# Statistical Learning

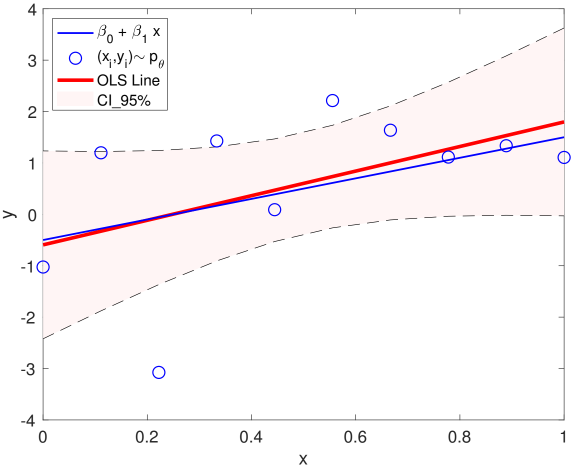
Master Spécialisé Intelligence Artificielle de Confiance (IAC)
@ Centrale Supélec en partenariat avec l'IRT SystemX
2024/2025.


Faïcel Chamroukhi

chamroukhi.com

# Supervised Learning

- Regression



# Learning Framework for Regression

- The data are represented by a random pair $(\boldsymbol{X}, Y) \in \mathcal{X} \times \mathcal{Y}$ where $\boldsymbol{X}$ is a vector of descriptors for some variable of interest $Y$
- The objective is **Prediction**, i.e. to seek for a prediction function $h : \mathcal{X} \rightarrow \mathcal{Y}$ for which $\widehat{y} = h(\boldsymbol{x})$ is a good approximation of the true output $y$
- In a **regression** problem : typically $\boldsymbol{X} \in \mathbb{R}^p$, $Y \in \mathcal{Y} = \mathbb{R}^d$

↪ We will mainly focus on parametric probabilistic models of the form

$$Y = h(X) + \epsilon, \epsilon \sim p_\theta$$

with the conditional distr. $P(Y|X,h)$ can be computed in terms of $P_\theta(Y - h(X))$.
$P(X,Y|h) = P(Y|X,h)P(X|h)$

- Data : a random sample $(\boldsymbol{X}_i, Y_i)_{i=1}^n$ with observed values $\mathcal{D}_n = (\boldsymbol{x}_i, y_i)_{i=1}^n$
- **Data-Scientist's role :** given the **data**, choose a **prediction function** $h$ that attempts to "minimize" the prediction error for of all possible data (**risk**) $R(h)$, under a chosen **loss** function $\ell$ measuring the error of predicting $Y$ by $h(X)$.
  - ↪ minimize the empirical (data-$\mathcal{D}_n$-driven) risk $R_n(h)$
  - ↪ Minimizing $R_n(h)$ may require an optimization **algorithm** $\mathcal{A}$
- Data-Scientist's "**Toolbox**" **:** {Data, hypothesis, loss, risk, algorithm}

# Learning Framework for Regression

- The data are represented by a random pair $(\boldsymbol{X}, Y) \in \mathcal{X} \times \mathcal{Y}$ where $\boldsymbol{X}$ is a vector of descriptors for some variable of interest $Y$
- The objective is **Prediction**, i.e. to seek for a prediction function $h : \mathcal{X} \rightarrow \mathcal{Y}$ for which $\widehat{y} = h(\boldsymbol{x})$ is a good approximation of the true output $y$
- In a **regression** problem : typically $\boldsymbol{X} \in \mathbb{R}^p$, $Y \in \mathcal{Y} = \mathbb{R}^d$

↪ We will mainly focus on parametric probabilistic models of the form

$$Y = h(X) + \epsilon, \epsilon \sim p_\theta$$

with the conditional distr. $P(Y|X, h)$ can be computed in terms of $P_\theta(Y - h(X))$. $P(X, Y|h) = P(Y|X, h)P(X|h)$

- Data : a random sample $(\boldsymbol{X}_i, Y_i)_{i=1}^n$ with observed values $\mathcal{D}_n = (\boldsymbol{x}_i, y_i)_{i=1}^n$
- **Data-Scientist's role :** given the **data**, choose a **prediction function** $h$ that attempts to "minimize" the prediction error for of all possible data (**risk**) $R(h)$, under a chosen **loss** function $\ell$ measuring the error of predicting $Y$ by $h(X)$.
  - ↪ minimize the empirical (data-$\mathcal{D}_n$-driven) risk $R_n(h)$
  - ↪ Minimizing $R_n(h)$ may require an optimization **algorithm** $\mathcal{A}$
- Data-Scientist's "**Toolbox**" **:** {Data, hypothesis, loss, risk, algorithm}

# Learning Framework for Regression

- The data are represented by a random pair $(\boldsymbol{X}, Y) \in \mathcal{X} \times \mathcal{Y}$ where $\boldsymbol{X}$ is a vector of descriptors for some variable of interest $Y$
- The objective is **Prediction**, i.e. to seek for a prediction function $h : \mathcal{X} \to \mathcal{Y}$ for which $\widehat{y} = h(\boldsymbol{x})$ is a good approximation of the true output $y$
- In a **regression** problem : typically $\boldsymbol{X} \in \mathbb{R}^p$, $Y \in \mathcal{Y} = \mathbb{R}^d$

↪ We will mainly focus on parametric probabilistic models of the form

$$Y = h(X) + \epsilon, \epsilon \sim p_\theta$$

with the conditional distr. $P(Y|X,h)$ can be computed in terms of $P_\theta(Y - h(X))$.
$P(X,Y|h) = P(Y|X,h)P(X|h)$

- Data : a random sample $(\boldsymbol{X}_i, Y_i)_{i=1}^n$ with observed values $\mathcal{D}_n = (\boldsymbol{x}_i, y_i)_{i=1}^n$
- **Data-Scientist's role :** given the **data**, choose a **prediction function** $h$ that attempts to "minimize" the prediction error for of all possible data (**risk**) $R(h)$, under a chosen **loss** function $\ell$ measuring the error of predicting $Y$ by $h(X)$.

  ↪ minimize the empirical (data-$\mathcal{D}_n$-driven) risk $R_n(h)$

  ↪ Minimizing $R_n(h)$ may require an optimization **algorithm** $\mathcal{A}$
- Data-Scientist's "**Toolbox**" : {Data, hypothesis, loss, risk, algorithm}

# Learning Framework for Regression

- The data are represented by a random pair $(\boldsymbol{X}, Y) \in \mathcal{X} \times \mathcal{Y}$ where $\boldsymbol{X}$ is a vector of descriptors for some variable of interest $Y$
- The objective is **Prediction**, i.e. to seek for a prediction function $h : \mathcal{X} \to \mathcal{Y}$ for which $\widehat{y} = h(\boldsymbol{x})$ is a good approximation of the true output $y$
- In a **regression** problem : typically $\boldsymbol{X} \in \mathbb{R}^p$, $Y \in \mathcal{Y} = \mathbb{R}^d$

↪ We will mainly focus on parametric probabilistic models of the form

$$Y = h(X) + \epsilon, \epsilon \sim p_\theta$$

with the conditional distr. $P(Y|X,h)$ can be computed in terms of $P_\theta(Y - h(X))$.
$P(X,Y|h) = P(Y|X,h)P(X|h)$

- Data : a random sample $(\boldsymbol{X}_i, Y_i)_{i=1}^n$ with observed values $\mathcal{D}_n = (\boldsymbol{x}_i, y_i)_{i=1}^n$
- **Data-Scientist's role :** given the **data**, choose a **prediction function** $h$ that attempts to "minimize" the prediction error for of all possible data (**risk**) $R(h)$, under a chosen **loss** function $\ell$ measuring the error of predicting $Y$ by $h(X)$.
  - ↪ minimize the empirical (data-$\mathcal{D}_n$-driven) risk $R_n(h)$
  - ↪ Minimizing $R_n(h)$ may require an optimization **algorithm** $\mathcal{A}$
- Data-Scientist's "**Toolbox**" **:** {Data, hypothesis, loss, risk, algorithm}

# Learning Framework for Regression

- The data are represented by a random pair $(\boldsymbol{X}, Y) \in \mathcal{X} \times \mathcal{Y}$ where $\boldsymbol{X}$ is a vector of descriptors for some variable of interest $Y$
- The objective is **Prediction**, i.e. to seek for a prediction function $h : \mathcal{X} \rightarrow \mathcal{Y}$ for which $\widehat{y} = h(\boldsymbol{x})$ is a good approximation of the true output $y$
- In a **regression** problem : typically $\boldsymbol{X} \in \mathbb{R}^p$, $Y \in \mathcal{Y} = \mathbb{R}^d$

↪ We will mainly focus on parametric probabilistic models of the form

$$Y = h(X) + \epsilon, \epsilon \sim p_\theta$$

with the conditional distr. $P(Y|X,h)$ can be computed in terms of $P_\theta(Y - h(X))$. $P(X,Y|h) = P(Y|X,h)P(X|h)$

- Data : a random sample $(\boldsymbol{X}_i, Y_i)_{i=1}^n$ with observed values $\mathcal{D}_n = (\boldsymbol{x}_i, y_i)_{i=1}^n$
- **Data-Scientist's role :** given the **data**, choose a **prediction function** $h$ that attempts to "minimize" the prediction error for of all possible data (**risk**) $R(h)$, under a chosen **loss** function $\ell$ measuring the error of predicting $Y$ by $h(X)$.
  - ↪ minimize the empirical (data-$\mathcal{D}_n$-driven) risk $R_n(h)$
  - ↪ Minimizing $R_n(h)$ may require an optimization **algorithm** $\mathcal{A}$
- Data-Scientist's "**Toolbox" :** {Data, hypothesis, loss, risk, algorithm}

# Learning Framework for Regression

- The data are represented by a random pair $(\boldsymbol{X}, Y) \in \mathcal{X} \times \mathcal{Y}$ where $\boldsymbol{X}$ is a vector of descriptors for some variable of interest $Y$
- The objective is **Prediction**, i.e. to seek for a prediction function $h : \mathcal{X} \to \mathcal{Y}$ for which $\widehat{y} = h(\boldsymbol{x})$ is a good approximation of the true output $y$
- In a **regression** problem : typically $\boldsymbol{X} \in \mathbb{R}^p$, $Y \in \mathcal{Y} = \mathbb{R}^d$

↪ We will mainly focus on parametric probabilistic models of the form

$$Y = h(X) + \epsilon, \epsilon \sim p_\theta$$

with the conditional distr. $P(Y|X,h)$ can be computed in terms of $P_\theta(Y - h(X))$. $P(X,Y|h) = P(Y|X,h)P(X|h)$

- Data : a random sample $(\boldsymbol{X}_i, Y_i)_{i=1}^n$ with observed values $\mathcal{D}_n = (\boldsymbol{x}_i, y_i)_{i=1}^n$
- **Data-Scientist's role :** given the **data**, choose a **prediction function** $h$ that attempts to "minimize" the prediction error for of all possible data (**risk**) $R(h)$, under a chosen **loss** function $\ell$ measuring the error of predicting $Y$ by $h(X)$.

