



Applied Microeconometrics

Fundamentals

Contents

Motivation	3
1 Statistical models	4
1.1 ⊃ Microeconometrics models	4
1.2 ⊃ Regression models	4
1.3 ⊃ Non-/Semi-/Parametric models	5
2 Data	7
2.1 Types of observational data	7
2.2 Sampling procedures	7
3 Statistical inference [under a frequentist approach]	8
3.1 Frequentist vs Bayesian inference	8
3.2 Estimation	9
3.2.1 Regression analysis	9
3.2.2 Estimators	9
3.2.3 Estimator properties	11
3.2.4 Uncertainty in the estimate: computing SEs & CIs	11
3.3 Hypothesis testing	14
3.3.1 Statistical tests	14
3.3.2 Null Hypothesis Significance Testing (NHST) paradigm	14
3.3.3 Type I/II errors, size and power	15
3.3.4 Criticisms of the NHST and ‘statistical significance’	16
4 Statistical inference [under a Bayesian approach]	17
4.1 Steps of Bayesian inference	17
4.2 Choosing θ ’s prior distribution	18
4.3 Estimating θ ’s posterior distribution	18
5 Prediction	20
6 Workflows for inference and prediction	21
6.1 Model comparison and selection	21
6.1.1 Comparing nested models: F tests	21
6.1.2 Comparing non-nested models: IC, CV	22
7 Other branches of statistical modeling	25
7.1 Statistical Inference Using Agent-based models (ABMs)	25
Key ideas [one pager]	26
References	27

Appendix A A small library of regression models	28
A.1 Expanding from the CLRM	28
A.2 Limited outcome models	29
A.3 Multilevel models	32
A.4 Nonparametric models	34

*Disclaimer: Sections and lines in brown correspond to content which is **very much** ‘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., prediction vs 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 methods of applied (micro)economists, 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 them 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 $\{y_i, X_i\}_{i=1}^n$.

- What distinguishes it from other mathematical models is that it is non-deterministic: some variables are stochastic or “random”, they have probability distributions.
- 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.

^aFormally, 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).

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).

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

- **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)**.

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

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 .

It has 4 components: y, X , unknown parameters β and an error term e .

The variables $\{y, x_1, \dots, x_k\}$ 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 distribution¹ of y , given X** .

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

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

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

In Bayesian inference, parameters are considered random variables, therefore the data generating process is written conditional on β : $y = f(X, e|\beta)$. In frequentist inference, the parameters are considered fixed, therefore we write $y = f(X, e, \beta)$.

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

$$y_i = X_i' \beta + e_i = \beta_0 + \sum_{j=1}^k \beta_j x_{ij} + e_i, \quad e_i \stackrel{\text{iid}}{\sim} (0, \sigma^2), \quad i = 1, \dots, n$$

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

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

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) \rightarrow *overspecified* model;
- excluding an x that is theoretically *relevant* (it may cause y) \rightarrow *underspecified* model.

1.3 \supset 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.² 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,³ 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 $\{\mu, \sigma\}$.

²Recall 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 $\mathcal{P} = \{P_\theta | \theta \in \Theta \subseteq \mathbb{R}^k\}$.

³I.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.⁴ 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: $\mathbb{E}[e] = 0$ and $\mathbb{V}[e] = \mathbb{E}[ee'] = \Omega$.

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

⁴Nonparametric 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: <ul style="list-style-type: none"> • inspect the entire population: perform a census; • inspect a subset: take sample data S_t from the population probability distribution $F(W_t \theta_t)$.

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**.⁵ 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 $N_1 \neq N_2, \dots, N_L$ units. Simple random samples of sizes n_1, n_2, \dots, 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.

⁵“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 *inferring* 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,⁶ 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, and 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 <i>probability</i> \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	
$\implies \theta$ is <i>fixed</i> . We can't assign \mathcal{P} s to events such as $\theta \leq \tilde{\theta}$. We handle our uncertainty in the value of θ by limiting error rates (over imaginary experiments).	$\implies \theta$ is a <i>random variable</i> . 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 $\text{CI}[\theta]$) 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 <i>posterior</i> \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 $\hat{\theta}$.

