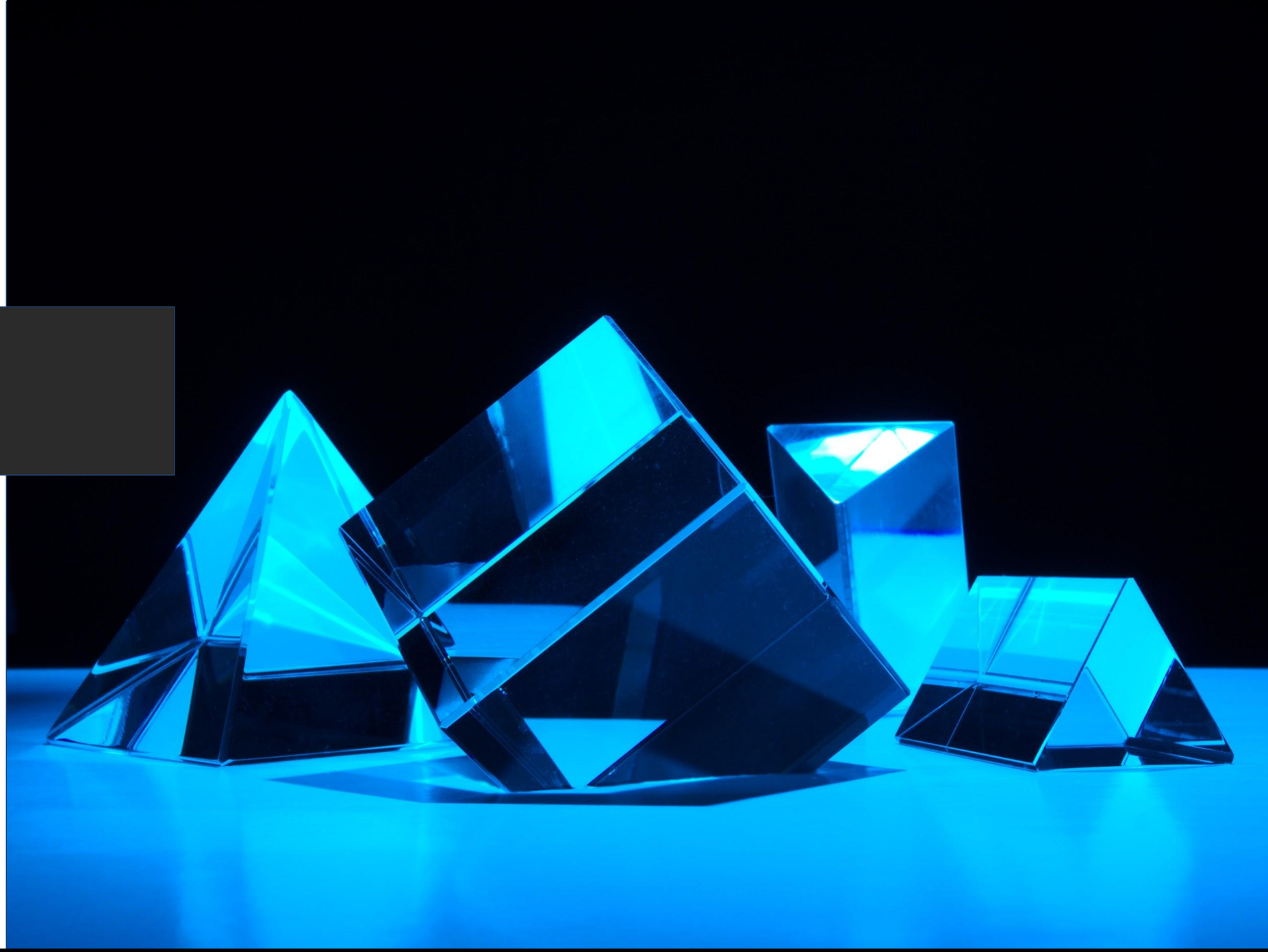
LES FACULTÉS
DE L'UNIVERSITÉ
CATHOLIQUE DE LILLE

Linear Models examples

SIMPLE LINEAR REGRESSION



Baptiste Mokas

baptiste.mokas@gmail.com

weeki.io/dynamical-système

linktr.ee/baptistemokas

+33 7 69 08 54 19

KNOWLEDGE TREE

Part 1: Foundations of Simple Linear Regression

1.1 Introduction to the Model

- Random variable
- Endogenous variable

1.2 Model Estimation

- Model form

1.3 Model Assumptions

1.4 Ordinary Least Squares Line

- Line equation
- Sum of squared residuals
- Partial derivatives
- Exercise
- Line equation
- Interpretation of the slope

1.5 Properties of OLS Estimators

- Theorem
- Unbiasedness of OLS estimators
- Theorem 2
- Gauss-Markov theorem
- Linear combination

1.6 Graphical Illustrations

1.7 1.7 Residual Analysis

- Measuring the quality of fit
- Analysis of variance equation
- Decomposition
- Equalities

