Introduction
Definitions, probabilities, distributions, estimators

Simon Fraser University
ECON 483
Summer 2023

Disclaimer

I do not allow this content to be published without my consent.

Regression

- In this course, we are interested in the following model:

$$\boldsymbol{Y_i = f(X_{i,1},\ X_{i,2}, ...,\ X_{i,K}) + u_i} \tag{1}$$

  where:
  - $\boldsymbol{Y_i}$ is a outcome variable of interest (also called **explained variable** or **dependent variable**)
  - $\boldsymbol{u_i}$ is a **random disturbance**, also called the **error term**
  - $\boldsymbol{X_{i,1}, ..., X_{i,K}}$ are the **explanatory variables** or the **independent variables** or the **covariates** or the **predictors**
- In words: we are interested in the relationship between the explanatory variables and the explained variable
- $\boldsymbol{Y_i}$ and $\boldsymbol{X_{i,1}, ..., X_{i,K}}$ are random variables. We observe realizations of them in the data
- This is called the **population regression equation**

Outline

- Random variables, distributions, moments
- Covariance and correlation
- Conditional probabilities, expectation and variance
- Estimators: Definitions and properties
- Estimation vs prediction
- What is a statistical model
- What is machine learning?

Random variables and probabilities

Random variables

Definition: Random variable

A **random variable** (r.v. for short) is a variable whose value (= realization) depends on outcomes of a **random phenomenon**. It is generally denoted by upper case letters ($\boldsymbol{X}$, $\boldsymbol{Y}$) while their realization are denoted by lower case letters ($\boldsymbol{x}$, $\boldsymbol{y}$). An **event** is a set of one or more outcomes involving random variables.

- Take the event "obtain Heads 5 times out of 7 coin tosses"
- Many outcomes lead to that event:
  - 5 Heads in a row, then 2 Tails
  - 3 Heads, then 2 Tails, then 2 Heads
  - etc (in fact, we could use some combinatorics formulas to figure out how many outcomes corresponds to that event)

Types of random variables

- Random variables can be **discrete** (0, 1, 2, ...) or **continuous** (continuum of values)
- Examples of discrete random variables: Age/Number of kids of the next person you meet, number obtained after throwing a die, score obtained at a roulette/Black jack table,...
- Examples of continuous random variables: Weight/height, temperature, score in a course, ...
- **Qualitative** variables (Yes/No, Man/Woman, Smoker/Non-smoker, car/bicycle/train/bus) can be turned into **quantitative** variables: 1 for Man, 0 for Woman, etc $\Rightarrow$ back to a **discrete** random variable that a software can understand!

Probabilities

Definition: Probability

A **probability** is a number that represents the proportion of the time an event occurs. For an event $\mathbf{A}$, it is denoted $\mathbb{P}(\boldsymbol{A})$.

Examples:

- Take the event $\mathbf{A}$ = "get a number bigger or equal to 3 when throwing a die". Then $\mathbb{P}(\boldsymbol{A}) = \mathbf{4/6}$ (get a 3, a 4, a 5 or a 6. 4 possible successes out of 6 possible cases)
- Let $\boldsymbol{X}$ denote the weight of a person and $\boldsymbol{A = X > 150}$ lbs. Then $\mathbb{P}(\boldsymbol{A})$ is the probability that a person weighs more than 150 pounds

Probability distributions: Discrete random variables

Probability mass function (pmf)

The **probability mass function** of a discrete random variable is the list of all possible values of the variable along with the probability they occur.

- Example: The probability mass function of a die throw assigns a probability of 1/6 to each number we can obtain: 1, 2, 3, 4, 5, 6
- The probability mass function of a fair coin toss assigns a probability of 50% to each outcome: Heads or Tails

Definition: Cumulative distribution function (cdf)

The **cumulative distribution function** is the function that gives the probability that a random variable is lower or equal to a particular value.

- Example: The probability that a die throw returns a number lower or equal to 5, i.e. $\mathbb{P}(\boldsymbol{X \leq 5})$

Probability distributions: Continuous random variables

- Since a continuous variable can take an infinite amount of values, $\mathbb{P}(\boldsymbol{X} = \boldsymbol{x}) = \mathbf{0}$ and the previous definition does not make sense here

Definition: Probability density function (pdf)

The **probability density function (pdf)** is the function representing the **relative likelihood** of a value a random variable can take vs others. The area under the pdf between two points shows the **probability** that the random variable equals a value between these two points.