⁶E.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 $y = f(X, e|\theta)$.

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

^aRecall the definition of a [regression model](#).

^bRegression 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* $f_{y|X}$. We can then choose to focus on its conditional mean $\mathbb{E}[y|X]$, its conditional quantiles $\mathbb{Q}_{y|X}()$...

3.2.2 Estimators

We have a set of observations x_1, \dots, x_n , i.e., a realization of the sample of random variables X_1, \dots, X_n .

An estimand θ 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 $\hat{\theta}_n$ of an estimand is a sample statistic, i.e., a function of the random sample (and therefore a random variable): $T_n = t(X_1, \dots, X_n)$. 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: $t_n = t(x_1, \dots, x_n)$.

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

- **Maximum Likelihood (ML)**⁷

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

⁷The 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:

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

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

• Least Squares (LS)

The fit of a model $y = g(X, e)$ to each data point i is measured by its residual $r_i \equiv y_i - g(x_i, \hat{\beta})$. 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.⁸

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

When the model is linear, i.e., 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:

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

In the simple case of the univariate regression model ($y = \alpha + \beta x + e$), the estimand is $\beta_{\text{OLS}} = \frac{\operatorname{cov}[x, y]}{\operatorname{V}[x]}$ 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 $\beta_{\text{OLS}}^k = \frac{\operatorname{cov}[\tilde{x}_k, y]}{\operatorname{V}[\tilde{x}_k]}$ 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 $w_i \propto \frac{1}{\sigma_i}$.

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

* Generalized Least Squares (GLS)

When errors are heteroscedastic or correlated, i.e., when $x_1, \dots, x_n \stackrel{iid}{\sim} f(x|\theta)$ doesn't hold (the

⁸Indeed, let $e \equiv y - \hat{y}$ be the unobserved error, $L(e)$ the loss. We want to minimize the expected loss $\mathbb{E}[L(e)|X]$. So we look for the fit $g(X, \hat{\beta})$ that minimizes the mean of that function $L()$ of the residuals. For a squared error loss function $L(e) = e^2$, it means minimizing the sum of squared residuals $\sum_i r_i^2$. That fit is the **conditional mean**: $g(X, \hat{\beta}_{\text{LS}}) \equiv \underset{g(\cdot)}{\operatorname{argmin}} \mathbb{E}[(y - g(X, \beta))^2] = \dots = \mathbb{E}[y|X]$.

covariance matrix $\Omega \equiv \text{cov}[e|X]$ is not diagonal with values σ^2 , OLS will again be inefficient. We minimize instead the squared *Mahalanobis length*⁹ 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: $\hat{\beta}_{\text{GLS}} = \underset{\beta}{\text{argmin}} (y - X\beta)' \Omega^{-1} (y - X\beta) = \dots = (X' \Omega^{-1} X)^{-1} X' \Omega^{-1} y$

* **Two-Stage Least Squares (2SLS)**

When regressors are correlated with the errors, we need a matrix of instruments Z s.t. $\mathbb{E}[z_i e_i] = 0$.

$$\hat{\beta}_{\text{2SLS}} = (X' Z (Z' Z)^{-1} Z' X)^{-1} X' Z (Z' Z)^{-1} Z' y$$

• **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**: $g(X, \hat{\beta}_{\text{LSA}}) \equiv \underset{g(\cdot)}{\text{argmin}} \mathbb{E}[|y - g(X, \beta)|] = \dots = \text{med}_{y|X}$.

Least asymmetric absolute error

We can generalize to an asymmetric loss function: $L_\alpha(e) \equiv \begin{cases} (1-\alpha)|e| & \text{if } e < 0 \\ \alpha|e| & \text{if } e \geq 0 \end{cases} = (\alpha - \mathbb{1}\{e < 0\}) \times e$. The optimal fit is the **conditional quantile** $g(X, \hat{\beta}_{\text{LAA}, \alpha}) \equiv \underset{g(\cdot)}{\text{argmin}} \mathbb{E}[L_\alpha(y - g(X, \beta))] = \dots = \mathbb{Q}_{y|X}(\alpha)$.