1.8 Estimation of σ^2

- Definition
- Theorem
- Proof

1.9 Exercise on Empirical Mean

1.10 Confidence Interval

- Results from the previous theorem
- Confidence interval

1.11 Confidence Interval for σ^2

- Percentiles of a distribution

1.12 Confidence Region for σ^2

- Order of a Fisher-Snedecor distribution

1.13 Hypothesis Testing / Model Significance

1.14 Overall Model Significance Test

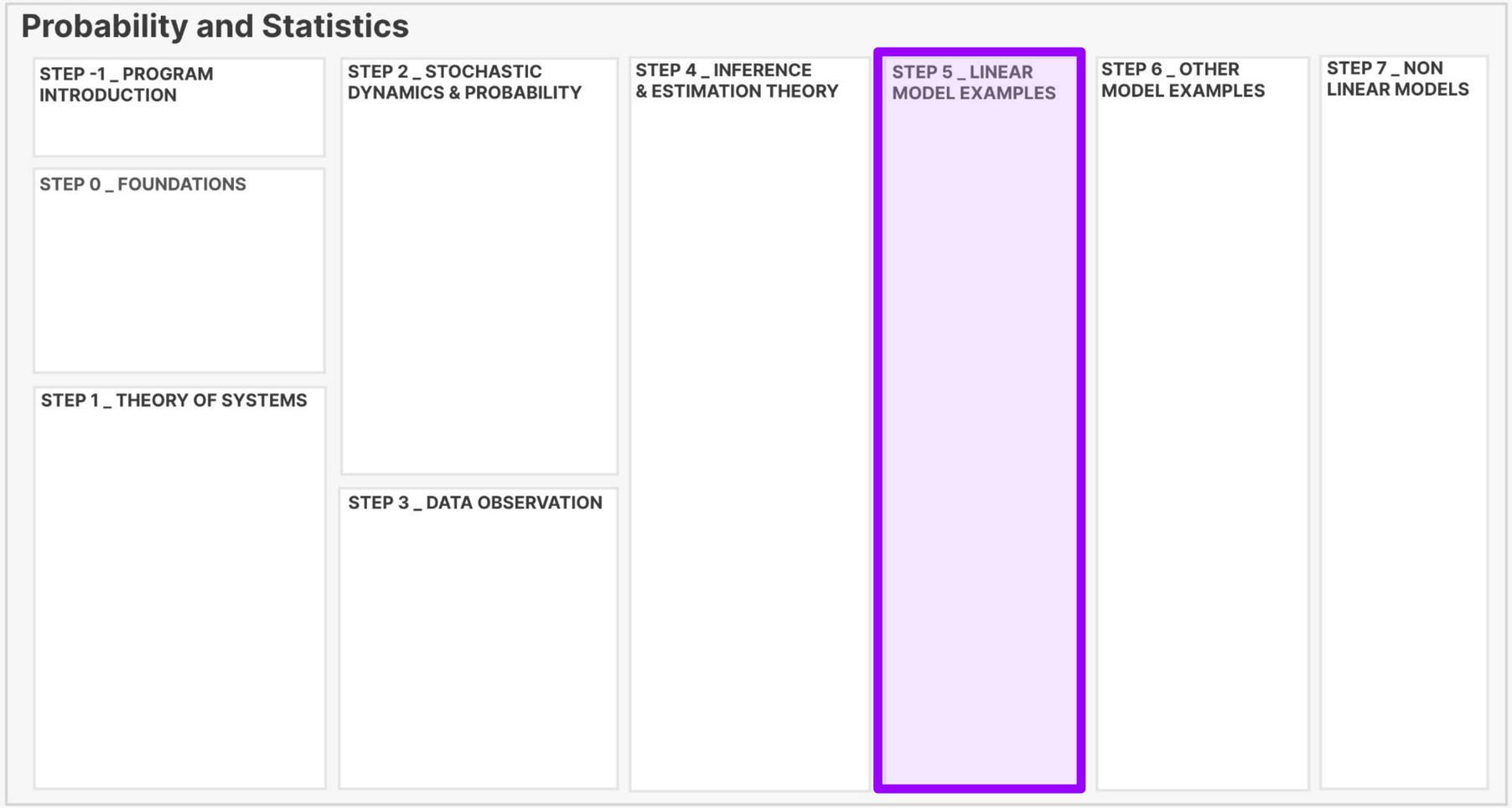
- Decision rule
- DROITEREG()