Definition: Cumulative distribution function (cdf)

The **cumulative distribution function (cdf)** is the function that gives the probability that a random variable is lower or equal to a particular value.

Common distributions

- **Uniform distribution**: It gives the same probability to each possible number $\boldsymbol{X}$ can take. Exists in the discrete case and the continuous case
- **Bernoulli distribution**: Corresponds to r.v. that can be equal to 1 with some probability, and 0 otherwise. Qualitative variables turned into binary (0-1) variables are Bernoulli r.v.
- **Binomial distribution**: Measures the probability of having $\boldsymbol{k}$ successes out of $\boldsymbol{n}$ independent trials involving a Bernoulli r.v. Example: Getting Heads $\boldsymbol{k = 2}$ times out of $\boldsymbol{n = 10}$ coin flips
- **Geometric distribution**: Measures the probability of the first success after $\boldsymbol{k}$ attempts
- **Hypergeometric distribution**: Measures the probability of having $\boldsymbol{k}$ successes out of $\boldsymbol{n}$ draws from a finite population of size $\boldsymbol{N}$ with $\boldsymbol{K}$ objects corresponding to a success (draw $\boldsymbol{k}$ blue balls out of $\boldsymbol{n}$ draws in a urn with $\boldsymbol{N}$ balls overall and $\boldsymbol{K}$ blue balls)
- **Normal distribution**: Continuous distribution that describes many variables in real life, and used in many statistical theorems

The normal distribution

- Also called **Gaussian** distribution after Carl Friedrich Gauss (genius mathematician)
- Invented by Gauss and Laplace (French) almost simultaneously, but independently
- Important parameters: The mean $\boldsymbol{\mu}$ and the variance $\boldsymbol{\sigma^2}$. They are enough to characterize the full probability distribution so it is denoted $\boldsymbol{X \sim \mathcal{N}(\mu, \sigma^2)}$.
- Notable features:
  - Symmetric and centered around the mean (the mean is the same as the median)
  - The sum of normal r.v. is itself a normal r.v.
  - The "Bell curve": Higher probability of being around the mean than around the extremes
- Equation of the density function: $\boldsymbol{\phi(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(x-\mu)^2}{2\sigma^2})}$
- No need to remember the density function expression

Expectation of a random variable

- In order to understand the distribution of a r.v., it is relevant to look at measures of location and dispersion
- These measures are called moments: of order 1 like the expectation, order 2 like the variance, and more

Definition: Expectation

The expectation of $\boldsymbol{X}$, denoted $\mathbb{E}[\boldsymbol{X}]$ (or often $\boldsymbol{\mu_X}$), is the average value $\boldsymbol{X}$ takes.

- For a discrete r.v : $\mathbb{E}[\boldsymbol{X}] = \sum_{i=1}^{n} \boldsymbol{x_i} \mathbb{P}(\boldsymbol{X} = \boldsymbol{x_i})$
- For a continuous r.v : $\mathbb{E}[\boldsymbol{X}] = \int_{-\infty}^{+\infty} \boldsymbol{x f(x) dx}$ where $\boldsymbol{f(x)}$ is the pdf of $\boldsymbol{X}$ (no need to know this one)

- For a 6-faced die throw ($\boldsymbol{X}$ is the obtained number): $\mathbb{E}[\boldsymbol{X}] = \frac{1}{6} \times \mathbf{1} + \frac{1}{6} \times \mathbf{2} + \mathbf{...} + \frac{1}{6} \times \mathbf{6} = \mathbf{3.5}$
- Note: When $\boldsymbol{X}$ is discrete, the expectation of $\boldsymbol{X}$ is often not equal to any value that $\boldsymbol{X}$ can take (e.g. a die throw)

Properties of the expectation

Properties of expectations

Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be two random variables, and $\boldsymbol{a}$ and $\boldsymbol{b}$ be two constants. Then:

- $\mathbb{E}[\boldsymbol{aX}] = \boldsymbol{a}\mathbb{E}[\boldsymbol{X}]$
- $\mathbb{E}[\boldsymbol{X} + \boldsymbol{Y}] = \mathbb{E}[\boldsymbol{X}] + \mathbb{E}[\boldsymbol{Y}]$ (linearity)
- $\mathbb{E}[\boldsymbol{a}] = \boldsymbol{a}$ ($\boldsymbol{a}$ is not random)
- In general: $\mathbb{E}[\boldsymbol{XY}] \neq \mathbb{E}[\boldsymbol{X}]\mathbb{E}[\boldsymbol{Y}]$

Variance of a random variable

Definition: Variance

The variance of $\boldsymbol{X}$, denoted $\mathbb{V}[\boldsymbol{X}]$ (often denoted $\boldsymbol{\sigma^2}$), is the expectation of the squared deviation from the mean:

$$\begin{aligned}\mathbb{V}[\boldsymbol{X}] &\equiv \mathbb{E}[(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])^2] \\ &= \sum_{i=1}^{n} (\boldsymbol{x}_i - \mathbb{E}[\boldsymbol{X}])^2 \mathbb{P}(\boldsymbol{X} = \boldsymbol{x}_i)\end{aligned}$$

An **equivalent** formula is

$$\mathbb{V}[\boldsymbol{X}] \equiv \mathbb{E}[\boldsymbol{X}^2] - (\mathbb{E}[\boldsymbol{X}])^2$$

and the **standard deviation** of $\boldsymbol{X}$ (often denoted $\boldsymbol{\sigma}$) is the square root of the variance.

Properties of the variance

- Because of the square, $\mathbb{V}[\boldsymbol{X}]$ is expressed in the square of the unit of $\boldsymbol{X}$. But the standard deviation is expressed in the original unit
- The second formula is more useful for r.v. that take a few values

Properties of the variance

Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be two random variables, and $\boldsymbol{a}$ and $\boldsymbol{b}$ be two constants. Then:

- $\mathbb{V}[\boldsymbol{aX}] = \boldsymbol{a^2}\mathbb{V}[\boldsymbol{X}]$ (the constant can come out, but since the variance is a square, we need to square the constant)
- $\mathbb{V}[\boldsymbol{a}] = \boldsymbol{0}$ ($\boldsymbol{a}$ is a constant, so its average is $\boldsymbol{a}$, and it never varies)
- $\mathbb{V}[\boldsymbol{X+Y}] = \mathbb{V}[\boldsymbol{X}] + \mathbb{V}[\boldsymbol{Y}] + \boldsymbol{2cov(X,Y)}$

Two random variables

- Often, we do not deal with one random variable but two or more
- Some random variables are **correlated**, i.e. the variation of one influences the variation of the other (especially in Economics!)
- Examples:
  - The proportion of smokers is not the same among men vs women. So the probability of meeting a smoker is different if it is a man we meet or a woman
  - The proportion of smokers is (definitely!) not the same among Canadian vs French people. So the probability of meeting a smoker is different if we are in Canada or in France
  - Consumption is not the same between rich and poor households. So Consumption is affected by income
- We need to look at measures of **covariation** between two random variables

Covariance between two random variables

Definition: Covariance

The **covariance** between $\boldsymbol{X}$ and $\boldsymbol{Y}$, denoted $\boldsymbol{Cov(X,Y)}$ (or $\boldsymbol{\sigma_{XY}}$), is the expectation of the products of the deviations from the mean:

$$\begin{aligned} \boldsymbol{Cov(X,Y)} &\equiv \mathbb{E}[(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])(\boldsymbol{Y} - \mathbb{E}[\boldsymbol{Y}])] \\ &= \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \sum_{\boldsymbol{j=1}}^{\boldsymbol{n}} \mathbb{P}(\boldsymbol{Y} = \boldsymbol{y_i}, \boldsymbol{X} = \boldsymbol{x_j})(\boldsymbol{x_j} - \mathbb{E}[\boldsymbol{X}])(\boldsymbol{y_i} - \mathbb{E}[\boldsymbol{Y}]) \end{aligned}$$

An **equivalent** formula is

$$\boldsymbol{Cov(X,Y)} \equiv \mathbb{E}[\boldsymbol{XY}] - \mathbb{E}[\boldsymbol{X}]\mathbb{E}[\boldsymbol{Y}]$$

Properties of the covariance

Properties of the covariance

Let $\boldsymbol{X}$, $\boldsymbol{Y}$ and $\boldsymbol{Z}$ be three random variables, and $\boldsymbol{a}$ and $\boldsymbol{b}$ be two constants. Then:

- A negative covariance means the two random variables evolve in opposite directions
- A positive covariance means the two random variables evolve in the same direction
- $\boldsymbol{Cov(aX, Y) = a\,Cov(X, Y)}$
- $\boldsymbol{Cov(X + Z, Y) = Cov(X, Y) + Cov(Z, Y)}$ **(linearity)**
- $\boldsymbol{Cov(aX + Z, bY) = ab\,Cov(X, Y) + b\,Cov(Z, Y)}$
- $\mathbb{V}[\boldsymbol{X + Y}] = \mathbb{V}[\boldsymbol{X}] + \mathbb{V}[\boldsymbol{Y}] + \boldsymbol{2Cov(X, Y)}$
- $\boldsymbol{Cov(X, X)} = \mathbb{V}[\boldsymbol{X}]$ (check the covariance formula when $\boldsymbol{X}$ is in both places!)

Correlation between two random variables

- The scale of the covariance is tricky: It is the unit of $\boldsymbol{X}$ times the unit of $\boldsymbol{Y}$
- A "unit-free" measure is the **correlation** between $\boldsymbol{X}$ and $\boldsymbol{Y}$

Definition: Correlation

The **correlation** between $\boldsymbol{X}$ and $\boldsymbol{Y}$, denoted $\boldsymbol{Corr(X,Y)}$ (or $\boldsymbol{\rho_{XY}}$), is defined as:

$$\boldsymbol{Corr(X,Y)} \equiv \frac{\boldsymbol{Cov(X,Y)}}{\sqrt{\mathbb{V}[\boldsymbol{X}]\mathbb{V}[\boldsymbol{Y}]}}$$

Properties of the correlation

- The units at the top and bottom cancel!
- $\boldsymbol{Corr(X,Y)}$ is always between **-1** and **1**
- If $\boldsymbol{Corr(X,Y)=0}$ then we say $\boldsymbol{X}$ and $\boldsymbol{Y}$ are **uncorrelated**
- Note: if $\boldsymbol{Cov(X,Y)=0}$ then $\boldsymbol{Corr(X,Y)=0}$
- Correlation does not imply causality! The whole difficulty of Econometrics is to disentangle the two: Correlation is easy to find (just a formula), but causality is more challenging to establish, and requires a deeper analysis
- Check out the **spurious correlations website** for weird correlations that have nothing to do with causality

Conditional probabilities

Conditional probabilities

Definition: Conditional probability

The **conditional probability** of a random variable $\boldsymbol{Y}$ **given** the value of a random variable $\boldsymbol{X}$, denoted $\mathbb{P}(\boldsymbol{Y} = \boldsymbol{y}|\boldsymbol{X} = \boldsymbol{x})$ or $\mathbb{P}(\boldsymbol{Y}|\boldsymbol{X})$ is the probability of $\boldsymbol{Y}$ taking a value $\boldsymbol{y}$ when the value of $\boldsymbol{X}$ is fixed at some value $\boldsymbol{x}$.

- Example: Consider the world as being composed of circles vs non circles, and smokers vs non smokers. We know that **20**% of people are circles, and **80**% are not. Among circles, **30**% are smokers and among non circles, **60**% are smokers. Let $\boldsymbol{X}$ be the circle status and $\boldsymbol{Y}$ be the smoking status (we could make them equal to **0** or **1**). We know:
  - $\mathbb{P}(\boldsymbol{Y} = \textbf{smoker}|\boldsymbol{X} = \textbf{circle}) = \textbf{0.3}$ and $\mathbb{P}(\boldsymbol{Y} = \textbf{non smoker}|\boldsymbol{X} = \textbf{circle}) = \textbf{0.7}$
  - $\mathbb{P}(\boldsymbol{Y} = \textbf{smoker}|\boldsymbol{X} = \textbf{non circle}) = \textbf{0.6}$ and $\mathbb{P}(\boldsymbol{Y} = \textbf{non smoker}|\boldsymbol{X} = \textbf{non circle}) = \textbf{0.4}$
  - $\mathbb{P}(\boldsymbol{X} = \textbf{circle}) = \textbf{0.2}$ and $\mathbb{P}(\boldsymbol{X} = \textbf{non circle}) = \textbf{0.8}$

Conditional probabilities (cont'd)