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:

Objective: fit	Objective: prediction	Optimal estimator/predictor
$\min L(e) \equiv e^2$	<i>predict</i> $\mathbb{E}[y X]$	$\hat{\beta}_{\text{LS}} \equiv \underset{\beta}{\text{argmin}} \sum_i (y_i - g(X_i, \beta))^2$
$\min L(e) \equiv e $	<i>predict</i> $\text{med}_{y X}$	$\hat{\beta}_{\text{LSA}} \equiv \underset{\beta}{\text{argmin}} \sum_i y_i - g(X_i, \beta) $
$\min L_\alpha(e) \equiv (\alpha - \mathbb{1}\{e < 0\})e$	<i>predict</i> $\mathbb{Q}_{y X}(\alpha)$	$\hat{\beta}_{\text{LAA}, \alpha} \equiv \underset{\beta}{\text{argmin}} \sum_i L_\alpha(y_i - g(X_i, \beta))$

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 & 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.

⁹The 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 $\text{SE}[\hat{\theta}]$** = an estimate of the standard deviation of its distribution.
- The **95% Confidence Interval $\text{CI}_{95\%}[\theta]$** = 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.*”¹⁰ 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 $\text{SE}[\hat{\theta}]$

- Example 1: θ is the population mean μ_x , $\hat{\theta}$ is the sample mean \bar{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 .¹¹ 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 β_{OLS} in the multivariate linear regression model.
 - Population: parameter β and error variance σ^2 are unobserved.
 - Sample: We measure the parameter estimate $\hat{\beta} \sim (\beta, \mathbb{V}[\hat{\beta}])$. 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 $\mathbb{V}[\hat{\beta}] = \sigma^2(X'X)^{-1}$. We can consistently estimate the population variance σ^2 by the bias-adjusted *sample* variance $s^2 \equiv \frac{1}{n-k} \sum_i r_i^2$.¹² We can thereby estimate $\mathbb{V}[\hat{\beta}]$ by $\hat{\mathbb{V}}[\hat{\beta}] \equiv s^2(X'X)^{-1} = \frac{1}{n-k} \sum_i r_i^2 (X'X)^{-1}$, and $\text{SD}[\hat{\beta}]$ by $\text{SE}[\hat{\beta}] \equiv \sqrt{\frac{1}{n-k} \sum_i r_i^2 (X'X)^{-1}}$.

¹⁰This 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.

¹¹ \triangle 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 $n \rightarrow \infty$, the second to 0.

¹²The bias here refers to that from the reduced degrees of freedom stemming from estimating the sample means.

2. Confidence Intervals $CI[\theta]$

We want to give a range of estimates for the unknown parameter θ . Consider *the estimate of the centered and standardized estimate* $\frac{\hat{\theta}-\theta}{SE[\hat{\theta}]}$.¹³

- 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:

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

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

$$CI_{95\%}[\mu] \equiv \left[\bar{x} - q_{\mathcal{T}_{n-1}}(0.975) \frac{s}{\sqrt{n}} ; \bar{x} - q_{\mathcal{T}_{n-1}}(0.025) \frac{s}{\sqrt{n}} \right] = \left[\bar{x} \pm q_{\mathcal{T}_{n-1}}(0.975) \frac{s}{\sqrt{n}} \right]$$

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

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

We can thus define the 95% CI:

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

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:

$$CI_{95\%}[\theta] \simeq \left[\hat{\theta} \pm q_{\mathcal{N}}(0.975) SE[\hat{\theta}] \right] = \left[\hat{\theta} \pm 1.96 SE[\hat{\theta}] \right]$$

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 $SE[\hat{\theta}]$ 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.

¹³The standardized estimate is $\frac{\hat{\theta}-\theta}{SD[\hat{\theta}]}$. The scaling term $SD[\hat{\theta}]$ is unknown, therefore we replace it by its estimate $SE[\hat{\theta}]$.

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

- + 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 deviation of the data from what would be expected under H_0 , and has a specific distribution under H_0 . Ex:
 - a t-test is a test in which the test statistic has a Student's \mathcal{T} distribution under H_0 ;¹⁵
 - an F-test is a test in which the test statistic has an \mathcal{F} 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.