1.15 Prediction

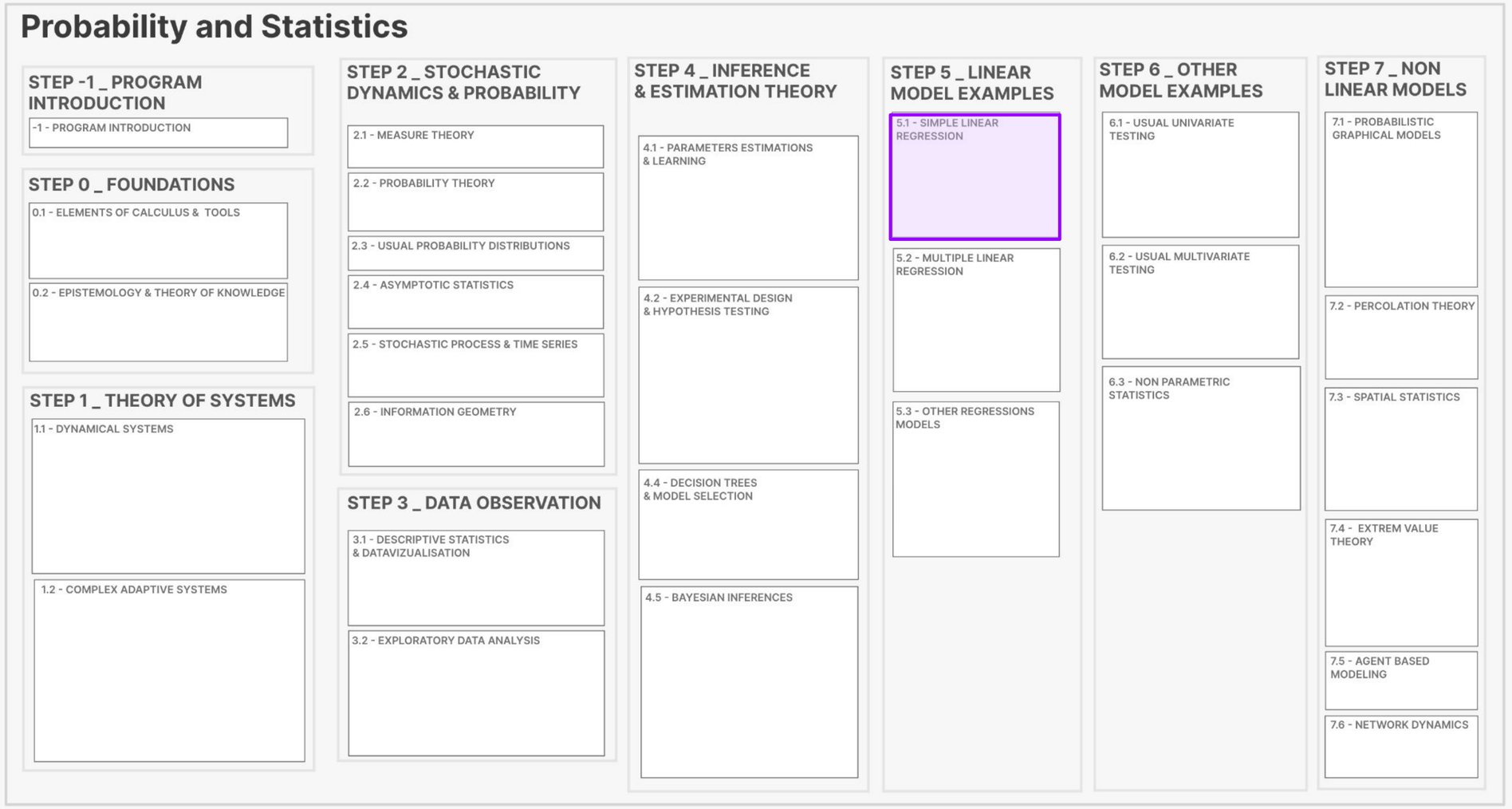
- Predicting the value of Y
- Prediction error
- Estimated variance of prediction error

1.16 Precision Interval

OUR STORY LINE



OUR STORY LINE



Random variable

Part 1: Foundations of Simple Linear Regression

1.1 Introduction to the Model

Let Y be the real random variable to be explained, and X be the explanatory random variable for Y.

The simple linear regression model assumes that, on average, Y is an affine function of X:

$$\mathbb{E}(Y) = f(X) \text{ where } f(t) = \beta_0 + \beta_1 t.$$

Alternatively, we can write it as:

$$Y = \beta_0 + \beta_1 X + \varepsilon \text{ where } \mathbb{E}(\varepsilon) = 0.$$

Endogenous variable

Part 1: Foundations of Simple Linear Regression

1.1 Introduction to the Model

Y is called the endogenous variable,
which is the variable we are trying to predict its values (predicted variable,
dependent variable, explained variable, response variable).

X is called the exogenous variable,

which is the variable used to explain and predict the values of Y (predictor variable,
independent variable, explanatory variable, fixed effect).

$f(t) = \beta_0 + \beta_1 t$ is the regression curve.

ε represents the residuals of the model,
the error term (a random variable that captures everything the model does not explain).

Note
For simplicity, we will assume that X is deterministic
(i.e., perfectly controlled by the experimenter) and denote it as x.
In the contrary case, the model (4) is then written conditionally on the observations of

X $(\mathbb{E}(\varepsilon/X) = 0)$

and leads to the same estimations with more advanced computational tools.

Model form

Therefore, we have a sample of $n(n > 1)$ pairs of points (x_i, y_i),
and we want to explain (predict) the values taken by Y based on the values taken by x.

Once the model form is validated, it can be written as:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i = 1, \ldots, n.$$

We need to find the pair (β_0, β_1) that allows us to obtain
$\beta_0 + \beta_1 x_i$ "closest" to the observations y_i of the variable Y.

The parameters of interest, β_0 and β_1, respectively represent the constant term (intercept)
and the slope of the regression line: for $x_i = 0, \mathbb{E}(Y_i) = \beta_0$.

Model Assumptions

In order to determine the properties of the estimators β_0 and β_1
and establish the tools of inferential statistics, we need to make some assumptions about the residuals.

H_ε
(i) The ε_i are i.i.d. (independent and identically distributed).
(ii) For all $i = 1, \ldots, n$, $\mathbb{E}(\varepsilon_i) = 0$:
On average, the errors cancel out.

There is an equal chance of making an error below or above the true value of y.
(iii) For all $i = 1, \ldots, n$, $\mathbb{V}(\varepsilon_i) = \sigma^2$: Homoscedasticity.
(iv) $\mathbb{C}\mathbf{ov}(\varepsilon_i, \varepsilon_j) = 0$ for $i \neq j$: Independence of errors (also known as uncorrelated errors).
(v) For all $i = 1, \ldots, n$, $\varepsilon_i \sim \mathcal{N}\left(0, \sigma^2\right)$: Normality of errors.