  ↪ minimize the empirical (data-$\mathcal{D}_n$-driven) risk $R_n(h)$

  ↪ Minimizing $R_n(h)$ may require an optimization **algorithm** $\mathcal{A}$
- Data-Scientist's "**Toolbox**" : {Data, hypothesis, loss, risk, algorithm}

# Learning Framework for Regression

- The data are represented by a random pair $(\boldsymbol{X}, Y) \in \mathcal{X} \times \mathcal{Y}$ where $\boldsymbol{X}$ is a vector of descriptors for some variable of interest $Y$
- The objective is **Prediction**, i.e. to seek for a prediction function $h : \mathcal{X} \to \mathcal{Y}$ for which $\widehat{y} = h(\boldsymbol{x})$ is a good approximation of the true output $y$
- In a **regression** problem : typically $\boldsymbol{X} \in \mathbb{R}^p$, $Y \in \mathcal{Y} = \mathbb{R}^d$

↪ We will mainly focus on parametric probabilistic models of the form

$$Y = h(X) + \epsilon, \epsilon \sim p_\theta$$

with the conditional distr. $P(Y|X,h)$ can be computed in terms of $P_\theta(Y - h(X))$. $P(X,Y|h) = P(Y|X,h)P(X|h)$

- Data : a random sample $(\boldsymbol{X}_i, Y_i)_{i=1}^n$ with observed values $\mathcal{D}_n = (\boldsymbol{x}_i, y_i)_{i=1}^n$
- **Data-Scientist's role :** given the **data**, choose a **prediction function** $h$ that attempts to "minimize" the prediction error for of all possible data (**risk**) $R(h)$, under a chosen **loss** function $\ell$ measuring the error of predicting $Y$ by $h(X)$.

  ↪ minimize the empirical (data-$\mathcal{D}_n$-driven) risk $R_n(h)$

  ↪ Minimizing $R_n(h)$ may require an optimization **algorithm** $\mathcal{A}$
- Data-Scientist's "**Toolbox**" **:** {Data, hypothesis, loss, risk, algorithm}

# Learning Framework for Regression

- The data are represented by a random pair $(\boldsymbol{X}, Y) \in \mathcal{X} \times \mathcal{Y}$ where $\boldsymbol{X}$ is a vector of descriptors for some variable of interest $Y$
- The objective is **Prediction**, i.e. to seek for a prediction function $h : \mathcal{X} \rightarrow \mathcal{Y}$ for which $\widehat{y} = h(\boldsymbol{x})$ is a good approximation of the true output $y$
- In a **regression** problem : typically $\boldsymbol{X} \in \mathbb{R}^p$, $Y \in \mathcal{Y} = \mathbb{R}^d$

↪ We will mainly focus on parametric probabilistic models of the form

$$Y = h(X) + \epsilon, \epsilon \sim p_\theta$$

with the conditional distr. $P(Y|X,h)$ can be computed in terms of $P_\theta(Y - h(X))$. $P(X,Y|h) = P(Y|X,h)P(X|h)$

- Data : a random sample $(\boldsymbol{X}_i, Y_i)_{i=1}^n$ with observed values $\mathcal{D}_n = (\boldsymbol{x}_i, y_i)_{i=1}^n$
- **Data-Scientist's role :** given the **data**, choose a **prediction function** $h$ that attempts to "minimize" the prediction error for of all possible data (**risk**) $R(h)$, under a chosen **loss** function $\ell$ measuring the error of predicting $Y$ by $h(X)$.
  - ↪ minimize the empirical (data-$\mathcal{D}_n$-driven) risk $R_n(h)$
  - ↪ Minimizing $R_n(h)$ may require an optimization **algorithm** $\mathcal{A}$
- Data-Scientist's "**Toolbox**" **:** {Data, hypothesis, loss, risk, algorithm}

# Learning Framework for Regression

- The data are represented by a random pair $(\boldsymbol{X}, Y) \in \mathcal{X} \times \mathcal{Y}$ where $\boldsymbol{X}$ is a vector of descriptors for some variable of interest $Y$
- The objective is **Prediction**, i.e. to seek for a prediction function $h : \mathcal{X} \to \mathcal{Y}$ for which $\widehat{y} = h(\boldsymbol{x})$ is a good approximation of the true output $y$
- In a **regression** problem : typically $\boldsymbol{X} \in \mathbb{R}^p$, $Y \in \mathcal{Y} = \mathbb{R}^d$

↪ We will mainly focus on parametric probabilistic models of the form

$$Y = h(X) + \epsilon, \epsilon \sim p_\theta$$

with the conditional distr. $P(Y|X,h)$ can be computed in terms of $P_\theta(Y - h(X))$. $P(X,Y|h) = P(Y|X,h)P(X|h)$

- Data : a random sample $(\boldsymbol{X}_i, Y_i)_{i=1}^n$ with observed values $\mathcal{D}_n = (\boldsymbol{x}_i, y_i)_{i=1}^n$
- **Data-Scientist's role :** given the **data**, choose a **prediction function** $h$ that attempts to "minimize" the prediction error for of all possible data (**risk**) $R(h)$, under a chosen **loss** function $\ell$ measuring the error of predicting $Y$ by $h(X)$.

  ↪ minimize the empirical (data-$\mathcal{D}_n$-driven) risk $R_n(h)$

  ↪ Minimizing $R_n(h)$ may require an optimization **algorithm** $\mathcal{A}$
- Data-Scientist's "**Toolbox**" **:** {Data, hypothesis, loss, risk, algorithm}

# General regression form

## Objective

Regression models the relationship between two variables $X$ and $Y$

- Temperature ($Y$) of some water source, given the air temperate ($x$)
- Price ($Y$) of an apartment given its surface ($x_1$) and number of rooms ($x_2$)

Vocabulary :

- The $x$'s are called inputs/predictors/covariates/features/descriptors/ exogenous/Explanatory/independent variables
- The $y$'s are called output/outcome/response/endogenous/variable of interest/Explained/dependent variable

- **Simple** regression : $x \in \mathbb{R}$
- **Multiple** regression : $\boldsymbol{x} \in \mathbb{R}^p$
- **Univariate** regression : $y \in \mathbb{R}$
- **Multivariate** regression : $\boldsymbol{y} \in \mathbb{R}^d$

- **Functional** regression : when $x$ and/or $y$ are functional data

# General regression form

## Objective

Regression models the relationship between two variables $X$ and $Y$

- Temperature ($Y$) of some water source, given the air temperate ($x$)
- Price ($Y$) of an apartment given its surface ($x_1$) and number of rooms ($x_2$)

Vocabulary :

- The $x$'s are called inputs/predictors/covariates/features/descriptors/ exogenous/Explanatory/independent variables
- The $y$'s are called output/outcome/response/endogenous/variable of interest/Explained/dependent variable

- **Simple** regression : $x \in \mathbb{R}$
- **Multiple** regression : $\boldsymbol{x} \in \mathbb{R}^p$
- **Univariate** regression : $y \in \mathbb{R}$
- **Multivariate** regression : $\boldsymbol{y} \in \mathbb{R}^d$

- **Functional** regression : when $x$ and/or $y$ are functional data

# General regression form

- Consider the random pair $(X, Y)$ where $X \in \mathcal{X} \subset \mathbb{R}^p$ is the predictor and $Y \in \mathcal{Y} \subset \mathbb{R}^d$ is the response.
- A regression model can be phrased as

$$Y = f(X) + \varepsilon$$

where $f : \mathcal{X} \to \mathcal{Y}$ is the regression function (parametric or not, linear or not..)
$\varepsilon$ is a random variable : noise/residual/error

- Standard hypotheses : The error terms $\varepsilon$ are

(i) centered : $\mathbb{E}(\varepsilon_i) = 0$ (for all $i$)

(ii) uncorrelated with the covariates : $\mathbb{E}(\varepsilon_i X_j) = 0$ (for all $i, j$)

(iii) homoskedastic, with limited variance : $\mathbb{E}(\varepsilon_i^2 | X_i) = \sigma^2 < \infty$ (for all $i$)

(iv) uncorrelated with each other : $\mathbb{E}[\varepsilon_i \varepsilon_j | X] = 0$ (for all $i \neq j$)

(v) identically distributed : $\varepsilon_i \underset{\text{id}}{\sim} p$ (for all $i$)

The caovariates $X$ can be deterministic (fixed design) or random

The errors $\varepsilon_i$ can be supposed normal for statistical inference

# General regression form

- Consider the random pair $(X, Y)$ where $X \in \mathcal{X} \subset \mathbb{R}^p$ is the predictor and $Y \in \mathcal{Y} \subset \mathbb{R}^d$ is the response.
- A regression model can be phrased as

$$Y = f(X) + \varepsilon$$

where $f : \mathcal{X} \to \mathcal{Y}$ is the regression function (parametric or not, linear or not..)
$\varepsilon$ is a random variable : noise/residual/error

- Standard hypotheses : The error terms $\varepsilon$ are

(i) centered : $\mathbb{E}(\varepsilon_i) = 0$ (for all $i$)

(ii) uncorrelated with the covariates : $\mathbb{E}(\varepsilon_i X_j) = 0$ (for all $i, j$)

(iii) homoskedastic, with limited variance : $\mathbb{E}(\varepsilon_i^2 | X_i) = \sigma^2 < \infty$ (for all $i$)

(iv) uncorrelated with each other : $\mathbb{E}[\varepsilon_i \varepsilon_j | X] = 0$ (for all $i \neq j$)

(v) identically distributed : $\varepsilon_i \underset{\text{id}}{\sim} p$ (for all $i$)

The caovariates $X$ can be deterministic (fixed design) or random

The errors $\varepsilon_i$ can be supposed normal for statistical inference

# Prediction/decision function

## Def. Regression function

$$h\colon \mathcal{X} \subset \mathbb{R}^p \to \mathcal{Y} \subset \mathbb{R}^d$$
$$x \mapsto h(x)$$

is a regression function, parametric or not, linear or not, ...

Example : Linear prediction functions : Consider $\mathcal{H} = \{h(x) = \langle x, \theta \rangle = \theta^T x\}$, the set of linear functions in $X$ of the form $h(x; \boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}[Y|X] = \beta_0 + \beta^T X$ and $\boldsymbol{\theta} = (\beta_0, \beta^T)^T$.

$$h\colon \mathbb{R}^p \to \mathbb{R}$$
$$x \mapsto \langle x, \theta \rangle = \theta^T x$$

The **predicted** values of $Y_i$'s for new covariates $X_i = x_i$s correspond to

$$\widehat{y}_i = h(x_i)$$

Example : Linear prediction functions (cont.) : $\widehat{y}_i = \langle x_i, \theta \rangle = \theta^T x_i$

**Q :** How good we are in prediction on a particular pair $(x, y)$ ?

# Prediction/decision function

## Def. Regression function

$$h\colon \mathcal{X} \subset \mathbb{R}^p \to \mathcal{Y} \subset \mathbb{R}^d$$
$$x \mapsto h(x)$$

is a regression function, parametric or not, linear or not, ...

Example : Linear prediction functions : Consider $\mathcal{H} = \{h(x) = \langle x, \theta\rangle = \theta^T x\}$, the set of linear functions in $X$ of the form $h(x; \boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}[Y|X] = \beta_0 + \beta^T X$ and $\boldsymbol{\theta} = (\beta_0, \beta^T)^T$.

$$h\colon \mathbb{R}^p \to \mathbb{R}$$
$$x \mapsto \langle x, \theta\rangle = \theta^T x$$

The **predicted** values of $Y_i$'s for new covariates $X_i = x_i$s correspond to

$$\widehat{y}_i = h(x_i)$$

Example : Linear prediction functions (cont.) : $\widehat{y}_i = \langle x_i, \theta\rangle = \theta^T x_i$

**Q** : How good we are in prediction on a particular pair $(x, y)$ ?

# Prediction/decision function

## Def. Regression function

$$h \colon \mathcal{X} \subset \mathbb{R}^p \to \mathcal{Y} \subset \mathbb{R}^d$$
$$x \mapsto h(x)$$

is a regression function, parametric or not, linear or not, ...

Example : Linear prediction functions : Consider $\mathcal{H} = \{h(x) = \langle x, \theta \rangle = \theta^T x\}$, the set of linear functions in $X$ of the form $h(x; \boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}[Y|X] = \beta_0 + \beta^T X$ and $\boldsymbol{\theta} = (\beta_0, \beta^T)^T$.

$$h \colon \mathbb{R}^p \to \mathbb{R}$$
$$x \mapsto \langle x, \theta \rangle = \theta^T x$$

The **predicted** values of $Y_i$'s for new covariates $X_i = x_i$s correspond to

$$\widehat{y}_i = h(x_i)$$

Example : Linear prediction functions (cont.) : $\widehat{y}_i = \langle x_i, \theta \rangle = \theta^T x_i$

**Q** : How good we are in prediction on a particular pair $(x, y)$ ?

# Prediction/decision function

## Def. Regression function

$$h\colon \mathcal{X} \subset \mathbb{R}^p \to \mathcal{Y} \subset \mathbb{R}^d$$
$$x \mapsto h(x)$$

is a regression function, parametric or not, linear or not, ...