¹⁵t-tests are commonly applied for test statistics that would follow a normal distribution if the value of the scaling term were known. When the scaling term is unknown and is replaced by an estimate based on the data, these test statistics (under certain conditions) follow a Student's \mathcal{T} distribution.

1. Write H_0 we define the null hypothesis $H_0: \beta = 0$
2. Design T we define the t -statistic $T \equiv \frac{\hat{\beta} - \beta_0}{\text{SE}[\hat{\beta}]} = \frac{\hat{\beta} - 0}{\text{SE}[\hat{\beta}]}$. If errors are normal, $T \sim \mathcal{T}_{n-2}$.¹⁶
3. Compute T_{obs} $T_{\text{obs}} \equiv T(\text{observed data})$
4. Interpret we define the 2-tailed¹⁷ p -value $\equiv \text{P}(|T| > |T_{\text{obs}}| \mid H_0)$, i.e., the probability of observing data as extreme as that actually observed, assuming H_0 .¹⁸

p -value small $\iff T_{\text{obs}}$ falls in the tail of the Student's \mathcal{T} -distribution
 \iff observing our T_{obs} is highly unlikely under H_0
 \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 $p \leq 0.05$.

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*: $\{+ \mid H_0\}$ don't reject H_0 when could (*overcautious*)

We define a test's:

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

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 size 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.¹⁹ Post-estimation, it is useful to perform a retrospective design analysis and ask: “*Was my study sufficiently*

¹⁶This is a very strong assumption! And it means that if errors aren't normal, the result of the t -test has no interpretation...

¹⁷We 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.

¹⁸⚠ The p -value is often misinterpreted to be the probability that H_0 is true, when it is the probability of observing data as extreme or more extreme than that actually observed, assuming H_0 . $p\text{-value} = \text{P}(\text{obs} \mid \text{hyp}) \neq \text{P}(\text{hyp} \mid \text{obs})$.

¹⁹Lacasse 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 .*”

powered?”, especially if we found a statistically significant non-null effect. But it must be done correctly:
 \triangle 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.
 - \hookrightarrow 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.
 - \hookrightarrow 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.

A. Gelman also notes that using 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.

4 Statistical inference [under a Bayesian approach]

As aforementioned, one of the main distinguishing features of Bayesian inference is the expression of uncertainty using probability. From this paradigm stem new possibilities at every step of inferring properties of a population from data.

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:

$$\text{posterior density} \rightarrow f(\theta | \text{data}) = \frac{\text{prior density} \cdot f(\theta) \cdot \text{likelihood}^{20}}{f(\text{data})} \leftarrow \text{scaling factor or "evidence"}$$

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 applied to each possible value of θ to compute a posterior distribution of θ :

- $y = \{y_1, \dots, y_n\}$ 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

$$\text{posterior} \rightarrow P(\tilde{\theta} | y, \alpha) = \frac{P(\tilde{\theta}, y | \alpha)}{P(y | \alpha)} = \frac{P(y | \tilde{\theta}, \alpha) P(\tilde{\theta} | \alpha)}{P(y | \alpha)} \propto \text{likelihood} \cdot P(y | \tilde{\theta}, \alpha) \cdot \text{prior}$$

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 predictive distribution.

²⁰Note 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 $f(y|\theta, x)$ instead of $f(y|x)$.

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

- *Intuitive interpretation of findings*

Because uncertainty is encoded probabilistically, our uncertainty about θ post-observing data is represented by a distribution of values. We can effortlessly compute a 95% credible interval from the posterior, whose interpretation is the most intuitive: “after seeing the data, there is a 95% chance that this CI contains the true θ .”

In comparison, as the frequentist 95% confidence interval refers to our uncertainty about the *sampling method* — not θ — it is 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...?”

- *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 information 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 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* $f(\tilde{y}|Y, \alpha) = \int_S f(\tilde{y}|\theta) f(\theta|Y, \alpha) d\theta$.