- In words: we know the probability of meeting a circle or non circle, and the probability of meeting a smoker **among (or given)** circles or non circles
- What we don't know:
  - $\mathbb{P}(\boldsymbol{Y = \text{smoker}})$ and $\mathbb{P}(\boldsymbol{Y = \text{non smoker}})$
  - $\mathbb{P}(\boldsymbol{X = \text{circle}|Y = \text{smoker}})$ and $\mathbb{P}(\boldsymbol{X = \text{circle}|Y = \text{non smoker}})$
- In words: we don't know the probability of meeting a smoker (we don't know their total proportion), nor the probability of meeting a circle **among (or given)** smokers or non smokers

Conditional expectation

- The same way we consider conditional probabilities, we can look at conditional expectations, i.e. the expectation of a r.v. when another r.v. is fixed
- Example: Let $\boldsymbol{Y}$ be the temperature in Celsius degrees. Let $\boldsymbol{X}$ be either Vancouver, or Hasparren (beautiful town in the Basque country). The average annual temperatures are $\mathbb{E}[\boldsymbol{Y}|\boldsymbol{X} = \textbf{Hasparren}] = \boldsymbol{13.5°C}$ and $\mathbb{E}[\boldsymbol{Y}|\boldsymbol{X} = \textbf{Vancouver}] = \boldsymbol{11°C}$

Definition: Conditional expectation

Consider the random variables $\boldsymbol{X}$ and $\boldsymbol{Y}$. The expectation of $\boldsymbol{Y}$ conditional on $\boldsymbol{X}$ is defined as:

$$\mathbb{E}[\boldsymbol{Y}|\boldsymbol{X} = \boldsymbol{x}] = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{y_i}\mathbb{P}(\boldsymbol{Y} = \boldsymbol{y_i}|\boldsymbol{X} = \boldsymbol{x})$$

- Note: Since it depends on what $\boldsymbol{X}$ is equal to, a conditional expectation is **random**!

The law of iterated expectations

- If Hasparren and Vancouver were the only 2 locations in the world, then $\mathbb{E}[\boldsymbol{Y}]$, the world annual average temperature, would be the average of $\mathbb{E}[\boldsymbol{Y}|\boldsymbol{X} = \textbf{Hasparren}]$ and $\mathbb{E}[\boldsymbol{Y}|\boldsymbol{X} = \textbf{Vancouver}]$
- So the expectation of $\boldsymbol{X}$ is an average of the conditional expectations!

Theorem: Law of iterated expectations

The expectation of a random variable is the expectation of conditional expectations:

$$\mathbb{E}[\boldsymbol{Y}] = \mathbb{E}[\mathbb{E}[\boldsymbol{Y}|\boldsymbol{X}]]$$

- When looking at $\mathbb{E}[\boldsymbol{Y}|\boldsymbol{X}]$, $\boldsymbol{X}$ is fixed. The outside expectation is going over the distribution of $\boldsymbol{X}$

Conditional variance

Definition: Conditional variance

The conditional variance of $\boldsymbol{Y}$ given $\boldsymbol{X} = \boldsymbol{x}$, denoted $\mathbb{V}[\boldsymbol{Y}|\boldsymbol{X}]$, is the expectation of the squared deviation from the mean, conditional on $\boldsymbol{X} = \boldsymbol{x}$:

$$\begin{aligned}\mathbb{V}[\boldsymbol{Y}|\boldsymbol{X}] &\equiv \mathbb{E}[(\boldsymbol{X} - \mathbb{E}[\boldsymbol{Y}|\boldsymbol{X}])^{\boldsymbol{2}}|\boldsymbol{X}] \\ &= \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (\boldsymbol{x_i} - \mathbb{E}[\boldsymbol{Y}|\boldsymbol{X}])^{\boldsymbol{2}})\mathbb{P}(\boldsymbol{Y} = \boldsymbol{y_i}|\boldsymbol{X})\end{aligned}$$

An equivalent formula is

$$\mathbb{V}[\boldsymbol{Y}|\boldsymbol{X}] = \mathbb{E}[\boldsymbol{Y}^{\boldsymbol{2}}|\boldsymbol{X}] - (\mathbb{E}[\boldsymbol{Y}|\boldsymbol{X}])^{\boldsymbol{2}}$$