Example : Linear prediction functions : Consider $\mathcal{H} = \{h(x) = \langle x, \theta\rangle = \theta^T x\}$, the set of linear functions in $X$ of the form $h(x;\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}[Y|X] = \beta_0 + \beta^T X$ and $\boldsymbol{\theta} = (\beta_0, \beta^T)^T$.

$$h\colon \mathbb{R}^p \to \mathbb{R}$$
$$x \mapsto \langle x, \theta\rangle = \theta^T x$$

The **predicted** values of $Y_i$'s for new covariates $X_i = x_i$s correspond to

$$\widehat{y}_i = h(x_i)$$

Example : Linear prediction functions (cont.) : $\widehat{y}_i = \langle x_i, \theta\rangle = \theta^T x_i$

**Q :** How good we are in prediction on a particular pair $(x, y)$ ?

# Loss

## Def. Loss function

$$\ell \colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$$
$$(y, h(x)) \mapsto \ell(y, h(x))$$

It measures how good we are on a particular $(x, y)$ pair.
(We assume that the distribution of the test data is the same as for the training data).

**Examples of loss functions in regression**

- Square ($\ell_2$)-loss : $\ell_2(y, h(x)) = (y - h(x))^2$
- Absolute ($\ell_1$)-loss : $\ell_1(y, h(x)) = |y - h(x)|$
- Huber loss : $\ell_\delta(y, h(x)) = \begin{cases} \frac{1}{2}(y - h(x))^2 \text{ if } |y - f(x)| \leq \delta, \\ \delta\left(|y - h(x)| - \frac{1}{2}\delta\right), \text{otherwise.} \end{cases}$

# Loss

## Def. Loss function

$$\ell \colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$$
$$(y, h(x)) \mapsto \ell(y, h(x))$$

It measures how good we are on a particular $(x, y)$ pair.
(We assume that the distribution of the test data is the same as for the training data).

**Examples of loss functions in regression**

- Square ($\ell_2$)-loss : $\ell_2(y, h(x)) = (y - h(x))^2$
- Absolute ($\ell_1$)-loss : $\ell_1(y, h(x)) = |y - h(x)|$
- Huber loss : $\ell_\delta(y, h(x)) = \begin{cases} \frac{1}{2}(y - h(x))^2 \text{ if } |y - f(x)| \leq \delta, \\ \delta\left(|y - h(x)| - \frac{1}{2}\delta\right), \text{otherwise.} \end{cases}$

## Def. Loss function

$$\ell\colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$$
$$(y, h(x)) \mapsto \ell(y, h(x))$$

It measures how good we are on a particular $(x, y)$ pair.
(We assume that the distribution of the test data is the same as for the training data).

**Examples of loss functions in regression**

- Square ($\ell_2$)-loss : $\ell_2(y, h(x)) = (y - h(x))^2$
- Absolute ($\ell_1$)-loss : $\ell_1(y, h(x)) = |y - h(x)|$
- Huber loss : $\ell_\delta(y, h(x)) = \begin{cases} \frac{1}{2}(y - h(x))^2 \text{ if } |y - f(x)| \leq \delta, \\ \delta\left(|y - h(x)| - \frac{1}{2}\delta\right), \text{otherwise.} \end{cases}$

# Examples of loss functions

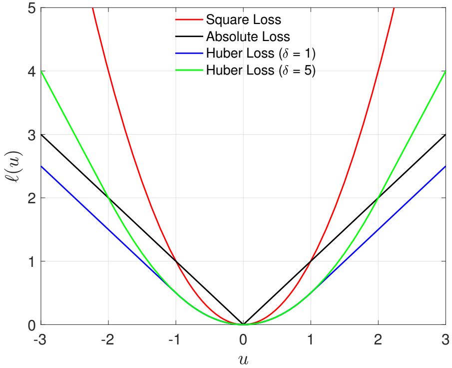



FIGURE – Some loss functions in regression : curves of $\ell(u)$ for $u = y - h(x)$ ; $y \in \mathbb{R}$.

- Square loss : $\ell_2(u) = (u)^2$
- Absolute loss : $\ell_1(u) = |u|$
- Huber loss : $\ell_\delta(u) = \begin{cases} \frac{1}{2}(u)^2 \text{ if } |u| \leq \delta, \\ \delta\,(|u| - \frac{1}{2}\delta), \text{otherwise.} \end{cases}$

# Risk

- **Risk** : the *Expected loss* :

$$R(h) = \mathbb{E}_P[\ell(Y, h(X))] = \int_{\mathcal{X}\times\mathcal{Y}} \ell(y, h(x))dP(x, y)$$

$\hookrightarrow$ the error of approximating $Y$ by model/hypothesis $h(X)$ as measured by a chosen loss function $\ell(Y, h(X))$ given the pair $(X, Y)$ with (unknown) joint distribution $P$,

$\hookrightarrow$ prediction error : measures the generalization performance of the function $h$.

- **Squared Risk :** Under the squared loss $\ell(y, h(x)) = (y - h(x))^2$ :

$$R(h) = \mathbb{E}_P[(Y - h(X))^2] = \int_{\mathcal{X}\times\mathcal{Y}} (y - h(x))^2 dP(x, y).$$

$\hookrightarrow$ This is the most used risk in regression

**Q** : what is the best function $h$ ? or equivalently, when the risk $R(h)$ is optimal ?

# Risk

- **Risk** : the *Expected loss* :

$$R(h) = \mathbb{E}_P[\ell(Y, h(X))] = \int_{\mathcal{X}\times\mathcal{Y}} \ell(y, h(x))dP(x, y)$$

$\hookrightarrow$ the error of approximating $Y$ by model/hypothesis $h(X)$ as measured by a chosen loss function $\ell(Y, h(X))$ given the pair $(X, Y)$ with (unknown) joint distribution $P$,

$\hookrightarrow$ prediction error : measures the generalization performance of the function $h$.

- **Squared Risk :** Under the squared loss $\ell(y, h(x)) = (y - h(x))^2$ :

$$R(h) = \mathbb{E}_P[(Y - h(X))^2] = \int_{\mathcal{X}\times\mathcal{Y}} (y - h(x))^2 dP(x, y).$$

$\hookrightarrow$ This is the most used risk in regression

**Q** : what is the best function $h$ ? or equivalently, when the risk $R(h)$ is optimal ?

# Risk

- **Risk** : the *Expected loss* :

$$R(h) = \mathbb{E}_P[\ell(Y, h(X))] = \int_{\mathcal{X}\times\mathcal{Y}} \ell(y, h(x))dP(x, y)$$

$\hookrightarrow$ the error of approximating $Y$ by model/hypothesis $h(X)$ as measured by a chosen loss function $\ell(Y, h(X))$ given the pair $(X, Y)$ with (unknown) joint distribution $P$,

$\hookrightarrow$ prediction error : measures the generalization performance of the function $h$.

- **Squared Risk :** Under the squared loss $\ell(y, h(x)) = (y - h(x))^2$ :

$$R(h) = \mathbb{E}_P[(Y - h(X))^2] = \int_{\mathcal{X}\times\mathcal{Y}} (y - h(x))^2 dP(x, y).$$

$\hookrightarrow$ This is the most used risk in regression

**Q** : what is the best function $h$ ? or equivalently, when the risk $R(h)$ is optimal ?

# Optimal prediction function

**Def. regression function**

$$h\colon \mathcal{X} \to \mathcal{Y}$$
$$x \mapsto h(x)$$

**Theorem (Bayes regression function under the square loss)**

*Consider the $\ell_2$-loss, $\ell(Y, h(X)) = (h(X)) - Y)^2$, then, the Bayes rule minimizing the corresponding regression risk (the best prediction function) of a regression function $h(x)$*

$$R(h) = \mathbb{E}_X((Y - h(X))^2 | X = x)$$

*is given by the conditional expectation*

$$h^*(x) = \mathbb{E}(Y|X = x).$$

# Optimal prediction function

**Def. regression function**

$$\begin{aligned} h \colon \mathcal{X} &\to \mathcal{Y} \\ x &\mapsto h(x) \end{aligned}$$

**Theorem (Bayes regression function under the square loss)**

*Consider the $\ell_2$-loss, $\ell(Y, h(X)) = (h(X)) - Y)^2$, then, the Bayes rule minimizing the corresponding regression risk (the best prediction function) of a regression function $h(x)$*

$$R(h) = \mathbb{E}_X((Y - h(X))^2 | X = x)$$

*is given by the conditional expectation*

$$h^*(x) = \mathbb{E}(Y|X = x).$$

# Optimal prediction function

## Proof : Bayes predictor under the squared loss.

Under the square loss, $\ell(y, h(x)) = (h(x) - y)^2$, the best prediction function is

$$h^*(x) = \mathbb{E}[Y|X = x]$$

For $X = x$, consider the conditional risk :
$\mathbb{E}[\ell(Y, h(X))|X = x] = \mathbb{E}_{Y|X=x}[(h(X) - Y)^2|X = x] = \int_{\mathcal{Y}}(h(x) - y)^2 p(y|x)dy$,
optimizing the Risk by differentiating w.r.t $h(x)$ and setting the derivative to $0$ :

$$\begin{aligned}\frac{\partial R(h(x))}{\partial h(x)} = 2\int(h(x) - y)p(y|x)dy &= 2\big[h(x)\int p(y|x)dy - \int yp(y|x)dy\big] \\ &= 2\big(h(x) - \mathbb{E}[Y|X = x]\big)\end{aligned}$$

which is zero at $h(x)^* = \mathbb{E}[Y|X = x]$. Then

$$h^*(x) = \arg\min_{h(x)\in\mathcal{Y}} \mathbb{E}[l(Y, h(X))|X = x] = \mathbb{E}[Y|X = x]$$

□

**Goal** : estimate $h^*$, knowing only the data sample $D_n = (X_i, Y_i)_{i=1}^n$ and loss $\ell$.

- Then *Expected loss* $R(h)$ depends on the joint distribution $P$ of the pair $(X, Y)$. In real situations $P$ is in unknown, as we only have a sample $D_n = (X_i, Y_i)_{1 \leq i \leq n}$,

↪ We attempt to minimize the **Empirical Risk** $R_n(h) = \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, h(X_i))$ to estimate $h^*$ (within a family $\mathcal{H}$) :

$$\widehat{h}_n \in \arg\min_{h \in \mathcal{H}} R_n(h).$$

**MSE and Ordinary Least Squares (OLS)** :

- Squared error : Under the squared loss (the standard in regression) : $\ell_2(y, h(x)) = (y - h(x))^2$, the empirical risk $R_n(h)$ is the empirical square loss [1]

$$R_n(h) = \frac{1}{n}\sum_{i=1}^{n} \|Y_i - h(X_i)\|_2^2$$

- ERM : $\widehat{h}_n = \arg\min_{h \in \mathcal{H}} R_n(h)$ is the **Ordinary Least Squares Estimator** of $h$
- Liner regression : Consider $\mathcal{H} = \{h_\theta(x) = \alpha + \beta^T x\}$, the set of linear functions in $x$

1. also called the Mean Squared Error (MSE), or the mean Residual Squared Sum (RSS) when the ML problem is phrased as an error model $Y = h(X) + \epsilon$, $\epsilon \sim p$

- Then *Expected loss* $R(h)$ depends on the joint distribution $P$ of the pair $(X, Y)$. In real situations $P$ is in unknown, as we only have a sample $D_n = (X_i, Y_i)_{1 \leq i \leq n}$,