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 estimator.
- A weakly informative family of prior distributions is the normal with mean 0 and scale 2.5 times that of the predictor.²¹
- “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. $\rightarrow \text{SE}_{\text{Bayes}} < \text{both } \text{SE}_{\text{prior}} \text{ and } \text{SE}_{\text{data}}$

4.3 Estimating θ 's posterior distribution

- 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. To estimate the posterior distribution, we can use the method of Markov Chain Monte-Carlo (MCMC): a family of iterative sampling algorithms²² that yield approximate draws from the posterior distribution:

²¹This distribution provides moderate regularization and stabilizes computation. It is the default used in the R function `rstanarm::stan_glm()`.

²²For example, Stan uses Hamiltonian Monte Carlo.

- “*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.
- Most of the time, **fitting a Bayesian model = generating a set of posterior simulations** (representing different possible values of the parameter vector θ).

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. Model building: 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. Inference conditional on a model (estimation and assessment of uncertainty in the fit)
3. Model checking/validation and improvement (checking that the underlying assumptions of the model are satisfied, but also evaluating using prediction, CV...)
 - 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
4. 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.²³ Much of this section is taken from [Gelman et al. \(2014\)](#).

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?

²³In 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.

6.1.2 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):²⁴

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

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

- A more general²⁵ summary is the *log likelihood* or *log predictive density* (LPD). For any data $y=y_1, \dots, y_m$ produced from the true DGP, i.e., taken from the *unknown* data distribution f , $LPD(y) \equiv \ln P(y|\theta) = \ln \prod_i P(y_i|\theta)$.

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

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 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.

²⁴The 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 R^2 , 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).

²⁵It is proportional to the MSE if the model is normal.

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: $\ln \hat{\mathcal{L}} \equiv \ln P(y|\hat{\theta})$;
- adjustment for overfitting: uses the simplest bias correction, based on the asymptotic normal posterior distribution, for which²⁶ simply subtracting k corrects for the number of parameters:

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

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

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 subtract 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 $\theta^1, \dots, \theta^S$. We can then compute the log predictive density for y_v as: $\text{LPD}(y_v) \equiv \ln P(y_v|\theta^{\text{post}}) \equiv \frac{1}{S} \sum_{s=1}^S \ln P(y_v|\theta^s)$

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. The CV process is repeated K times, each time using one subsample for validation, and the K results are averaged into one estimate:

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

- **'Leave-one-out' CV = n-fold CV**

In the extreme case of n partitions, each validation set represents a single data point:

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

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).

²⁶This is also true in the special case of a normal linear model with a uniform prior distribution.

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**.

^aComputational 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.

^bThe 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.

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 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.	<i>(linear regression model w. normal errors, logistic regression model, SEM...)</i>
★ Bring in data ★	
2. Estimate the model, i.e., estimate the conditional distribution. When the specification is parametric, it means estimating parameters.	<i>(OLS, 2SLS, ML,...)</i>
a. estimation \implies “ $\hat{\beta} = \dots$ ”	
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 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](https://doi.org/10.1080/00031305.2021.1900914).
- Gelman, A. and Greenland, S. (2019). Are confidence intervals better termed “uncertainty intervals”? *BMJ*, 366, DOI: [10.1136/bmj.15381](https://doi.org/10.1136/bmj.15381).
- 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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/10.1007/978-3-319-42700-3).

A A small library of regression models

The textbook Classical Linear Regression Model (CLRM) can be generalized in various dimensions, such as:

- the power of the independent variables (\rightarrow polynomial regression);
- the link function relating the linear predictor $X\beta$ to the outcome $\mathbb{E}[y|X]$ (\rightarrow generalized linear model);
- the number of levels in the data (\rightarrow 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 $y|X$.²⁷ 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* $+/ \times$ *error* form.