Population vs sample

- $\mathbb{P}(\boldsymbol{X} = \boldsymbol{x})$, $\mathbb{E}[\boldsymbol{X}]$ and $\mathbb{V}[\boldsymbol{X}]$ are **population** values: They are **not** random
- Note: not all random variables have an expectation or a variance! Check the Cauchy distribution for an eccentric case
- Any statistic derived from a **sample** is random. Give me another sample, and these measures will be different
- Since sample values (like the sample mean) are random, they can have an expectation and their variance is different from 0 (reminder: the variance of a nonrandom variable is 0)
- So $\bar{\boldsymbol{X}}$, $\hat{\boldsymbol{p}}$ or $\hat{\boldsymbol{y}}_{\boldsymbol{i}}$ are random: they change with every sample. This is why we care about their **bias**, **variance** and **consistency**

Estimators

Estimators

- An estimator is a rule to compute an estimate of a given quantity given some data
- It produces no more than a guess, educated or not
- If we had access to the population data, we would use it to find, say, $\mathbb{E}[\boldsymbol{X}]$
- Since we only have access to a sample of it, we are going to use an estimator to compute an estimate
- **Ideally**, the bigger the sample, the closer we should get to the truth
- Estimates are sample values, so an estimator is random, and hence has an **expectation**
- Computing estimates over different samples should tell us something reliable on **average**
- An estimator being random, we would like to use its distribution to infer something about the true value of the parameter of interest

Desirable properties of estimators

Definition: Consistency

An estimator $\hat{\boldsymbol{\theta}}$ of a nonrandom quantity $\boldsymbol{\theta}$ is **consistent** (or **asymptotically unbiased**) if it converges in probability towards the value it estimates:

$$\hat{\boldsymbol{\theta}} \xrightarrow{\mathbb{P}} \boldsymbol{\theta}$$

where $\xrightarrow{\mathbb{P}}$ denotes convergence in probability, i.e. $\forall \boldsymbol{\varepsilon} > \mathbf{0}$, $\mathbb{P}\left(|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}| > \varepsilon\right) \to \mathbf{0}$ as $\boldsymbol{n} \to \boldsymbol{\infty}$

Definition: Unbiasedness

An estimator $\hat{\boldsymbol{\theta}}$ of a nonrandom quantity $\boldsymbol{\theta}$ is **unbiased** if on **average**, it equals the true value of the quantity of interest:

$$\mathbb{E}[\hat{\boldsymbol{\theta}}] = \boldsymbol{\theta}$$

Desirable properties of estimators (cont'd)

Definition: Asymptotic normality

An estimator $\hat{\boldsymbol{\theta}}$ is said to be **asymptotically normal** if, as the sample size $n \to \infty$:

$$\sqrt{n}(\hat{\theta}) \xrightarrow{d} \mathcal{N}(\theta, \Omega)$$

where $\boldsymbol{\theta}$ and $\boldsymbol{\Omega}$ are some nonrandom quantities

Desirable properties of estimators (cont'd)

- Many (most) estimators are **biased**. For some, we can have an idea of the bias direction
- But being **consistent** makes them reliable given the sample used is big enough. If an estimator is consistent, the bigger the sample, the more accurate the estimator, and the closer the estimate is to the true value
- **Asymptotic normality** allows to make **inference** about the true value of the parameter, i.e. to draw probabilistic conclusions about the true parameter, such as **confidence intervals**
- Both **consistency** and **asymptotic normality** rely on having a large sample

Estimation vs prediction

- The whole course is about estimating $\boldsymbol{f(x_{i,1}, x_{i,2}, ..., x_{i,K})}$
- But what assumptions to make or tools to use depends on whether one wants to **estimate** or **predict**
- **Estimation** is about learning about a specific parameter of interest. Estimation problems make assumptions about the nature of the function $\boldsymbol{f()}$ in (1) and look for the effect of specific variables. Often, adding other variables is meant to satisfy assumptions ensuring the accuracy of the estimation method
- **Prediction** is about guessing, predicting the value of the dependent variable given some values of the covariates. So it is about making accurate predictions $\boldsymbol{\hat{y}_i = \hat{f}(x_{i,1}, x_{i,2}, ..., x_{i,K})}$. Whichever variables are improving predictions will be added, they do not have to have a particular relevance theoretically speaking
- The two problems often overlap. But keep in mind what each method covered in the course is used for!