$\hookrightarrow$ We attempt to minimize the **Empirical Risk** $R_n(h) = \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, h(X_i))$ to estimate $h^*$ (within a family $\mathcal{H}$) :

$$\widehat{h}_n \in \arg\min_{h \in \mathcal{H}} R_n(h).$$

**MSE and Ordinary Least Squares (OLS)** :

- Squared error : Under the squared loss (the standard in regression) : $\ell_2(y, h(x)) = (y - h(x))^2$, the empirical risk $R_n(h)$ is the empirical square loss [1]

$$R_n(h) = \frac{1}{n}\sum_{i=1}^{n} \|Y_i - h(X_i)\|_2^2$$

- ERM : $\widehat{h}_n = \arg\min_{h \in \mathcal{H}} R_n(h)$ is the **Ordinary Least Squares Estimator** of $h$
- Liner regression : Consider $\mathcal{H} = \{h_\theta(x) = \alpha + \beta^T x\}$, the set of linear functions in $x$

1. also called the Mean Squared Error (MSE), or the mean Residual Squared Sum (RSS) when the ML problem is phrased as an error model $Y = h(X) + \epsilon$, $\epsilon \sim p$

- Then *Expected loss* $R(h)$ depends on the joint distribution $P$ of the pair $(X, Y)$. In real situations $P$ is in unknown, as we only have a sample $D_n = (X_i, Y_i)_{1 \leq i \leq n}$,

↪ We attempt to minimize the **Empirical Risk** $R_n(h) = \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, h(X_i))$ to estimate $h^*$ (within a family $\mathcal{H}$) :

$$\widehat{h}_n \in \arg\min_{h \in \mathcal{H}} R_n(h).$$

**MSE and Ordinary Least Squares (OLS)** :

- Squared error : Under the squared loss (the standard in regression) : $\ell_2(y, h(x)) = (y - h(x))^2$, the empirical risk $R_n(h)$ is the empirical square loss [1]

$$R_n(h) = \frac{1}{n}\sum_{i=1}^{n} \|Y_i - h(X_i)\|_2^2$$

- ERM : $\widehat{h}_n = \arg\min_{h \in \mathcal{H}} R_n(h)$ is the **Ordinary Least Squares Estimator** of $h$
- Liner regression : Consider $\mathcal{H} = \{h_\theta(x) = \alpha + \beta^T x\}$, the set of linear functions in $x$

1. also called the Mean Squared Error (MSE), or the mean Residual Squared Sum (RSS) when the ML problem is phrased as an error model $Y = h(X) + \epsilon$, $\epsilon \sim p$

# Simple Linear Regression

# Simple Linear Regression

- We model the pair $(X, Y)$ where the predictor $X \in \mathbb{R}$ and the response $Y \in \mathbb{R}$
- An observed $Y$ given a single scalar predictor $x$, is said to satisfy the simple linear regression model when

$$h(x) = \mathbb{E}[Y|X = x] = \int_{\mathcal{Y}} yp(y|x)dy = \beta_0 + \beta_1 x$$

i.e., equivalently

$$\boxed{\begin{aligned} &Y = \beta_0 + \beta_1 X + \epsilon, \\ &\quad \mathbb{E}[\epsilon|X] = 0 \text{ and } \mathbb{V}[\epsilon|X] = \sigma^2 \end{aligned}}$$

$\beta_0$ (the intercept) and $\beta_1$ (the slope) : unknown regression **coefficients**
$\sigma^2$ an unknown noise variance

**Bayes Risk**
$R^* = R(\theta^*) = \mathbb{E}[(Y - h^*(X))^2] = \mathbb{E}_X \mathbb{E}_{Y|X}[(Y - h^*(X))^2|X] = \mathbb{E}_X \mathbb{E}[\epsilon^2|X] = \sigma^2$

# Simple Linear Regression

- We model the pair $(X, Y)$ where the predictor $X \in \mathbb{R}$ and the response $Y \in \mathbb{R}$
- An observed $Y$ given a single scalar predictor $x$, is said to satisfy the simple linear regression model when

$$h(x) = \mathbb{E}[Y|X = x] = \int_{\mathcal{Y}} yp(y|x)dy = \beta_0 + \beta_1 x$$

i.e., equivalently

$$\boxed{\begin{aligned} &Y = \beta_0 + \beta_1 X + \epsilon, \\ &\mathbb{E}[\epsilon|X] = 0 \text{ and } \mathbb{V}[\epsilon|X] = \sigma^2 \end{aligned}}$$

$\beta_0$ (the intercept) and $\beta_1$ (the slope) : unknown regression **coefficients**

$\sigma^2$ an unknown noise variance

**Bayes Risk**

$R^* = R(\theta^*) = \mathbb{E}[(Y - h^*(X))^2] = \mathbb{E}_X \mathbb{E}_{Y|X}[(Y - h^*(X))^2|X] = \mathbb{E}_X \mathbb{E}[\epsilon^2|X] = \sigma^2$

# Simple Linear Regression

- We model the pair $(X, Y)$ where the predictor $X \in \mathbb{R}$ and the response $Y \in \mathbb{R}$
- An observed $Y$ given a single scalar predictor $x$, is said to satisfy the simple linear regression model when

$$h(x) = \mathbb{E}[Y|X = x] = \int_{\mathcal{Y}} yp(y|x)dy = \beta_0 + \beta_1 x$$

i.e., equivalently

$$\boxed{\begin{aligned} &Y = \beta_0 + \beta_1 X + \epsilon, \\ &\quad \mathbb{E}[\epsilon|X] = 0 \text{ and } \mathbb{V}[\epsilon|X] = \sigma^2 \end{aligned}}$$

$\beta_0$ (the intercept) and $\beta_1$ (the slope) : unknown regression **coefficients**

$\sigma^2$ an unknown noise variance

**Bayes Risk**

$R^* = R(\theta^*) = \mathbb{E}[(Y - h^*(X))^2] = \mathbb{E}_X \mathbb{E}_{Y|X}[(Y - h^*(X))^2|X] = \mathbb{E}_X \mathbb{E}[\epsilon^2|X] = \sigma^2$

# Simple Linear Regression

- We model the pair $(X, Y)$ where the predictor $X \in \mathbb{R}$ and the response $Y \in \mathbb{R}$
- An observed $Y$ given a single scalar predictor $x$, is said to satisfy the simple linear regression model when

$$h(x) = \mathbb{E}[Y|X = x] = \int_{\mathcal{Y}} yp(y|x)dy = \beta_0 + \beta_1 x$$

i.e., equivalently

$$\boxed{\begin{aligned} Y &= \beta_0 + \beta_1 X + \epsilon, \\ \mathbb{E}[\epsilon|X] &= 0 \text{ and } \mathbb{V}[\epsilon|X] = \sigma^2 \end{aligned}}$$

$\beta_0$ (the intercept) and $\beta_1$ (the slope) : unknown regression **coefficients**

$\sigma^2$ an unknown noise variance

**Bayes Risk**

$R^* = R(\theta^*) = \mathbb{E}[(Y - h^*(X))^2] = \mathbb{E}_X \mathbb{E}_{Y|X}[(Y - h^*(X))^2|X] = \mathbb{E}_X \mathbb{E}[\epsilon^2|X] = \sigma^2$

# Ordinary Least Squares (OLS) for SLR

Let $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ be the empirical (sample) means.

## Theorem (Ordinary Least Squares (OLS) for SLR)

*If* $(\widehat{\beta}_0^{OLS}, \widehat{\beta}_1^{OLS})$ *are the OLS Estimators of* $(\beta_0, \beta_1)$*, then*

$$\widehat{\beta}_0^{OLS} = \overline{Y} - \widehat{\beta}_1 \overline{X},$$

$$\widehat{\beta}_1^{OLS} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}.$$

- We have $\widehat{\beta}_1^{\text{OLS}} = S_{XY} \Big/ S_X^2$ where $S_{XY} = n^{-1}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})$ is the sample covariance and $S_X^2 = n^{-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ is the sample variance.
- An estimator $\widehat{\sigma}^2$ of the variance $\sigma^2$ can be taken as the empirical variance

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{h}(X_i))^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - (\widehat{\beta}_0^{\text{OLS}} + \widehat{\beta}_1^{\text{OLS}} X_i))^2$$

We'll see its construction later in connection with Gaussian regression, as the Maximum-Likelihood Estimator (MLE)

# Ordinary Least Squares (OLS) for SLR

Let $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ be the empirical (sample) means.

## Theorem (Ordinary Least Squares (OLS) for SLR)

*If* $(\widehat{\beta}_0^{OLS}, \widehat{\beta}_1^{OLS})$ *are the OLS Estimators of* $(\beta_0, \beta_1)$, *then*

$$\widehat{\beta}_0^{OLS} = \overline{Y} - \widehat{\beta}_1 \overline{X},$$

$$\widehat{\beta}_1^{OLS} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}.$$

- We have $\widehat{\beta}_1^{\text{OLS}} = S_{XY}\Big/S_X^2$ where $S_{XY} = n^{-1}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})$ is the sample covariance and $S_X^2 = n^{-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ is the sample variance.
- An estimator $\widehat{\sigma}^2$ of the variance $\sigma^2$ can be taken as the empirical variance

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{h}(X_i))^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - (\widehat{\beta}_0^{\text{OLS}} + \widehat{\beta}_1^{\text{OLS}} X_i))^2$$

We'll see its construction later in connection with Gaussian regression, as the Maximum-Likelihood Estimator (MLE)

# Ordinary Least Squares (OLS) for SLR

## Proof of the OLS for SLR.

The OLS estimates $(\widehat{\beta}_0, \widehat{\beta}_1)$ are the ERM, i.e minimizing the residual squared sum (RSS) $Q(\beta_0, \beta_1) = \sum_{i=1}^{n} \big(y_i - (\beta_0 + \beta_1 x_i)\big)^2$, i.e. $(\widehat{\beta}_0, \widehat{\beta}_1) = \arg\min_{(\beta_0,\beta_1)\in\mathbb{R}^2} Q(\beta_0, \beta_1)$.
F.O.C : Deriving $Q$ w.r.t $(\beta_0, \beta_1)$ we get

$$\begin{aligned}
\frac{\partial Q}{\partial \beta_0} &= \frac{\partial \sum_{i=1}^{n} \big(y_i - (\beta_0 + \beta_1 x_i)\big)^2}{\partial \beta_0} = -2\sum_{i=1}^{n} \big(y_i - (\beta_0 + \beta_1 x_i)\big) \\
\frac{\partial Q}{\partial \beta_1} &= \frac{\partial \sum_{i=1}^{n} \big(y_i - (\beta_0 + \beta_1 x_i)\big)^2}{\partial \beta_1} = -2\sum_{i=1}^{n} x_i\big(y_i - (\beta_0 + \beta_1 x_i)\big) \cdot
\end{aligned}$$

and setting to zero we obtain

$$\begin{aligned}
\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \widehat{\beta}_0 - \widehat{\beta}_1 \sum_{i=1}^{n} x_i &= 0 \\
\sum_{i=1}^{n} x_i y_i - \widehat{\beta}_0 \sum_{i=1}^{n} x_i - \widehat{\beta}_1 \sum_{i=1}^{n} x_i^2 &= 0 \cdot
\end{aligned}$$

which gives the *normal equations* :

$$\begin{aligned}
n\widehat{\beta}_0 &= \sum_{i=1}^{n} y_i - \widehat{\beta}_1 \sum_{i=1}^{n} x_i \\
\widehat{\beta}_0 \sum_{i=1}^{n} x_i + \widehat{\beta}_1 \sum_{i=1}^{n} x_i^2 &= \sum_{i=1}^{n} x_i y_i
\end{aligned}$$

□

# Ordinary Least Squares (OLS) for SLR

## Proof of the OLS for SLR (cont.)

The first normal equation one gives

$$\begin{aligned} n\widehat{\beta}_0 &= \textstyle\sum_{i=1}^n y_i - \widehat{\beta}_1 \sum_{i=1}^n x_i \\ \widehat{\beta}_0 &= \frac{1}{n}\sum_{i=1}^n y_i - \widehat{\beta}_1 \left(\frac{1}{n}\sum_{i=1}^n x_i\right) = \overline{y} - \widehat{\beta}_1 \overline{x} \cdot \end{aligned}$$

The second gives

$$\begin{aligned} \widehat{\beta}_0 \textstyle\sum_{i=1}^n x_i + \widehat{\beta}_1 \sum_{i=1}^n x_i^2 &= \textstyle\sum_{i=1}^n x_i y_i \\ \textstyle\sum_{i=1}^n x_i \overline{y} - \widehat{\beta}_1 \sum_{i=1}^n x_i \overline{x} + \widehat{\beta}_1 \sum_{i=1}^n x_i^2 &= \textstyle\sum_{i=1}^n x_i y_i \,, \end{aligned}$$

we finally obtain : $\widehat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \overline{y}}{\sum_{i=1}^n x_i^2 - \sum_{i=1}^n x_i \overline{x}} = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^n (x_i - \overline{x})^2}$ .