Note
Assumption (i) implies the independence of the random variables Y_i.
However, it is important to note that they do not have the same distribution.

Indeed:

$$\mathbb{E}(Y_i) = \beta_0 + \beta_1 x_i,$$

so the expectation of Y_i varies with i.

Line equation

Part 1: Foundations of Simple Linear Regression

1.4 Ordinary Least Squares Line

$\Rightarrow$ The objective is to fit a line with the equation $y = \beta_0 + \beta_1 x$
within the scatter plot of points (x_i, y_i).
We aim to find the "best" line (i.e., the one that passes closest to the points).

Example
We want to explain the income in tens of thousands of euros of an employee
based on the number of years after obtaining their diploma.

	1	2	3	4	5	6	7	8	9	10
Years	2.00	3.00	6.00	7.00	8.00	....10.00	11.00	12.00	14.00	16.00
Salary	0.20	0.40	0.60	0.80	1.00	1.20	1.40	1.60	1.80	2.00

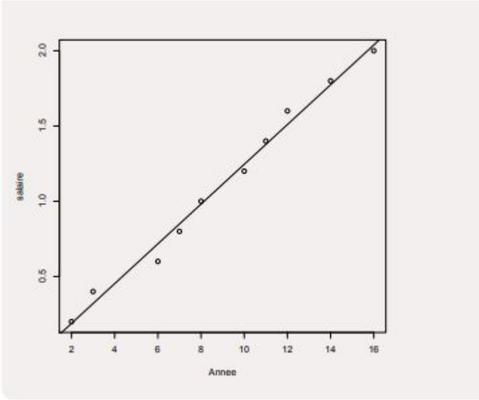
Sum of squared residuals

⇒ We define:

$$F(\beta_0, \beta_1) = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x)^2$$

⇒ ... ... $F(\beta_0, \beta_1)$ is the sum of squared differences
between the true values of the variable Y and the values predicted by the model.

Goal:
Find $\hat{\beta}_0$ and $\hat{\beta}_1$ such that $F\left(\hat{\beta}_0, \hat{\beta}_1\right)$
is the minimum of $F(\beta_0, \beta_1)$.



Partial derivatives

We set the partial derivatives to zero:

$$\frac{\partial F(\beta_0, \beta_1)}{\partial \beta_0} = 0; \frac{\partial F(\beta_0, \beta_1)}{\partial \beta_1} = 0$$

Normal equations:

$$\begin{cases} \sum_{i=1}^{n} x_i Y_i - \beta_1 \sum_{i=1}^{n} x_i^2 - n\beta_0 \bar{x} = 0 \\ \bar{Y} - \beta_1 \bar{x} - \beta_0 = 0 \end{cases}.$$

Theorem:

After solving the normal equations,
we obtain the Ordinary Least Square Estimators (OLS) as the estimates:

$$\begin{cases} \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{S_{xy}}{s_{xx}} \end{cases}.$$

Exercise

Part 1: Foundations of Simple Linear Regression

1.4 Ordinary Least Squares Line

Exercise:
(1) Derive the normal equations.
Solve this system of two equations with two unknowns to obtain the expressions for the OLS estimators.
(2) Show that ... ... :

$$\hat{\beta}_1 = \beta_1 + \frac{1}{ns_{xx}} \sum_{i=1}^{n} (x_i - \bar{x}) \ldots\ldots \varepsilon_i$$

and

$$\hat{\beta}_0 = \beta_0 + \frac{1}{ns_{xx}} \sum_{i=1}^{n} [s_{xx} - (x_i - \bar{x})] \varepsilon_i .$$

Line equation

Note
(i) The line with the equation $y(x) = \hat{\beta}_0 + \hat{\beta}_1 x$ is called the regression line
least squares line, or trend line.

(ii) It necessarily passes through the centroid of the scatter plot $(\bar{x}, \bar{Y})$.
Indeed:

$$y(\bar{x}) = \hat{\beta}_0 + \hat{\beta}_1 \ldots\ldots \bar{x} = \bar{Y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 \bar{x} = \bar{Y}.$$

(iii) All statistical software provides the values of $\hat{\beta}_0$ and $\hat{\beta}_1$.
The calculations above are therefore unnecessary in practice, but necessary for understanding the method.

Interpretation of the slope

Interpretation of the slope $\hat{\beta}_1$

$\hat{\beta}_1$ measures the expected change in Y for a one-unit increase in x.

Indeed, let's consider the predicted value of Y at x:

$$\hat{y}_1 = \hat{\beta}_0 + \hat{\beta}_1 x.$$

After an increase in x by one unit, we have:

$$\hat{y}_2 = \hat{\beta}_0 + \hat{\beta}_1 (x + 1).$$

Therefore:

$$\Delta\hat{y} = \hat{y}_2 - \hat{y}_1 = \hat{\beta}_1.$$

Theorem

Theorem

(i) $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased estimators of β_0 and β_1, respectively. That is,

$$\mathbb{E}\left(\hat{\beta}_0\right) = \beta_0 \text{ and } \mathbb{E}\left(\hat{\beta}_1\right) = \beta_1.$$

(ii) The variances of the OLS estimators are given by:

$$\mathbb{V}\left(\hat{\beta}_1\right) = \frac{\sigma^2}{n s_{xx}} \text{ and } \mathbb{V}\left(\hat{\beta}_0\right) = \frac{\sigma^2}{n}\left(1 + \frac{\bar{x}^2}{s_{xx}}\right).$$

Unbiasedness of OLS estimators

Part 1: Foundations of Simple Linear Regression

1.5 Properties of OLS Estimators

- Unbiasedness of OLS estimators

Note
Only the assumption $\mathbb{E}(\varepsilon_i) = 0$ is necessary to prove the unbiasedness of the OLS estimators.
For the calculation of variances, we use the assumption that ε_i are iid with mean zero and variance σ^2.

