whereas we often see overdispersion in the data (ex: a few traders will do many trades, many traders will do a few). Some softwares (e.g., `R`) have packages that permit Poisson regression with an adjustment for overdispersion. Or we can also turn to the negative binomial distribution, which can accommodate overdispersion.

- **Negative binomial model**
  The negative binomial $NB(p,r)$ distribution includes an additional parameter $r > 0$ to capture overdispersion, s.t. that it does not impose mean = variance and makes for a generalization or robust alternative to the Poisson. It has larger variance than the Poisson for small r (overdispersion), and converges to Poisson as $r \to \infty$.

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

$$y_i|X_i \sim NB\left(\mu_i,\phi\right)$$

$$\mu_i = g^{-1}(X_i'\beta)$$

  Estimation by Maximum Likelihood:

$$\begin{aligned}\hat{\theta}_{\text{MLE}} = \underset{\theta}{\text{argmax}}\ \log L\big(y|X,\theta\big) &= \underset{\theta}{\text{argmax}}\ \log \prod_{i=1}^{n} P[y_i|X_i,\theta] \\ &= \underset{\theta}{\text{argmax}}\ \sum_{i=1}^{n} \ln ...\end{aligned}$$

Modelling heterogeneity

Testing homogeneity in slopes The Hausman and Swamy's tests

[27] Or "incidence rate ratio".

References

Gelman, A. and Greenland, S. Are confidence intervals better termed "uncertainty intervals"? BMJ, 366, Sept. 2019. doi: 10.1136/bmj.l5381.

Gelman, A., Hwang, J., and Vehtari, A. Understanding predictive information criteria for Bayesian models. Statistics and Computing, 24, 07 2013. doi: 10.1007/s11222-013-9416-2.

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. Randomized trial of nocturnal oxygen in chronic obstructive pulmonary disease. New England Journal of Medicine, 383(12):1129–1138, 2020. doi: 10.1056/NEJMoa2013219.

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