S.O.C :

$$\det \begin{pmatrix} \frac{\partial^2 Q}{\partial^2 \beta_0} & \frac{\partial^2 Q}{\partial \beta_0 \partial \beta_1} \\ \frac{\partial^2 Q}{\partial \beta_0 \partial \beta_1} & \frac{\partial^2 Q}{\partial^2 \beta_1} \end{pmatrix} = \det \begin{pmatrix} 2n & 2\sum_i x_i \\ 2\sum_i x_i & 2\sum_i x_i^2 \end{pmatrix} = 4n \sum_i (x_i - \overline{x})^2 > 0 \cdot$$

This determinant is zero if all the $x_i$'s take the same value. At least two distinct $x_i$'s are necessary to estimate the coefficients $(\beta_0, \beta_1)$ (to fit the line).

# Statistical Properties of the OLS Estimator

Let $S_X^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2$ be the empirical variance of the $n$ covariates $X^n = (X_1, \ldots, X_n)$.

## Theorem : Linearity, Unbiasedness and Variance of the OLS

The OLS Estimators $(\widehat{\beta}_0, \widehat{\beta}_1)$ of $(\beta_0, \beta_1)$ are **linear** in $Y_i$ and **unbiased**, with

$$\mathbb{V}(\widehat{\beta}_0|X^n) = \sigma^2 \left(\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\right) = \frac{\sigma^2}{n}\left(1 + \frac{\overline{X}^2}{S_X^2}\right)$$

$$\mathbb{V}(\widehat{\beta}_1|X^n) = \frac{\sigma^2}{\sum_{i=1}^n (X_i - \overline{X})^2} = \frac{\sigma^2}{nS_X^2}$$

$$\mathrm{Cov}(\widehat{\beta}_0, \widehat{\beta}_1|X^n) = -\frac{\overline{X}\sigma^2}{\sum_{i=1}^n (X_i - \overline{X})^2} = -\frac{\overline{X}\sigma^2}{nS_X^2}$$

- Estimates of these statistics are obtained by replacing the variance $\sigma^2$ by its estimator $\widehat{\sigma}^2$ (eg., the corrected MLE). The estimated standard errors $\widehat{\text{se}}$ of $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are given by

$$\widehat{\text{se}}(\widehat{\beta}_0|X^n) = \frac{\widehat{\sigma}}{\sqrt{n}}\sqrt{\left(1 + \frac{\overline{X}^2}{S_X^2}\right)} = \frac{\widehat{\sigma}}{\sqrt{n}S_X}\sqrt{\frac{\sum_{i=1}^n X_i^2}{n}}$$

$$\widehat{\text{se}}(\widehat{\beta}_1|X^n) = \frac{\widehat{\sigma}}{\sqrt{n}S_X}; \quad \widehat{\mathrm{Cov}}(\widehat{\beta}_0, \widehat{\beta}_1|X^n) = -\frac{\overline{X}\widehat{\sigma}^2}{nS_X^2}$$

# Statistical Properties of the OLS Estimator

Let $S_X^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2$ be the empirical variance of the $n$ covariates $X^n = (X_1, \ldots, X_n)$.

## Theorem : Linearity, Unbiasedness and Variance of the OLS

The OLS Estimators $(\widehat{\beta}_0, \widehat{\beta}_1)$ of $(\beta_0, \beta_1)$ are **linear** in $Y_i$ and **unbiased**, with

$$\mathbb{V}(\widehat{\beta}_0|X^n) = \sigma^2\left(\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\right) = \frac{\sigma^2}{n}\left(1 + \frac{\overline{X}^2}{S_X^2}\right)$$

$$\mathbb{V}(\widehat{\beta}_1|X^n) = \frac{\sigma^2}{\sum_{i=1}^n (X_i - \overline{X})^2} = \frac{\sigma^2}{nS_X^2}$$

$$\mathrm{Cov}(\widehat{\beta}_0, \widehat{\beta}_1|X^n) = -\frac{\overline{X}\sigma^2}{\sum_{i=1}^n (X_i - \overline{X})^2} = -\frac{\overline{X}\sigma^2}{nS_X^2}$$

- Estimates of these statistics are obtained by replacing the variance $\sigma^2$ by its estimator $\widehat{\sigma}^2$ (eg., the corrected MLE). The estimated standard errors $\widehat{\mathsf{se}}$ of $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are given by

$$\widehat{\mathsf{se}}(\widehat{\beta}_0|X^n) = \frac{\widehat{\sigma}}{\sqrt{n}}\sqrt{\left(1 + \frac{\overline{X}^2}{S_X^2}\right)} = \frac{\widehat{\sigma}}{\sqrt{n}S_X}\sqrt{\frac{\sum_{i=1}^n X_i^2}{n}}$$

$$\widehat{\mathsf{se}}(\beta_1|X^n) = \frac{\widehat{\sigma}}{\sqrt{n}S_X}; \quad \widehat{\mathrm{Cov}}(\widehat{\beta}_0, \widehat{\beta}_1|X^n) = -\frac{\overline{X}\widehat{\sigma}^2}{nS_X^2}$$

# Statistical Properties of the OLS Estimator

Let $S_X^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$ be the empirical variance of the $n$ covariates $X^n = (X_1, \ldots, X_n)$.

## Theorem : Linearity, Unbiasedness and Variance of the OLS

The OLS Estimators $(\widehat{\beta}_0, \widehat{\beta}_1)$ of $(\beta_0, \beta_1)$ are **linear** in $Y_i$ and **unbiased**, with

$$\mathbb{V}(\widehat{\beta}_0|X^n) = \sigma^2\left(\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}\right) = \frac{\sigma^2}{n}\left(1 + \frac{\overline{X}^2}{S_X^2}\right) = \frac{\sigma^2}{nS_X^2}\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right)$$

$$\mathbb{V}(\widehat{\beta}_1|X^n) = \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{\sigma^2}{nS_X^2}$$

$$\mathrm{Cov}(\widehat{\beta}_0, \widehat{\beta}_1|X^n) = -\frac{\overline{X}\sigma^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = -\frac{\overline{X}\sigma^2}{nS_X^2}$$

That is, $\mathrm{Cov}\left((\widehat{\beta}_0, \widehat{\beta}_1)^T\right) = \frac{\sigma^2}{nS_X^2}\begin{pmatrix}\frac{1}{n}\sum_{i=1}^{n} X_i^2 & -\overline{X} \\ -\overline{X} & 1\end{pmatrix}$

- Estimates of these statistics are obtained by replacing the variance $\sigma^2$ by its estimator $\widehat{\sigma}^2$ (eg., the corrected MLE). The estimated standard errors $\widehat{\mathrm{se}}$ of $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are given by

$$\widehat{\mathrm{se}}(\widehat{\beta}_0|X^n) = \frac{\widehat{\sigma}}{\sqrt{n}S_X}\sqrt{\frac{1}{n}\sum_{i=1}^{n} X_i^2}; \quad \widehat{\mathrm{se}}(\beta_1|X^n) = \frac{\widehat{\sigma}}{\sqrt{n}S_X}; \quad \widehat{\mathrm{Cov}}(\widehat{\beta}_0, \widehat{\beta}_1|X^n) = -\frac{\overline{X}\widehat{\sigma}^2}{nS_X^2}$$

# Statistical Properties of the OLS Estimator

Let $S_X^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2$ be the empirical variance of the $n$ covariates $X^n = (X_1, \ldots, X_n)$.

## Theorem : Linearity, Unbiasedness and Variance of the OLS

The OLS Estimators $(\widehat{\beta}_0, \widehat{\beta}_1)$ of $(\beta_0, \beta_1)$ are **linear** in $Y_i$ and **unbiased**, with

$$\mathbb{V}(\widehat{\beta}_0|X^n) = \sigma^2 \left(\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\right) = \frac{\sigma^2}{n}\left(1 + \frac{\overline{X}^2}{S_X^2}\right) = \frac{\sigma^2}{nS_X^2}\left(\frac{1}{n}\sum_{i=1}^n X_i^2\right)$$

$$\mathbb{V}(\widehat{\beta}_1|X^n) = \frac{\sigma^2}{\sum_{i=1}^n (X_i - \overline{X})^2} = \frac{\sigma^2}{nS_X^2}$$

$$\mathrm{Cov}(\widehat{\beta}_0, \widehat{\beta}_1|X^n) = -\frac{\overline{X}\sigma^2}{\sum_{i=1}^n (X_i - \overline{X})^2} = -\frac{\overline{X}\sigma^2}{nS_X^2}$$

That is, $\mathrm{Cov}\left((\widehat{\beta}_0, \widehat{\beta}_1)^T\right) = \frac{\sigma^2}{nS_X^2}\begin{pmatrix} \frac{1}{n}\sum_{i=1}^n X_i^2 & -\overline{X} \\ -\overline{X} & 1 \end{pmatrix}$

- Estimates of these statistics are obtained by replacing the variance $\sigma^2$ by its estimator $\widehat{\sigma}^2$ (eg., the corrected MLE). The estimated standard errors $\widehat{\text{se}}$ of $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are given by

$$\widehat{\text{se}}(\widehat{\beta}_0|X^n) = \frac{\widehat{\sigma}}{\sqrt{n}S_X}\sqrt{\frac{1}{n}\sum_{i=1}^n X_i^2}; \quad \widehat{\text{se}}(\beta_1|X^n) = \frac{\widehat{\sigma}}{\sqrt{n}S_X}; \quad \widehat{\mathrm{Cov}}(\widehat{\beta}_0, \widehat{\beta}_1|X^n) = -\frac{\overline{X}\widehat{\sigma}^2}{nS_X^2}$$

# Statistical Properties of the OLS Estimator

## Proof : Linearity of the OLS.

To simplify notation, let $w_i = \dfrac{X_i - \overline{X}}{\sum_{i=1}^n (X_i - \overline{X})^2}$.