Theorem 2

We finally obtain a more general result in the following theorem. Theorem

Under the assumption H_ε, we have:

$$\hat{\beta}_0 \sim \ldots\ldots\ldots \mathcal{N}\left(\beta_0, \frac{\sigma^2}{n}\left[1+\frac{\bar{x}^2}{s_{xx}}\right]\right) \text{ and } \hat{\beta}_1 \sim \ldots\ldots\ldots \mathcal{N}\left(\beta_1, \frac{\sigma^2}{n s_{xx}}\right).$$

Show that

$$\operatorname{Cov}\left(\hat{\beta}_1, \bar{Y}\right) = 0$$

Gauss-Markov theorem

Gauss-Markov Theorem

The OLS estimators of regression are unbiased and consistent.
- Among unbiased estimators, they have the minimum variance,
meaning that it is impossible to find another unbiased estimator with a smaller variance.
⇒ They are called BLUE (Best Linear Unbiased Estimators).
They are considered efficient estimators.

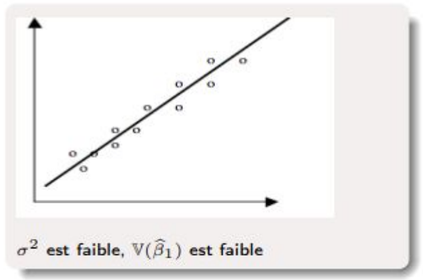
Linear combination

Proof:
$\Rightarrow$ Since $\hat{\beta}_1$ is a linear combination of independent Gaussian random variables,
$\hat{\beta}_1$ follows a normal distribution.
- The same reasoning applies to $\hat{\beta}_0$.

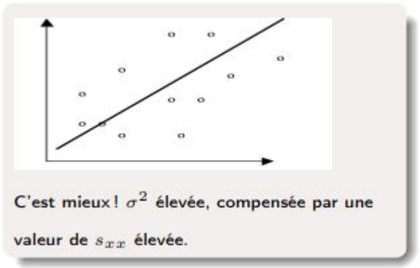
Graphical Illustrations

Note
(i) The larger the sample size, the better the fit (as the variances of the OLS estimators decrease with n).

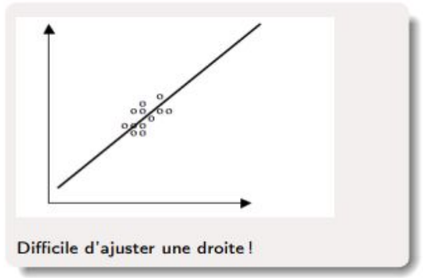
(ii) The larger s_{xx} is, the better the variances:
The experimenter is therefore advised to work with x_i values that are dispersed around $\bar{x}$.



σ^2 est faible, $\mathbb{V}(\widehat{\beta}_1)$ est faible



C'est mieux ! σ^2 élevée, compensée par une valeur de s_{xx} élevée.



Difficile d'ajuster une droite !

Measuring the quality of fit

- The OLS estimators allow us to calculate the estimated response $\widehat{Y}_i$ for each value x_i.
$\Rightarrow$ We compare $\widehat{Y}_i$ to the true value Y_i through the difference:

$$\hat{e}_i = Y_i - \widehat{Y}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i.$$

$\Rightarrow$ The residuals $\hat{e}_i$ ... ...
allow us to measure the quality of the fit and analyze whether the responses are well explained by the model.
- It is precisely based on the sum of squares of these differences that we have constructed $\hat{\beta}_0$ and $\hat{\beta}_1$.

Analysis of variance equation

Part 1: Foundations of Simple Linear Regression

1.7 Residual Analysis

Analysis of Variance Equation:

$\Rightarrow$ We can decompose the quantity $Y_i - \bar{Y}$ (deviation of Y_i from the mean) ... ... :

$$Y_i - \bar{Y} \ldots\ldots\ldots = \left(Y_i - \widehat{Y}_i\right) + \left(\widehat{Y}_i - \bar{Y}\right) = \widehat{Y}_i - \bar{Y} + \hat{e}_i$$

$$= \text{ part of the deviation explained by the model } + \text{ error made by the model.}$$

By squaring and summing, we obtain:

$$\underbrace{\sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2}_{SC_{\text{tot}}} = \ldots\ldots\ldots \underbrace{\sum_{i=1}^{n} \left(\widehat{Y}_i - \bar{Y}\right)^2}_{SC_{\text{reg}}} + 2 \ldots\ldots\ldots \underbrace{\sum_{i=1}^{n} \left(\widehat{Y}_i - \bar{Y}\right)\left(Y_i - \widehat{Y}_i\right)}_{=0 \text{ (to be done in exercise)}}$$

$$+ \underbrace{\sum_{i=1}^{n} \left(Y_i - \widehat{Y}_i\right)^2}_{SC_{\text{res}}}.$$

Decomposition

- The decomposition (3) is a way to measure the relevance of the regression line as a predictor of the response variable Y.