Conditional expectation and MSE

- In general, one wants to find the best function $\boldsymbol{f()}$, i.e. the one that **minimizes** the expected distance between $\boldsymbol{Y_i}$ and $\boldsymbol{f(X_i)}$
- That expected distance is called the **Mean Squared Error**:

$$\boldsymbol{MSE(f(X_i)) \equiv \mathbb{E}[(Y_i - f(X_i))^2]}$$

- By the law of iterated expectations that says $\mathbb{E}[\boldsymbol{Y}] = \mathbb{E}[\mathbb{E}[\boldsymbol{Y}|\boldsymbol{X}]]$, we have

$$\boldsymbol{MSE(f(X_i)) \equiv \mathbb{E}\left[\mathbb{E}[(Y_i - f(X_i))^2|X_i]\right]}$$

Conditional expectation and MSE (cont'd)

- From the general variance formula: $\mathbb{V}[\boldsymbol{X}_i] = \mathbb{E}[\boldsymbol{X}_i^2] - (\mathbb{E}[\boldsymbol{X}_i])^2$ so $\mathbb{E}[\boldsymbol{X}_i^2] = \mathbb{V}[\boldsymbol{X}_i] + (\mathbb{E}[\boldsymbol{X}_i])^2$ and that is the case in the conditional case too
- The MSE can be rewritten

$$
\begin{aligned}
\boldsymbol{MSE}(\boldsymbol{f}(\boldsymbol{X}_i)) &= \mathbb{E}\left[\mathbb{E}[(\boldsymbol{Y}_i - \boldsymbol{f}(\boldsymbol{X}_i))^2 | \boldsymbol{X}_i]\right] \\
&= \mathbb{E}\left[\mathbb{V}[\boldsymbol{Y}_i - \boldsymbol{f}(\boldsymbol{X}_i) | \boldsymbol{X}] + (\mathbb{E}[\boldsymbol{Y}_i - \boldsymbol{f}(\boldsymbol{X}_i) | \boldsymbol{X}])^2\right] \\
&= \mathbb{V}[\boldsymbol{Y}_i | \boldsymbol{X}_i] + (\mathbb{E}[\boldsymbol{Y}_i - \boldsymbol{f}(\boldsymbol{X}_i) | \boldsymbol{X}_i])^2 \\
&= \mathbb{V}[\boldsymbol{Y}_i | \boldsymbol{X}_i] + (\mathbb{E}[\boldsymbol{Y}_i | \boldsymbol{X}_i] - \mathbb{E}[\boldsymbol{f}(\boldsymbol{X}_i) | \boldsymbol{X}_i])^2 \\
&= \mathbb{V}[\boldsymbol{Y}_i | \boldsymbol{X}_i] + (\mathbb{E}[\boldsymbol{Y}_i | \boldsymbol{X}_i] - \boldsymbol{f}(\boldsymbol{x}_i))^2
\end{aligned}
$$

Conditional expectation and MSE (cont'd)

$$\boldsymbol{MSE}(\boldsymbol{f}(\boldsymbol{X}_i)) = \mathbb{V}[\boldsymbol{Y}_i|\boldsymbol{X}_i] + (\mathbb{E}[\boldsymbol{Y}_i|\boldsymbol{X}_i] - \boldsymbol{f}(\boldsymbol{x}_i))^2$$

- The first term does not contain $\boldsymbol{f}(\boldsymbol{X}_i)$, so it is irrelevant
- If we want to minimize $\boldsymbol{MSE}(\boldsymbol{f}(\boldsymbol{X}_i))$, we need to minimize the second term (the first term would vanish in the first order condition)
- That second term is a square, so it is always positive. Minimizing it means making it equal to 0:

$$\boldsymbol{f}(\boldsymbol{x}_i) = \mathbb{E}[\boldsymbol{Y}_i|\boldsymbol{X}_i]$$

Conditional expectation and MSE (cont'd)

- The best predictor of $\boldsymbol{Y_i}$ by a function of $\boldsymbol{X_i}$ in the MSE sense is the conditional expectation of $\boldsymbol{Y_i}$ given $\boldsymbol{X_i}$!!
- In the linear model, we make the assumption $\mathbb{E}[\boldsymbol{Y_i}|\boldsymbol{X_i}] = \boldsymbol{\beta_0 + \beta_1 x_i}$
- Plug that in $\boldsymbol{MSE(f(X_i))}$ and you get

$$\boldsymbol{MSE}(\beta_0, \beta_1) = \mathbb{E}[(\boldsymbol{Y_i} - \beta_0 - \beta_1 \boldsymbol{X_i})^2]$$

- Rings a bell? It is the population version of

$$\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{y_i} - \beta_0 - \beta_1 \boldsymbol{x_i})^2$$

- The objective function the OLS estimates minimize! (see lecture on linear models)

What is a statistical model?