Then we have :

$$\begin{aligned}
\sum_{i=1}^n w_i &= \frac{\sum_{i=1}^n X_i}{\sum_{i=1}^n (X_i - \overline{X})^2} - \frac{\sum_{i=1}^n \overline{X}}{\sum_{i=1}^n (X_i - \overline{x})^2} \\
&= \frac{n\overline{X}}{\sum_{i=1}^n (X_i - \overline{X})^2} - \frac{n\overline{X}}{\sum_{i=1}^n (X_i - \overline{X})^2} \\
&= 0
\end{aligned}$$

We can then write :

$$\begin{aligned}
\widehat{\beta}_1 &= \sum_{i=1}^n w_i (Y_i - \overline{Y}) = \sum_{i=1}^n w_i Y_i - \overline{Y} \sum_{i=1}^n w_i = \sum_{i=1}^n w_i Y_i \\
\widehat{\beta}_0 &= \overline{Y} - \widehat{\beta}_1 \overline{X} = \frac{1}{n} \sum_{i=1}^n Y_i - \overline{X} \sum_{i=1}^n w_i Y_i = \sum_{i=1}^n \left( \frac{1}{n} - \overline{X} w_i \right) Y_i
\end{aligned}$$

who are **linear** in $Y$. □

# Statistical Properties of the OLS Estimator

## Proof : Unbiasdness of the OLS.

We have $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$, with $\mathbb{E}[\epsilon_i|X] = 0$ and $\mathbb{V}[\epsilon_i|X] = \sigma^2$, then $\mathbb{E}[Y_i|X_i] = \beta_0 + \beta_1 X_i$
Knowing that $\widehat{\beta}_1 = \sum_{i=1}^n w_i Y_i$ and $\widehat{\beta}_0 = \sum_{i=1}^n \left(\frac{1}{n} - \overline{X} w_i\right) Y_i$, we can then write

$$\begin{aligned}
\mathbb{E}[\widehat{\beta}_1|X^n] &= \sum_{i=1}^n w_i(\beta_0 + \beta_1 x_i) = \beta_0 \sum_{i=1}^n w_i + \beta_1 \sum_{i=1}^n w_i x_i \\
&= \beta_1 \sum_{i=1}^n w_i x_i - \overline{x}\beta_1 \sum_{i=1}^n w_i = \beta_1 \sum_{i=1}^n w_i (x_i - \overline{x}) = \beta_1 \frac{\sum_{i=1}^n (x_i - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2} = \beta_1
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}[\widehat{\beta}_0|X^n] &= \sum_{i=1}^n \left(\frac{1}{n} - \overline{x} w_i\right)(\beta_0 + \beta_1 x_i) = \frac{1}{n}\sum_{i=1}^n (\beta_0 + \beta_1 x_i) - \overline{x} \sum_{i=1}^n w_i(\beta_0 + \beta_1 x_i) \\
&= \beta_0 + \beta_1 \overline{x} - \overline{x}\beta_0 \underbrace{\sum_{i=1}^n w_i}_{0} - \beta_1 \overline{x} \underbrace{\sum_{i=1}^n w_i x_i}_{1} = \beta_0
\end{aligned}$$

We used the fact that :

$$\sum_{i=1}^n w_i x_i = \sum_{i=1}^n w_i x_i - \overline{x} \sum_{i=1}^n w_i = \sum_{i=1}^n w_i (x_i - \overline{x}) = \frac{\sum_{i=1}^n (x_i - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2} = 1$$

# Statistical Properties of the OLS Estimator

**Proof : Variance of the OLS.**

$$
\begin{aligned}
w_i &= \frac{X_i - \overline{X}}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \\
\widehat{\beta}_1 &= \sum_{i=1}^{n} w_i Y_i \\
\mathrm{Var}(\widehat{\beta}_1|X^n) &= \mathbb{E}[\hat{\beta}_1^2 \big| X^n] - 0 \\
&= \mathbb{E}\left[\sum_{i=1}^{n} w_i Y_i \sum_{j=1}^{n} w_j Y_j \middle| X^n\right] \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j \mathbb{E}\left[Y_i Y_j \middle| X^n\right] = \sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j \sigma^2 \mathbb{1}_{i=j} \\
&= \sum_{i=1}^{n} w_i^2 \sigma^2 = \sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{(\sum_{i=1}^{n}(X_i - \overline{X})^2)^2}\sigma^2 \\
&= \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \\
&= \frac{\sigma^2}{n S_X^2}
\end{aligned}
$$

□

# Statistical Properties of the OLS Estimator

## Proof : Variance of the OLS.

$$\begin{aligned}
\hat{\beta}_0 &= \textstyle\sum_{i=1}^n \left(\frac{1}{n} - \bar{x}w_i\right) Y_i \\
\mathrm{Var}(\hat{\beta}_0|X^n) &= \mathbb{E}[\hat{\beta}_0^2|X^n] - 0 \\
&= \mathbb{E}\left[\textstyle\sum_{i=1}^n \left(\frac{1}{n} - \bar{x}w_i\right) Y_i \sum_{j=1}^n \left(\frac{1}{n} - \bar{x}w_j\right) Y_j\right] \\
&= \textstyle\sum_{i=1}^n \sum_{j=1}^n \left(\frac{1}{n} - \bar{x}w_i\right)\left(\frac{1}{n} - \bar{x}w_j\right) \mathbb{E}[Y_iY_j|X^n] \\
&= \textstyle\sum_{i=1}^n \sum_{j=1}^n \left(\frac{1}{n} - \bar{x}w_i\right)\left(\frac{1}{n} - \bar{x}w_j\right) \sigma^2 \mathbb{1}_{i=j} \\
&= \sigma^2 \textstyle\sum_{i=1}^n \left(\frac{1}{n} - \bar{x}w_i\right)^2 = \sigma^2 \left(\sum_{i=1}^n \frac{1}{n^2} - \frac{2\bar{x}}{n}\sum_{i=1}^n w_i + \bar{x}^2 \sum_{i=1}^n w_i^2\right) \\
&= \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right) \\
&= \frac{\sigma^2}{n}\left(1 + \frac{\bar{x}}{S_X^2}\right)
\end{aligned}$$

Since $S_X^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^n x_i^2 - \bar{x}$, then

$$\mathrm{Var}(\hat{\beta}_0|X^n) = \frac{\sigma^2}{nS_X^2}\left(\frac{1}{n}\sum_{i=1}^n x_i^2\right)$$

□

# Statistical Properties of the OLS Estimator

## Proof : Variance of the OLS.

$$
\begin{aligned}
\mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1 \big| X^n) &= \mathrm{Cov}\left( \sum_{i=1}^{n} \left(\frac{1}{n} - \bar{x} w_i\right) Y_i, \sum_{j=1}^{n} w_j Y_j \Big| X^n \right) \\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} \left(\frac{1}{n} - \bar{x} w_i\right) w_j \,\mathrm{Cov}\left(Y_i, Y_j \big| X^n\right) \\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} \left(\frac{1}{n} - \bar{x} w_i\right) w_j \sigma^2 \mathbb{1}_{i=j} \\
&= \sigma^2 \sum_{i=1}^{n} \left(\frac{1}{n} - \bar{x} w_i\right) w_i = \sigma^2 \left( \frac{1}{n} \sum_{i=1}^{n} w_i - \bar{x} \sum_{i=1}^{n} w_i^2 \right) \\
&= -\frac{\bar{x}\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \\
&= -\frac{\bar{x}\sigma^2}{n S_X^2}
\end{aligned}
$$

□

# Statistical Properties of the OLS Estimator

**Theorem (Gauss-Markov)**

The OLS estimator is the unique linear unibiased estimator with minimum variance :
The OLS estimator is the Best Linear Unbiased Estimator (BLUE)

We give the proof of this result in the multiple regression part.

# Statistical Properties of the OLS Estimator

### Theorem (Consistency)

*The OLS estimators $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are consistent.*

### Proof : Consistency.

Let $(\widehat{\beta}_j^{(n)})_{n\in\mathbb{N}^*}$, $j \in \{0,1\}$, be an estimator sequence derived from increasing sample sizes. By Bienaymé-Tchebychev ineq.

$$\forall \lambda > 0, \mathbb{P}\left(\left|\widehat{\beta}_j^{(n)} - \mathbb{E}[\widehat{\beta}_j^{(n)}]\right| \geqslant \lambda\right) \leqslant \frac{\mathbb{V}\left(\widehat{\beta}_j^{(n)}\right)}{\lambda^2}.$$

Then $\mathbb{P}\left(\left|\widehat{\beta}_1^{(n)} - \beta_1\right| \geqslant \lambda\right) \leqslant \frac{\frac{\sigma^2}{nS_X^2}}{\lambda^2}$ and $\mathbb{P}\left(\left|\widehat{\beta}_0^{(n)} - \beta_0\right| \geqslant \lambda\right) \leqslant \frac{\frac{\sigma^2}{n}\left(1+\frac{\bar{x}}{S_X^2}\right)}{\lambda^2}$.

Thus $0 \leqslant \lim_{n\to\infty} \mathbb{P}\left(\left|\widehat{\beta}_1^{(n)} - \beta_1\right| \geqslant \lambda\right) \leqslant \frac{\sigma^2}{\lambda^2 S_X^2} \lim_{n\to\infty} \frac{1}{n} = 0$,

and $0 \leqslant \lim_{n\to\infty} \mathbb{P}\left(\left|\widehat{\beta}_0^{(n)} - \beta_0\right| \geqslant \lambda\right) \leqslant \frac{\sigma^2\left(1+\frac{\bar{x}}{S_X^2}\right)}{\lambda^2} \lim_{n\to\infty} \frac{1}{n} = 0$,

That is $\lim_{n\to\infty} \mathbb{P}\left(\left|\widehat{\beta}_j^{(n)} - \beta_j\right| \geqslant \lambda\right) = 0$ ; Then the sequence $(\widehat{\beta}_j^{(n)})_{n\in\mathbb{N}^*}$ converges in probability to $\beta_j$ ($\text{plim}_{n\to\infty}\widehat{\beta}_j^{(n)}=\beta_j$) : the OLS Estimators $\widehat{\beta}_j$ are consistent. □

# Statistical Properties of the OLS Estimator

## Theorem (Consistency)

*The OLS estimators $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are consistent.*

## Proof : Consistency.

Let $(\widehat{\beta}_j^{(n)})_{n\in\mathbb{N}^*}$, $j \in \{0,1\}$, be an estimator sequence derived from increasing sample sizes. By Bienaymé-Tchebychev ineq.

$$\forall\lambda > 0,\ \mathbb{P}\left(\left|\widehat{\beta}_j^{(n)} - \mathbb{E}[\widehat{\beta}_j^{(n)}]\right| \geqslant \lambda\right) \leqslant \frac{\mathbb{V}\left(\widehat{\beta}_j^{(n)}\right)}{\lambda^2}.$$

Then $\mathbb{P}\left(\left|\widehat{\beta}_1^{(n)} - \beta_1\right| \geqslant \lambda\right) \leqslant \frac{\frac{\sigma^2}{nS_X^2}}{\lambda^2}$ and $\mathbb{P}\left(\left|\widehat{\beta}_0^{(n)} - \beta_0\right| \geqslant \lambda\right) \leqslant \frac{\frac{\sigma^2}{n}\left(1+\frac{\bar{x}}{S_X^2}\right)}{\lambda^2}$.
Thus $0 \leqslant \lim_{n\to\infty} \mathbb{P}\left(\left|\widehat{\beta}_1^{(n)} - \beta_1\right| \geqslant \lambda\right) \leqslant \frac{\sigma^2}{\lambda^2 S_X^2} \lim_{n\to\infty} \frac{1}{n} = 0,$

and $0 \leqslant \lim_{n\to\infty} \mathbb{P}\left(\left|\widehat{\beta}_0^{(n)} - \beta_0\right| \geqslant \lambda\right) \leqslant \frac{\sigma^2\left(1+\frac{\bar{x}}{S_X^2}\right)}{\lambda^2} \lim_{n\to\infty} \frac{1}{n} = 0,$

That is $\lim_{n\to\infty} \mathbb{P}\left(\left|\widehat{\beta}_j^{(n)} - \beta_j\right| \geqslant \lambda\right) = 0$; Then the sequence $(\widehat{\beta}_j^{(n)})_{n\in\mathbb{N}^*}$ converges in probability to $\beta_j$ (plim$_{n\to\infty}\widehat{\beta}_j^{(n)} = \beta_j$) : the OLS Estimators $\widehat{\beta}_j$ are consistent. □

# Statistical Properties of the OLS Estimator

**Theorem (Asymptotic normality)**

*The OLS estimators $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are asymptotically normal.*

**Proof : Asymptotic normality.**

By the Central Limit Theorem (CLT), the sequence $(Z_n)_{n\in\mathbb{N}^*}$ such that $Z_n = \frac{\widehat{\beta}_j^{(n)} - \mathbb{E}[\widehat{\beta}_j^{(n)}]}{\sqrt{\mathbb{V}(\widehat{\beta}_j^{(n)})}}$ converges to a standard normal random variable.