- It allows us to calculate the proportion of the variability of Y_i that is explained by the model, relative to the total variability.
- This proportion is expressed by the coefficient of determination.

Definition:
The coefficient of determination of the model is defined as:

$$R^2 = SC_{\text{reg}} / SC_{\text{tot}}$$

Equalities

Note
We also have the following equalities:

$$R^2 = 1 - \frac{SC_{res}}{SC_{tot}} = \hat{\beta}_1^2 \frac{s_{xx}}{S_{YY}} = r_{xy}^2 \text{ and } F\left(\hat{\beta}_0, \hat{\beta}_1\right) = S_{YY}\left(1 - R^2\right).$$

$\Rightarrow$ If $R^2 = 1$, the model explains everything.
$\Rightarrow$ If $R^2 = 0$, then $\sum_{i=1}^{n} \left(\widehat{Y}_i - \bar{Y}\right)^2 = 0$ meaning that $\widehat{Y}_i = \bar{Y}$ for all i.

The model is therefore inappropriate as it does not model anything better than the mean.
- If R^2 is close to 0, then x does not explain Y well.

Definition

σ^2 plays a very important role as the variances of the estimators depend on it. We seek to estimate it from the data.

Definition:
The residual variance is the estimated variance of the residuals $\hat{e}_i$, denoted $\hat{s}_e^2$.
Thus:

$$\hat{s}_e^2 = \frac{1}{n}\sum_{i=1}^{n}(\hat{e}_i)^2 \dots\dots\dots$$

Using this quantity, we can derive the estimator of σ^2 as:

$$\widehat{\sigma^2} = \frac{n}{n-2}\hat{s}_e^2 = \frac{1}{n-2}\sum_{i=1}^{n}(\hat{e}_i)^2 = \frac{SC_{res}}{n-2}.$$

Theorem

We obtain the following theorem:

Theorem

Under the assumption H_ε, we have:

(i) $\widehat{\sigma^2}$ is an unbiasedestimator of ... ... σ^2, independent of $\left(\bar{Y}, \hat{\beta}_0, \hat{\beta}_1\right)$.

(ii) $\frac{n-2}{\sigma^2}\widehat{\sigma^2} \sim \ldots\ldots\ldots \chi^2_{n-2}$.

(iii) $\frac{\hat{\beta}_0-\beta_0}{\sqrt{\frac{\widehat{\sigma^2}}{n}\left(1+\frac{\bar{x}^2}{s_{xx}}\right)}} \ldots\ldots\ldots \sim \mathcal{T}_{n-2}$ and $\frac{\hat{\beta}_1-\beta_1}{\sqrt{\frac{\widehat{\sigma^2}}{ns_{xx}}}} \sim \mathcal{T}_{n-2}$.

The quantities ... ... $\widehat{\sigma_{\hat{\beta}_0}} = \sqrt{\frac{\widehat{\sigma^2}}{n}\left(1+\frac{\bar{x}^2}{s_{xx}}\right)}$ and $\widehat{\sigma_{\hat{\beta}_1}} = \sqrt{\frac{\widehat{\sigma^2}}{ns_{xx}}}$ are called the standard errors of the OLS estimators.

Proof

Proof
⇒ (i): See tutorial, the independence is assumed.
⇒ (ii): The $\hat{e}_i$ are linear combinations of the ε_i which are Gaussian.

Moreover, it can be shown (see next exercise) that for all $i, \hat{e}_i \sim \mathcal{N}\left(0, \sigma^2\right)$. ... ...

Therefore,

$$\frac{n-2}{\sigma^2}\widehat{\sigma^2} = \sum_{i=1}^{n}\left(\frac{\hat{e}_i}{\sigma}\right)^2 \sim \chi^2_{n-2} \cdot \cdots \cdots$$

⇒ (iii):
We deduce:

$$\frac{\frac{\hat{\beta}_0 - \beta_0}{\sqrt{\frac{\sigma^2}{n}\left(1+\frac{\bar{x}^2}{s_{xx}}\right)}}}{\sqrt{\frac{n-2}{\sigma^2}\widehat{\sigma^2}/(n-2)}} \sim \mathcal{T}_{n-2}$$

Exercise on Empirical Mean

Part 1: Foundations of Simple Linear Regression

Show that unlike the ε_i, the residuals $\hat{e}_i$ are not independent, their empirical mean is zero, and they satisfy

$$\mathbb{E}(\hat{e}_i) = 0, i = 1, \ldots, n.$$

$\Rightarrow$ Reminder: $\hat{e}_i = Y_i - \widehat{Y}_i, i = 1, \ldots, n$ with $\widehat{Y}_i = \bar{Y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i$. Thus: $\sum_{i=1}^{n} \ldots\ldots\ldots \hat{e}_i = 0$
which shows that the $\hat{e}_i$ are not independent, since for example we can write:

$$\hat{e}_n = \sum_{i=1}^{n-1} \hat{e}_i.$$

The knowledge of $\hat{e}_1, \ldots, \hat{e}_{n-1}$ completely determines $\hat{e}_n$.