- A statistical model is a set of **assumptions** about the distribution of some data or about their relationships
- Example: Consider a pair of equally weighted six-sided dice. Underlying assumption: the probability of a die falling on any number is equal to **1/6**
- Example: Consider a pair of weighted six-sided dice. Underlying assumption; the probability of a die falling on any number differs depending on what face we are thinking of. The probabilities could be given to us, or not. And our job could be to figure out the weighting of the dice based on a sample of data
- These assumptions will have consequences on the properties of the statistical procedures used
- If we wrongly believe that the dice are equally weighted, our computations and conclusions will be wrong

Model types

3 types of statistical models are distinguished:

- **Parametric models**: models with a finite number of parameters (known or unknown). Parametric models specify the distribution of the error term typically (models estimated via **maximum likelihood estimators** are parametric)
- **Semiparametric models**: models with a finite-dimensional component and an infinite-dimensional component. Some assumptions are made about the distribution of some random variables, but the distribution is not fully specified (typically, linear models are semiparametric models as we generally only assume moments for the error term, not the full distribution)
- **Nonparametric models**: models with an infinite number of parameters. Minimal assumptions are made for such models. We remain agnostic about the shape of $\boldsymbol{f()}$, as we only care about making good predictions $\hat{\boldsymbol{y}}_{\boldsymbol{i}} = \hat{\boldsymbol{f}}(\boldsymbol{x}_{\boldsymbol{i,1}}, \boldsymbol{x}_{\boldsymbol{i,2}}, ..., \boldsymbol{x}_{\boldsymbol{i,K}})$

The role of assumptions in a model

- The assumptions made have a crucial role for the properties of the estimators
- The more restrictive the assumptions, the better the properties of the estimator. Is it always worth making restrictive assumptions though?
- Some assumptions are purely technical and needed for estimators to have the desired properties. Also hard to verify
- Other assumptions are based on common sense: given the data at hand, does it makes sense to assume this or that? There can be evidence in favor of an assumption or not
- As a consequence, one estimator might be preferred to another

What is machine learning?

- There are many definitions that differ on several aspects
- Wikipedia says: “Machine learning (ML) is the study of computer algorithms that improve automatically through experience.[1] It is seen as a part of artificial intelligence. Machine learning algorithms build a model based on sample data, known as”training data”, in order to make predictions or decisions without being explicitly programmed to do so”
- 1997 by Professor Tom M. Mitchel from Carnegie Mellon University, in his famous quote from (1997) “A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E”.
- Self-driving cars use AI systems, the automatic vision system that identifies an imminent accident is ML

Learning types

Broadly, 2 types of learning are distinguished:

- **Supervised learning**: There is a response variable $\boldsymbol{Y_i}$ with associated predictors $\boldsymbol{X_{i,j},\ j = 1, ..., K}$. The objective is to estimate a relationship between $\boldsymbol{Y_i}$ and the $\boldsymbol{X}$'s, or predict $\boldsymbol{Y_i}$ (Example: **regression methods**)
- **Unsupervised learning**: There is no response variable, only variables $\boldsymbol{X_{i,j},\ j = 1, ..., K}$. The objective is to understand the relationships between the variables or the observations (Example: **cluster analysis** seeks to group observations in categories according to the value of the variables)

Course outline

- We will go over many methods that are part of statistical learning
- Linear models are **semiparametric methods** that allow to estimate marginal effects as well as make predictions. But they impose constraints on the shape of $f()$
- **Nonparametric methods**: Generally used for prediction purposes, they assume very little about the shape of $f()$ and allow for arbitrary flexibility:
  - Nearest neighbors
  - kernel methods
  - Regression/classification trees
  - Neural networks
- **Unsupervised learning**
  - Principal component analysis
  - Clustering