Then the limit distribution of $\frac{\widehat{\beta}_j^{(n)} - \beta_j}{\sqrt{\mathbb{V}(\widehat{\beta}_j^{(n)})}}$ is $\mathcal{N}(0,1)$ :

$$\frac{\widehat{\beta}_1 - \beta_1}{\sqrt{\frac{\sigma^2}{nS_X^2}}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1) \text{ and } \frac{\widehat{\beta}_0 - \beta_0}{\sqrt{\frac{\sigma^2}{nS_X^2}\left(\frac{1}{n}\sum_{i=1}^n X_i^2\right)}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$

Equivalently :
$\sqrt{n}(\widehat{\beta}_1 - \beta_1) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \frac{\sigma^2}{S_X^2}\right)$ and $\sqrt{n}(\widehat{\beta}_0 - \beta_0) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \frac{\sigma^2}{S_X^2}\left(\frac{1}{n}\sum_{i=1}^n X_i^2\right)\right)$. □

# Statistical Properties of the OLS Estimator

## Theorem (Asymptotic normality)

*The OLS estimators $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are asymptotically normal.*

## Proof : Asymptotic normality.

By the Central Limit Theorem (CLT), the sequence $(Z_n)_{n\in\mathbb{N}^*}$ such that $Z_n = \frac{\widehat{\beta}_j^{(n)} - \mathbb{E}[\widehat{\beta}_j^{(n)}]}{\sqrt{\mathbb{V}(\widehat{\beta}_j^{(n)})}}$ converges to a standard normal random variable.

Then the limit distribution of $\frac{\widehat{\beta}_j^{(n)} - \beta_j}{\sqrt{\mathbb{V}(\widehat{\beta}_j^{(n)})}}$ is $\mathcal{N}(0,1)$ :

$$\frac{\widehat{\beta}_1 - \beta_1}{\sqrt{\frac{\sigma^2}{nS_X^2}}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1) \text{ and } \frac{\widehat{\beta}_0 - \beta_0}{\sqrt{\frac{\sigma^2}{nS_X^2}\left(\frac{1}{n}\sum_{i=1}^n X_i^2\right)}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$

Equivalently :
$\sqrt{n}(\widehat{\beta}_1 - \beta_1) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \frac{\sigma^2}{S_X^2}\right)$ and $\sqrt{n}(\widehat{\beta}_0 - \beta_0) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \frac{\sigma^2}{S_X^2}\left(\frac{1}{n}\sum_{i=1}^n X_i^2\right)\right).$ □

# Statistical Properties of the OLS Estimator

## Theorem (Efficiency)

*The OLS estimators $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are efficient (achieve minimum variance (CRLB)).*

## Proof : Efficiency.

Consider the SLR with normal errors $\epsilon_i \underset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$. Then $Y_i | x_i \underset{\text{iid}}{\sim} \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2)$.
Then, the Fisher information matrix can be defined as

$$\mathcal{I}_n(\beta_0, \beta_1) = -\mathbb{E}\left[\left(\frac{\partial^2 L(\beta_0, \beta_1)}{\partial \beta_i \partial \beta_j}\right)_{i,j=0,1}\right]$$

with $L(\beta_0, \beta_1)$ is the conditional log-likelihood function as given by

$$\begin{aligned} L(\beta_0, \beta_1) &= \log p(y_1, \ldots, y_n | x_1, \ldots, x_n; \beta_0, \beta_1) = \log \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}\left(\frac{y_i - (\beta_0 + \beta_1 x_i)}{\sigma}\right)^2} \\ &= \sum_{i=1}^{n} \log \frac{1}{\sigma\sqrt{2\pi}} - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2 . \end{aligned}$$

□

# Statistical Properties of the OLS Estimator

## Theorem (Efficiency)

*The OLS estimators $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are efficient (achieve minimum variance (CRLB)).*

## Proof : Efficiency.

Consider the SLR with normal errors $\epsilon_i \underset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$. Then $Y_i | x_i \underset{\text{iid}}{\sim} \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2)$.
Then, the Fisher information matrix can be defined as

$$\mathcal{I}_n(\beta_0, \beta_1) = -\mathbb{E}\left[\left(\frac{\partial^2 L(\beta_0, \beta_1)}{\partial \beta_i \partial \beta_j}\right)_{i,j=0,1}\right]$$

with $L(\beta_0, \beta_1)$ is the conditional log-likelihood function as given by

$$\begin{aligned} L(\beta_0, \beta_1) &= \log p(y_1, \ldots, y_n | x_1, \ldots, x_n; \beta_0, \beta_1) = \log \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\,\pi}}\, \mathrm{e}^{-\frac{1}{2}\left(\frac{y_i - (\beta_0 + \beta_1 x_i)}{\sigma}\right)^2} \\ &= \sum_{i=1}^{n} \log \frac{1}{\sigma\sqrt{2\,\pi}} - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2 \,. \end{aligned}$$

□

# Statistical Properties of the OLS Estimator

## Proof : Efficiency (cont.)

$$\frac{\partial L(\beta_0,\beta_1)}{\partial \beta_0} = \frac{\partial}{\partial \beta_0}\{-\frac{1}{2\sigma^2}\sum_{i=1}^n \big(y_i - (\beta_0+\beta_1 x_i)\big)^2\} = \frac{1}{\sigma^2}\sum_{i=1}^n \big(y_i - (\beta_0+\beta_1 x_i)\big)$$

$$\frac{\partial L(\beta_0,\beta_1)}{\partial \beta_1} = \frac{\partial}{\partial \beta_1}\{-\frac{1}{2\sigma^2}\sum_{i=1}^n \big(y_i - (\beta_0+\beta_1 x_i)\big)^2\} = \frac{1}{\sigma^2}\sum_{i=1}^n \big(y_i - (\beta_0+\beta_1 x_i)\big)x_i.$$

$$\frac{\partial^2 L(\beta_0,\beta_1)}{\partial \beta_0^2} = \frac{\partial}{\partial \beta_0}\{\frac{1}{\sigma^2}\sum_{i=1}^n \big(y_i - (\beta_0+\beta_1 x_i)\big)\} = -\frac{n}{\sigma^2}$$

$$\frac{\partial^2 L(\beta_0,\beta_1)}{\partial \beta_1^2} = \frac{\partial}{\partial \beta_1}\{\frac{1}{\sigma^2}\sum_{i=1}^n \big(y_i - (\beta_0+\beta_1 x_i)\big)x_i\} = -\frac{1}{\sigma^2}\sum_{i=1}^n x_i^2.$$

$$\frac{\partial^2 L(\beta_0,\beta_1)}{\partial \beta_0 \partial \beta_1} = \frac{\partial}{\partial \beta_0}\{\frac{1}{\sigma^2}\sum_{i=1}^n \big(y_i - (\beta_0+\beta_1 x_i)\big)x_i\} = -\frac{1}{\sigma^2}\sum_{i=1}^n x_i = -\frac{n}{\sigma^2}\bar{x}.$$

Then the Fisher information is (fixed design here)

$$\mathcal{I}_n(\beta_0,\beta_1) = -\mathbb{E}\begin{pmatrix} -\frac{n}{\sigma^2} & -\frac{n}{\sigma^2}\bar{x} \\ -\frac{n}{\sigma^2}\bar{x} & -\frac{1}{\sigma^2}\sum_{i=1}^n x_i^2 \end{pmatrix} = \frac{n}{\sigma^2}\begin{pmatrix} 1 & \bar{x} \\ \bar{x} & \frac{1}{n}\sum_{i=1}^n x_i^2 \end{pmatrix}$$

□

# Statistical Properties of the OLS Estimator

## Proof : Efficiency (cont.)

The Cramer-Rao Lower Bound is given by the inverse of the Fisher information matrix

$$\mathcal{I}_n^{-1}(\beta_0,\beta_1) = \frac{\sigma^2}{n}\begin{pmatrix}1 & \bar{x}\\ \bar{x} & \frac{1}{n}\sum_{i=1}^n x_i^2\end{pmatrix}^{-1} = \frac{\sigma^2}{n}\times\frac{1}{\frac{1}{n}\sum_{i=1}^n x_i^2 - \bar{x}^2}\begin{pmatrix}\frac{1}{n}\sum_{i=1}^n x_i^2 & -\bar{x}\\ -\bar{x} & 1\end{pmatrix}$$

Since $S_X^2 = \frac{1}{n}\sum_{i=1}^n (x_i-\bar{x})^2 = \frac{1}{n}\sum_{i=1}^n x_i^2 - \bar{x}$, finally we get

$$\begin{aligned}\mathcal{I}_n^{-1}(\beta_0,\beta_1) &= \frac{\sigma^2}{nS_X^2}\begin{pmatrix}\frac{1}{n}\sum_{i=1}^n x_i^2 & -\bar{x}\\ -\bar{x} & 1\end{pmatrix}\\ &= \mathrm{Cov}\left((\widehat{\beta}_0,\widehat{\beta}_1)^T\right)\end{aligned}$$

The OLS Estimators $\widehat{\beta}_0$ and $\widehat{\beta}_1$ then achieve the Cramer-Rao Lower Bound. □

# Confidence Intervals

Let $\mathrm{se}(\widehat{\beta}_1) = \sqrt{\frac{\sigma^2}{n S_X^2}}$ and $\mathrm{se}(\widehat{\beta}_0) = \sqrt{\frac{\sigma^2}{n}\left(1 + \frac{\bar{x}}{S_X^2}\right)}$. We then have

$$Z_j = \frac{\widehat{\beta}_j - \beta_j}{\mathrm{se}(\widehat{\beta}_j)} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$

Since the variance $\sigma^2$ is unknown, we use instead its best estimator : the corrected MLE

$$\widehat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - (\widehat{\beta}_0^{\mathsf{OLS}} + \widehat{\beta}_1^{\mathsf{OLS}} X_i))^2$$

We then use instead the statistic

$$T_j = \frac{\widehat{\beta}_j - \beta_j}{\widehat{\mathrm{se}}(\widehat{\beta}_j)}$$

where $\widehat{\mathrm{se}}(\widehat{\beta}_j)$ corresponds to replacing $\sigma^2$ by $\widehat{\sigma}^2$ in $\mathrm{se}(\widehat{\beta}_j)$.