Results from the previous theorem

According to the result of the previous theorem, for a fixed $\alpha \in (0, 1)$,
if we denote $t_{n-2,1-\alpha/2}$ as the $(1-\alpha/2)$-th quantile of a Student's t-distribution
with $n-2$ degrees of freedom (these distributions are symmetric like the $\mathcal{N}(0,1)$ distribution,
and $t_{n-2,1-\alpha/2} = -t_{n-2,\alpha/2}$),

we have:

$$\begin{aligned}\mathbb{P}\left(-t_{n-2,1-\alpha/2} \leq \mathcal{T}_{n-2} \leq t_{n-2,1-\alpha/2}\right) \\ = 1 - \alpha/2 - [1 - (1 - \alpha/2)]\end{aligned}$$

Thus, we can conclude that:

$$\mathbb{P}\left(-t_{n-2,1-\alpha/2} \leq \frac{\hat{\beta}_0 - \beta_0}{\sqrt{\frac{\widehat{\sigma^2}}{n}\left(1 + \frac{\bar{x}^2}{s_{xx}}\right)}} \leq t_{n-2,1-\alpha/2}\right) = 1 - \alpha$$

Confidence interval

Part 1: Foundations of Simple Linear Regression

1.10 Confidence Interval

Hence, a confidence interval for β_0 at level $1-\alpha$ is given by:

$$\left[\hat{\beta}_0 - t_{n-2,1-\alpha/2}\sqrt{\frac{\widehat{\sigma^2}}{n}\left(1+\frac{\bar{x}^2}{s_{xx}}\right)}, \hat{\beta}_0 + t_{n-2,1-\alpha/2}\sqrt{\frac{\widehat{\sigma^2}}{n}\left(1+\frac{\bar{x}^2}{s_{xx}}\right)}\right].$$

Its width is equal to $2t_{n-2,1-\alpha/2}\sqrt{\frac{\widehat{\sigma^2}}{n}\left(1+\frac{\bar{x}^2}{s_{xx}}\right)}$.

Note:
The confidence interval will be smaller (thus more precise) when s_{xx} is large.

Using the same notations as before, we obtain for the confidence interval of β_1:

$$\left[\hat{\beta}_1 - t_{n-2,1-\alpha/2}\sqrt{\frac{\widehat{\sigma^2}}{ns_{xx}}}, \hat{\beta}_1 + t_{n-2,1-\alpha/2}\sqrt{\frac{\widehat{\sigma^2}}{ns_{xx}}}\right].$$

Percentiles of a distribution

By following the same procedure as for the confidence intervals, we can derive a confidence interval for σ^2 as:

$$\left[\frac{n-2}{c_{n-2,1-\alpha/2}}\widehat{\sigma^2}, \frac{n-2}{c_{n-2,\alpha/2}}\widehat{\sigma^2}\right],$$

where $c_{n-2,1-\alpha/2}$ and $c_{n-2,\alpha/2}$ are the quantiles of a χ^2_{n-2} distribution.

Order of a Fisher-Snedecor distribution

It is possible to construct a simultaneous confidence region for the pair (β_0, β_1) at level $1 - \alpha$.

$$R(\beta_0, \beta_1) = \left\{ (\beta_0, \beta_1) : \frac{1}{\widehat{\sigma^2}} \left[n\left(\hat{\beta}_0 - \beta_0\right)^2 + 2n\bar{x}\left(\hat{\beta}_0 - \beta_0\right)\left(\hat{\beta}_1 - \beta_1\right) + \left(\hat{\beta}_1 - \beta_1\right)^2 \sum_{i=1}^{n} x_i^2 \right] \leq f_{(1,n-2),1-\alpha} \right\}$$

where $f_{(1,n-2),1-\alpha}$ is the quantile of a Fisher-Snedecor $\mathcal{F}_{1,n-2}$ distribution.

Significance

Part 1: Foundations of Simple Linear Regression

1.13 Hypothesis Testing / Model

We aim to perform the following tests:

$$H_0 : \beta_j = b_j \text{ against } H_1 : \beta_j \neq b_j, j = 0, 1.$$

In particular, for $b_j = 0$, we conduct the test of model significance (measuring the impact of X in explaining Y through the model).

Decision Rule:
The test statistics are given by:

$$t_0 = \frac{\hat{\beta}_0 - b_0}{\sqrt{\frac{\widehat{\sigma^2}}{n}\left(1 + \frac{\bar{x}^2}{s_{xx}}\right)}} \text{ or } t_1 = \frac{\hat{\beta}_1 - b_1}{\sqrt{\frac{\widehat{\sigma^2}}{n s_{xx}}}}$$

Reject H_0 if $|t_j| > t_{n-2,1-\alpha/2}$ and do not reject H_0 otherwise.

Decision rule

We aim to perform the following test:

$$H_0 : \beta_1 = 0 \text{ against } H_1 : \beta_1 \neq 0.$$

The model does not contribute to the explanation of ... ... Y against The model is globally significant.

Decision Rule:

$$F = \frac{(n-2)R^2}{1-R^2}$$

$\Rightarrow$ Reject H_0 if $F > f_{(1,n-2),1-\alpha}$, we conclude that the model is globally significant.
$\Rightarrow$ Do not reject H_0 if $F \leq f_{(1,n-2),1-\alpha}$.