A.1 Expanding from the CLRM

- Classical Linear Regression Model

$$y|X \sim \mathcal{F}(X\beta, \sigma^2\mathbf{I})$$

$$\iff y = \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k + e, \quad e \stackrel{\text{iid}}{\sim} \mathcal{F}(0, \sigma^2\mathbf{I})$$

$$\iff y = \mathbb{E}[y|X] + e, \quad \mathbb{E}[y|X] = \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k, \quad e \stackrel{\text{iid}}{\sim} \mathcal{F}(0, \sigma^2\mathbf{I})$$

- Polynomial Regression

- Ex: LOESS (locally estimated scatterplot smoothing) 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).

$$\mathbb{E}[y|X] = \beta_0 + \beta_1 X_1 + \beta_2 X^2 + \dots + \beta_p X^p, \quad e \stackrel{\text{iid}}{\sim} \mathcal{F}(0, \sigma^2\mathbf{I})$$

- 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 we assume the outcome to be generated from $\mathcal{F}()$, a linear predictor $X\beta$, and an invertible link function $g()$ that relates $\mathbb{E}[y|X]$ to $X\beta$.

- Ex: the linear regression model is a GLM with normal data and identity link
- Ex: the logistic regression model is a GLM with Bernoulli data and logit link
- Ex: the Poisson regression model is a GLM with Poisson data and log link

$$y|X \sim \mathcal{F}_{\text{ExpFamily}}(\dots), \quad g(\mathbb{E}[y|X]) = \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k$$

- Generalized additive model (GAM)

GAMs generalize further to allow for $g(\mathbb{E}[y|X])$ to be a *nonlinear* smooth function of each predictor. The space of functions of which h is an element is the “basis”.

$$y|X \sim \mathcal{F}_{\text{ExpFamily}}(\dots), \quad g(\mathbb{E}[y|X]) = \beta_0 + h_1(X_1) + \dots + h_k(X_k)$$

²⁷For 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’d make it explicit that distributions of y are conditional on θ .

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.

- 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:

$$\begin{cases} y_i \sim \mathcal{F}(\alpha_{j[i]} + \beta X_i, \sigma_y^2) & \forall i = 1, \dots, N, \quad j = 1, \dots, J \\ \alpha_j \sim \mathcal{F}(\gamma_0 + \gamma_1 W_j, \sigma_\alpha^2) & \forall j = 1, \dots, J \end{cases}$$

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

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

- Nonparametric models

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

Ex: splines

- 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 regression, logit regression
count: $y \in \{0, 1, 2, 3, \dots\}$	Poisson regression, negative binomial regression
interval: $y \in [0, 1]$	fractional response
censored	censored regression

A.2.1 Binary outcome models

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

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

The conditional mean $\mathbb{E}[y|X]$ is equal to the conditional probability $\pi \equiv P(y=1|X)$.²⁸ A regression model is therefore formed by expressing π as a function of X and β :²⁹

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

Models

- **Linear probability model**

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

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 $X_i' \beta$ 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**

$$y_i|X_i \sim \text{Ber}(\pi_i)$$

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

We choose as link function $g()$ the logit $\text{logit}(\cdot) \equiv \ln\left(\frac{\cdot}{1-\cdot}\right)$ (i.e., we choose as $g^{-1}()$ the CDF of the logistic distribution: $\text{logit}^{-1}()$), 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.

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 $\text{log-odds}(y=1)$ ”
- odds³⁰ scale $[0, \infty]$ “a 1-unit difference in x corresponds to a $e^{\hat{\beta}_k}$ multiplicative difference in $\text{odds}(y=1)$ ”
- probability scale $[0,1]$ “a 1-unit difference in x corresponds to a $\frac{\hat{\beta}_k}{4}$ -unit maximum³¹ difference in $P(y=1)$ ”

- **Probit model**

$$y_i|X_i \sim \text{Ber}(\pi_i)$$

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

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: $\text{probit}^{-1}()$), which maps $[0,1]$ to $[-\infty, \infty]$.

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

²⁸As $\mathbb{E}[y|X] = 1 \times P(y=1|X) + 0 \times P(y=0|X) = P(y=1|X)$.

²⁹The function $g^{-1}()$ should be a *cumulative distribution function*, to ensure that $0 \leq \pi_i \leq 1$.

³⁰The odds of success are 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.

³¹ \triangle 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 $\beta/4$.