# Confidence Intervals II

We know that

$$U = \frac{(n-2)\widehat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2}$$

Then we finally have

$$T_j = \frac{\widehat{\beta}_j - \beta_j}{\widehat{\mathrm{se}}(\widehat{\beta}_j)} = \frac{Z_j}{\sqrt{\frac{U}{n-2}}} \sim \mathcal{T}_{n-2}$$

Let $\mathcal{T}_{1-\frac{\alpha}{2}} = \mathbb{P}(T_j \leq \frac{\alpha}{2})$, i.e the quantile of order $\frac{\alpha}{2}$ of the Student's law with $n-2$ degrees of freedom. An approximate $1-\alpha$ confidence interval for $\widehat{\beta}_j$ is then given by

$$\mathbb{P}(-\mathcal{T}_{n-2,\frac{\alpha}{2}} \leq T_j \leq \mathcal{T}_{n-2,\frac{\alpha}{2}}) = 1-\alpha$$

which corresponds to

$$\mathbb{P}\left(\beta_j - \mathcal{T}_{n-2,\frac{\alpha}{2}}\widehat{\mathrm{se}}(\widehat{\beta}_j) \leq \widehat{\beta}_j \leq \beta_j + \mathcal{T}_{n-2,\frac{\alpha}{2}}\widehat{\mathrm{se}}(\widehat{\beta}_j)\right) = 1-\alpha.$$

We finally obtain

$$\mathrm{CI}_{1-\alpha}(\widehat{\beta}_j) = \left[\beta_j - \mathcal{T}_{n-2,\frac{\alpha}{2}}\widehat{\mathrm{se}}(\widehat{\beta}_j)\,,\, \beta_j + \mathcal{T}_{n-2,\frac{\alpha}{2}}\widehat{\mathrm{se}}(\widehat{\beta}_j)\right].$$

# Confidence Intervals III

Finally :

$$\begin{aligned}
\mathrm{CI}_{1-\alpha}(\widehat{\beta}_0) &= \left[\widehat{\beta}_0 \pm \mathcal{T}^{n-2}_{1-\frac{\alpha}{2}} \sqrt{\sigma^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right)}\right] \\
\mathrm{CI}_{1-\alpha}(\widehat{\beta}_1) &= \left[\widehat{\beta}_1 \pm \mathcal{T}^{n-2}_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma^2}{\sum_{i=1}^n (x_i - \overline{x})^2}}\right]
\end{aligned}$$

# Confidence Intervals IV

## Confidence interval for the regression line

We can construct a confidence interval for $\widehat{h}(x_i) = \mathbb{E}[Y_i|X_i = x_i; \widehat{\beta}_0, \widehat{\beta}_1] = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$.
Since we have

$$\begin{aligned}
\mathrm{Var}(\widehat{\beta}_0 + \widehat{\beta}_1 x_i) &= \mathrm{Var}(\widehat{\beta}_0) + \mathrm{Var}(\widehat{\beta}_1 x_i)) + 2\,\mathrm{Cov}(\widehat{\beta}_0, \widehat{\beta}_1 x_i) \\
&= \mathrm{Var}(\widehat{\beta}_0) + x_i^2\,\mathrm{Var}(\widehat{\beta}_1) + 2\,\mathrm{Cov}(\widehat{\beta}_0, \widehat{\beta}_1) x_i \\
&= \frac{\sigma^2}{n}\left(1 + \frac{\overline{x}^2}{S_X^2}\right) + x_i^2 \frac{\sigma^2}{n S_X^2} - 2\frac{\overline{x}\sigma^2}{n S_X^2} x_i \\
&= \sigma^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{S_X^2} + \frac{x_i^2}{n S_X^2} - 2\frac{\overline{x}}{n S_X^2} x_i\right) \\
&= \sigma^2 \left(\frac{1}{n} + \frac{(x_i - \overline{x})^2}{n S_X^2}\right) \\
&= \sigma^2 \left(\frac{1}{n} + \frac{(x_i - \overline{x})^2}{\sum_{j=1}^n (x_j - \overline{x})^2}\right)
\end{aligned}$$

We then obtain $\mathrm{CI}_{1-\alpha}(\widehat{\beta}_0 + \widehat{\beta}_1 x_i) = \left[\widehat{\beta}_0 + \widehat{\beta}_1 x_i \pm \mathcal{T}_{1-\frac{\alpha}{2}}^{n-2} \sqrt{\sigma^2 \left(\frac{1}{n} + \frac{(x_i - \overline{x})^2}{\sum_{j=1}^n (x_j - \overline{x})^2}\right)}\right]$

# Confidence Intervals

## Prediction interval

Given a new input $x_*$, the predicted output $Y_*$ is given by

$$Y_* = \widehat{h}(x_*) + \epsilon_* = \widehat{\beta}_0 + \widehat{\beta}_1 x_* + \epsilon_*$$

Since $\epsilon_*$ is not observed, then independent from the training set, the variance of the predicted value is then

$$\begin{aligned}
\mathrm{Var}(Y_*) &= \mathrm{Var}(\widehat{\beta}_0 + \widehat{\beta}_1 x_*) + \mathrm{Var}(\epsilon_*) \\
&= \sigma^2 \left( \frac{1}{n} + \frac{(x_i - \overline{x})^2}{\sum_{j=1}^{n}(x_j - \overline{x})^2} \right) + \sigma^2 \\
&= \sigma^2 \left( 1 + \frac{1}{n} + \frac{(x_i - \overline{x})^2}{\sum_{j=1}^{n}(x_j - \overline{x})^2} \right)
\end{aligned}$$

and we have $\mathrm{CI}_{1-\alpha}(Y_*) = \left[ \widehat{\beta}_0 + \widehat{\beta}_1 x_* \pm \mathcal{T}^{n-2}_{1-\frac{\alpha}{2}} \sqrt{\sigma^2 \left( 1 + \frac{1}{n} + \frac{(x_* - \overline{x})^2}{\sum_{j=1}^{n}(x_* - \overline{x})^2} \right)} \right]$.

This one (on $Y_*$) is larger compared to the previous one (on $\mathbb{E}[Y_* | X = x_*]$).

## Regression with Gaussian errors

Let $\mathcal{X} = \mathbb{R}$, $y \in \mathbb{R}$ and $h\colon \mathcal{X} \to \mathcal{Y}$ s.t $x \mapsto \beta_0 + \beta_1 x$, and consider the following model

$$Y_i = h(X_i; \beta_0, \beta_1) + \varepsilon_i \quad \text{with} \quad \varepsilon_i | X \underset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$$

- Empirical Risk : under the square loss $R_n(\beta_0, \beta_1) = \frac{1}{n} \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2$
- Empirical Risk Minimizer : $(\widehat{\beta}_0, \widehat{\beta}_1)_n = \arg\min_{(\beta_0, \beta)} R_n(\beta_0, \beta_1)$
- Conditional Maximum Likelihood Risk

  Data model : $Y_i | X_i \underset{\text{iid}}{\sim} \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2)$ : $p_{\boldsymbol{\theta}}(y_i | x_i) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y - (\beta_0 + \beta_1 x_i)}{\sigma}\right)^2}$

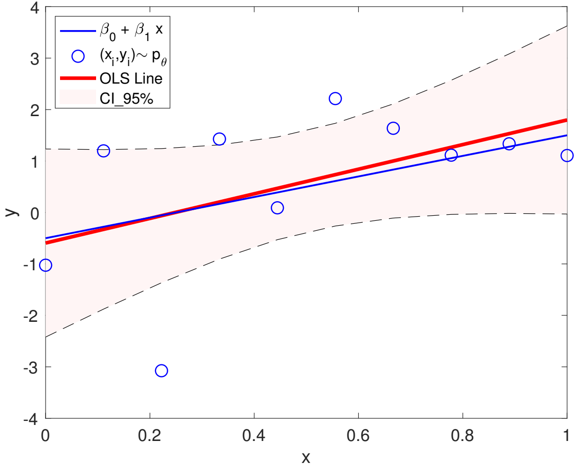
$$\log L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log p_{\boldsymbol{\theta}}(y_i | x_i) = -\frac{1}{2\sigma^2} \underbrace{\sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2}_{R_n(\boldsymbol{\beta})} - \frac{n}{2} \log \sigma^2 - \frac{n}{2} \log(2\pi)$$

- Conditional MLE : $= (\widehat{\beta}_0^{(n)}, \widehat{\beta}_1^{(n)}) = \arg\max_{(\beta_0, \beta_1)} \log L(\boldsymbol{\theta})$

$\hookrightarrow$ Then we have : $\arg\min_{(\beta_0, \beta)} R_n(\beta_0, \beta_1) = \arg\max_{(\beta_0, \beta)} \log L(\boldsymbol{\theta})$.

- For both we can take the sample variance as an estimator of the variance $\sigma^2$ : $\widehat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (Y_i - (\widehat{\beta}_0, \widehat{\beta}_1 x_i))^2$ which is the Maximum-Likelihood Estimator

# SLR illustration




$y_i = -\frac{1}{2} + 2x_i + \epsilon_i$, $\epsilon_i \sim \mathcal{N}(0,1)$ and $x_i$'s are $n$ values from a uniform grid in $[0,1]$

**Practical work**: Coding session to implement from the scratch the confidence interval calculation for regression

$y_i = -\frac{1}{2} + 2x_i + \epsilon_i,$
$\epsilon_i \sim \mathcal{N}(0, 1)$ and $x_i$'s are values from a uniform grid in $[0, 1]$

**Practical work session** : Confidence intervals and prediction using linear regression

- Simulated data
- Appart data prediction

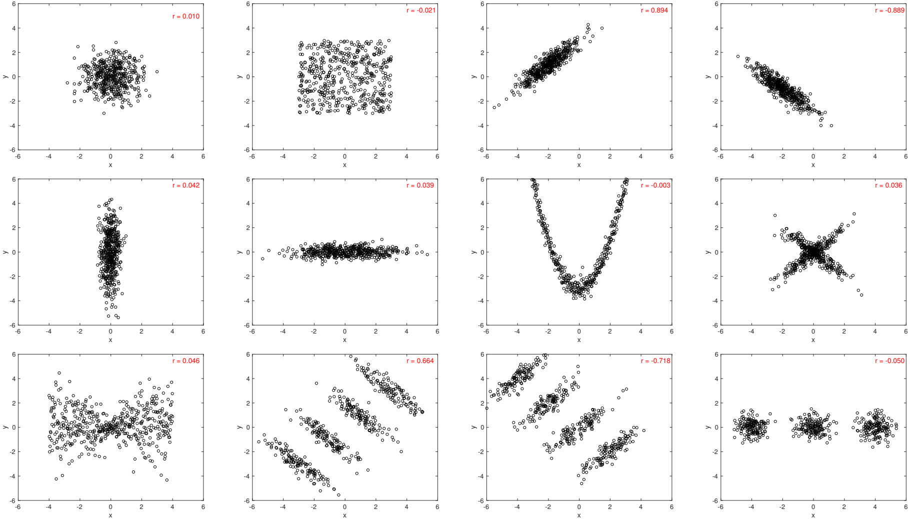
  Python, R, and MATLAB code provided during the session and accessible here on the course's page :

# Goodness of Fit

# Correlation coefficient

Correlation coefficient $\rho_{XY} = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)\,\text{Var}(Y)}} = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y} \in [-1, 1]$.

Sample Correlation coefficient : $r_{XY} = \frac{S_{XY}}{S_X S_Y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$.



# Coefficient of determination $R^2$

Measures the **quality of fit**

- $\mathrm{SST} = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$ : sum of the squares of the deviations around $\overline{Y}$ :

  a measure of the total variability for the $n$ given observations, $Y_i$'s.
- $\mathrm{SSE} = \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2$ : The sum of the squares of the deviations around the $\widehat{Y}_i$'s :

  A measure of the variability in $Y$ that remains after the regression is fitted.
- $\frac{\mathrm{SSE}}{\mathrm{SST}} = \frac{\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}$ : proportion of the total variability unexplained by the fitted regression
- $R^2$ : proportion of the total variability accounted for by the regression :

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2} = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}}$$

- For Simple Linear Regression, i.e. $\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$ : $R^2 = r^2$

  $R^2$ is the correlation coefficient (squared)
- NB : In the general case, $R^2$ is not a coefficient correlation (and should not be confused with).

# Coefficient of determination $R^2$

- In simple linear regression : $R^2 \simeq 1$ indicates that the empirical correlation coefficient between the response $y$ and the predictor $x$ is close to 1, so that a modeling by a line is satisfactory
- In general : High value of $R^2$ indicates that the regression model is well fitted to the data. However, it is not an indicator of how good the prediction capability of the fitted model is. For example, a model with $R^2 \approx 1$ will have high variance (and hence over-fits the data)

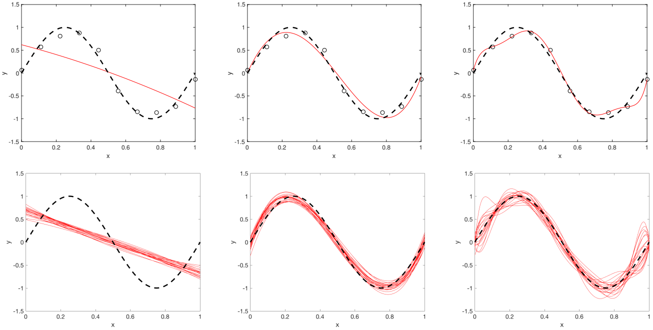



FIGURE – Sample (∘), True function (- -), realizations of the fitted prediction function (—)