Estimation by Maximum Likelihood, as the distribution of the data $y|X$ must be the Bernoulli. 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:

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

\triangle 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 $P(\text{winning})$.
- elections: predict vote differential and then map that to $P(\text{winning})$.
- health: model change in blood pressure, and then convert it to the binary disease state $P(\text{hypertension})$.

A.2.2 Count data models

$y_i \in \{0, 1, 2, \dots\}$: number of occurrences of an event. Ex: number of children in a household, number of doctor visits per year.

• 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 $P(y|\lambda) = \frac{e^{-\lambda} \lambda^y}{y!}$, which further implies $\mathbb{E}[y] = \mathbb{V}[y] = \lambda$. The general Poisson regression model is:

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

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: $y_i|X_i \sim Pois(e^{X_i' \beta})$.

Estimation by Maximum Likelihood:

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

A limitation of the Poisson model is that it implies equi-dispersion: $\mathbb{V}[y_i|x_i] = \mathbb{E}[y_i|x_i]$, whereas we often see overdispersion in the data (ex: a few traders will do many trades, many traders will do a few). 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 $\text{NB}(p, r)$ models the number of successes in a sequence of iid Bernoulli(p) trials before r failures occur. Its probability mass function is therefore $P(y|p, r) = \binom{y+r-1}{y} p^y (1-p)^r = \frac{\Gamma(y+r)}{y! \Gamma(r)} p^y (1-p)^r$.³² The NB distribution converges to Poisson as $r \rightarrow \infty$, but has larger variance (overdispersion) for small r , s.t. r is called the “reciprocal dispersion” parameter.

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

$$y_i | X_i \sim \text{NB}(\mu_i, r), \quad \mu_i = e^{X_i' \beta}$$

Estimation by Maximum Likelihood:

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

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 .”

A.3 Multilevel models

Multilevel models are generalizations of classical regression models, and are particularly useful when we want to model within-group variation *and* between-group variation. The general structure is that some coefficients of the lowest-level model are allowed to vary by group, and are themselves given a probability model. This higher (group)-level model can have predictors or not, and has parameters of its own (the “hyperparameters” of the full model) which are also estimated from data.

- Ex: 2-level, varying-intercept model; the group-level model has predictors:

$$\begin{aligned} y_i &\sim \mathcal{F}(\alpha_{j[i]} + \beta X_i, \sigma_y^2) & \forall i = 1, \dots, N, \quad j = 1, \dots, J \\ \alpha_j &\sim \mathcal{F}(\gamma + \delta W_j, \sigma_\alpha^2) & \forall j = 1, \dots, J \end{aligned}$$

- Ex: 2-level, varying-intercept & slope model; intercepts and slopes assumed to covary; the group-level model has no predictors:

$$\begin{aligned} y_{it} &\sim \mathcal{F}(\alpha_{j[i]} + \beta_{j[i]} X_i, \sigma_y^2) & \forall i = 1, \dots, N, \quad j = 1, \dots, J \\ \begin{bmatrix} \alpha_j \\ \beta_j \end{bmatrix} &\sim \text{MVN} \left(\begin{bmatrix} \mu_\alpha \\ \mu_\beta \end{bmatrix}, \begin{bmatrix} \sigma_\alpha^2 & \sigma_{\alpha\beta} \\ \sigma_{\alpha\beta} & \sigma_\beta^2 \end{bmatrix} \right) & \forall j = 1, \dots, J \end{aligned}$$

³²The negative binomial distribution is typically specified in terms of the parameters (p, r) . In a regression framework, it is more intuitive to specify it in terms of its mean $\mu = \frac{pr}{1-p}$ and r . We rewrite the probability mass function as $P(y|\mu, r) = \frac{\Gamma(y+r)}{y! \Gamma(r)} \left(\frac{\mu}{r+\mu} \right)^y \left(\frac{r}{r+\mu} \right)^r$.

- $\sigma_\alpha \rightarrow \infty$ means no pooling: a separate intercept per group. The risk is to overstate the variation between groups, i.e., overfit the data;
- $\sigma_\alpha \rightarrow 0$ means complete pooling: α_j all fixed at a common α (ignores variation betw. groups);
- the multilevel model is a compromise between the two: the group-level parameters α_j are partially pooled toward the group-level mean μ_α , by an amount that depends on the sample size of each group and on σ_α , which is estimated from the data rather than being arbitrarily set to 0 or ∞ .

Inference

- **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 estimates are unbiased for the group-level mean parameters (“fixed effects”) but downward-biased for the group-level variance parameters (“random effects”) (especially when the number of groups is small), because the mean parameters are assumed to be known with certainty when estimating the variance parameters. REML estimation takes account of the number of (fixed effects) parameters estimated, losing 1 degree of freedom for each, and so produces unbiased estimates of variance parameters. However, for model comparison: likelihood ratio tests for REML require exactly the same fixed effects specification in both models. So, to compare models with different fixed effects (a common scenario) with an LR test, ML must be used.

👍 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.

Why use multilevel models?

- *To do correct inferences in the case of correlated or structured errors (e.g., with time series, spatial correlation, networks...).* Traditional regression techniques treat the units of analysis as independent observations. Failing to recognize hierarchical structures will underestimate standard errors of regression coefficients (esp. that of higher-level explanatory variables) hence overstate statistical significance. A multilevel model can be viewed as a classical regression with correlated errors:

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

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

- *In causal inference, to adjust for unmeasured covariates more efficiently by partial pooling.* In observational studies it is typically necessary to adjust for unobserved differences between treated and control items. One way to do that if the observations are structured is to use group fixed effects, i.e., estimate separate regressions for each group (*no pooling*).³³ Especially when the number of observations per group is small, including group indicators in a least squares regression will give very noisy estimates. *Partially pooling* the varying coefficients, via multilevel regression, will give more efficient estimates of these J regression lines.
- *In causal inference, to model variation in treatment effect.* If there is interest not just in an average treatment effect but also in how the effect varies across the population, with a multilevel model, we can model the expected treatment effect 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: in order to eventually generalize to a larger population we would need to account for this somehow and could do so using multilevel regression and poststratification (MRP).
 - To estimate y for particular groups: with multilevel modeling we can get reasonable estimates even for groups with small sample sizes; this is difficult with classical regression.
 - To predict a new observation in a new group: multilevel regression enables to propagate the uncertainty about the new group into the uncertainty about the new individual in this group; we didn't have this distinction in classical regression. And with a fixed effects model, inferences cannot be made beyond the groups in the sample.

Multilevel modeling enables to quantify sources of variation, which is useful to make generalizations, e.g., when considering how a treatment will scale in the real world.

Note that when the number of groups J is small, it is difficult to estimate the between-group variation precisely, multilevel modeling adds little to no-pooling models.

More generally, multilevel modeling can be thought of as a method or even a framework for combining different sources of information.

A.4 Nonparametric models

Splines A spline is a piecewise polynomial function with additional constraints:

- piecewise polynomial: we divide X into intervals, fit each interval with a separate polynomial of low-degree m . There is seldom any good reason to go beyond a cubic spline ($m = 3$).
- additional constraints: we impose that the $m - 1$ first derivatives are continuous at the knots (equal values on both sides), so that the total fit is smooth. \implies For K knots, there are $K + 4$ total degrees of freedom.

As the behavior of polynomials tends to be erratic near the boundaries, we can use a “natural” or “restricted” spline which further imposes that the polynomial fits beyond the boundary knots are linear. I.e., for a natural cubic spline, it sets the cubic and quadratic parts there to 0. The total degrees of freedom are reduced from $K + 4$ to K .

Restricted cubic splines are an easy way of including an explanatory variable in a smooth non-linear way in a wide variety of models. With low-degree polynomials, we don't observe high oscillations of the curve around the data.

In practice: restricted cubic splines are just a transformation of an explanatory variable. This transformed variable can then be entered in any regression command (fixest, logit, glm...)

³³For example, with panel data, each unit has T observations and is therefore itself a “group”.

Interpreting the results:

- graph the adjusted predictions: the predicted outcome against the spline variable. If there are other covariates, show the predicted outcome for an observation with typical values on the other covariates.
- graph the marginal effects, i.e., the spline's first derivative. = How much does the predicted outcome change for a unit change in the explanatory variable?