Note
The test for the significance of the regression is equivalent to the test for the significance of the slope in simple regression. Indeed:

$$F = t_1^2$$

DROITEREG()

Part 1: Foundations of Simple Linear Regression

1.14 Overall Model Significance Test

$\hat{\beta}_1 = \frac{S_{xY}}{s_{xx}}$	$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$
$\widehat{\sigma_{\hat{\beta}_1}} = \frac{\hat{\sigma}}{\sqrt{n s_{xx}}}$	$\widehat{\sigma_{\hat{\beta}_0}} = \hat{\sigma}\sqrt{\frac{1}{n}\left(1 + \frac{\bar{x}^2}{s_{xx}}\right)}$
$R^2 = \frac{SC_{reg}}{SC_{tot}}$	$\hat{\sigma} = \sqrt{\frac{SC_{res}}{n-2}}$
$F = \frac{(n-2)R^2}{1-R^2}$	$n-2$
SC_{reg}	SC_{res}

Predicting the value of Y

We are given a known value x^* of the variable x (for example, $x^* = x_{n+1}$) and we aim to predict the value of Y.

Let ε^* be the associated error term:

$$\mathbb{E}(\varepsilon^*) = 0, \mathbb{V}(\varepsilon^*) = \sigma^2, \text{ and } \text{Cov}(\varepsilon^*, \varepsilon_i) = 0, i = 1, \dots, n.$$

We have: $\widehat{Y^*} = \hat{\beta}_0 + \hat{\beta}_1 x^*$ (Note that this is a random variable) ... ... $\widehat{Y^*}$ is an unbiased predictor for $Y^* = \beta_0 + \beta_1 x^* + \varepsilon^*$.

It is a random variable with an expectation of $y^* = \mathbb{E}(Y^*) = \beta_0 + \beta_1 x^*$.
(Note: we could also aim to predict y^* instead of Y^*)

Prediction error

Part 1: Foundations of Simple Linear Regression

1.15 Prediction

- The prediction error is defined as :

$$\widehat{e^*} = Y^* - \widehat{Y^*} = \beta_0 - \hat{\beta}_0 + \left(\beta_1 - \hat{\beta}_1\right) x^* + \varepsilon^*.$$

- The prediction error is a centered random variable, and its variance is given by

$$\mathbb{V}\left(\widehat{e^*}\right) = \sigma^2 \left[1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{n s_{xx}}\right]$$

Estimated variance of prediction error

By replacing σ^2 with its estimator, we obtain the estimated variance of the prediction error:

Estimated Variance of Prediction Error:

$$\widehat{\mathbb{V}\left(\widehat{e^*}\right)} = \widehat{\sigma^2}\left[1 + \frac{1}{n} + \frac{\left(x^* - \bar{x}\right)^2}{ns_{x.....x}}\right].$$

Note:
The quantity

$$h^* = \frac{1}{n} + \frac{\left(x^* - \bar{x}\right)^2}{ns_{xx}}$$

is called the leverage of the observation x^*. It plays an important role in the study of outliers.

Estimated variance of prediction error

The estimated variance of the prediction error is smaller when:
$\Rightarrow n$ is large (the model is built with a large number of observations)
$\Rightarrow s_{xx}$ is large (the scatter of the data points is large compared to the mean)
$\Rightarrow \widehat{\sigma^2}$ is small (the line fits the data points well)
$\Rightarrow (x^* - \bar{x})^2$ is small (the prediction is made at a point close to the center of gravity of the data).

Precision Interval

Part 1: Foundations of Simple Linear Regression

1.16 Precision Interval

Finally, we obtain a more general result in the following theorem.

Theorem:
Under the assumption H_ε, we have:

$$\widehat{e^*} = \widehat{Y^*} - Y^* \sim \mathcal{N}\left(0, \sigma^2(1+h^*)\right)$$

and

$$\frac{\widehat{Y^*} - Y^*}{\sqrt{\widehat{\sigma^2}(1+h^*)}} \sim \mathcal{T}_{n-2}.$$

Thus, a prediction interval for Y^* is given by:

$$\left[\widehat{Y^*} - t_{n-2,1-\alpha/2}\sqrt{\widehat{\sigma^2}(1+h^*)}, \widehat{Y^*} + t_{n-2,1-\alpha/2}\sqrt{\widehat{\sigma^2}(1+h^*)}\right].$$

KNOWLEDGE TREE

Part 1: Foundations of Simple Linear Regression

1.1 Introduction to Linear Regression

- Definition of Linear Regression
- Common Applications in Statistics

1.2 Dependent and Independent Variables

- Identifying Variables in the Context of Linear Regression
- Concept of Dependent and Independent Variables

1.3 Simple Linear Regression Model

- Basic Equation of Simple Linear Regression
- Interpretation of Coefficients

1.4 Assumptions of Linear Regression

- Linearity, Independence, Homoscedasticity, and Residual Normality
- Diagnosing Assumption Violations

Part 2: Estimation Methods

2.1 Ordinary Least Squares (OLS) Method

- Principle of Minimizing the Sum of Residual Squares
- Calculating Regression Coefficients

2.2 Interpretation of Coefficients

- Slope Coefficient (Regression Slope)
- Intercept Coefficient (Regression Intercept)

2.3 Standard Error of the Estimate (SEE)

- Computing SEE
- Using SEE to Assess Model Fit

Part 3: Evaluation of Simple Linear Regression

3.1 Coefficient of Determination (R-squared)

- Definition and Interpretation of R-squared
- Limitations of R-squared

3.2 Hypotheses and Statistical Tests

- Global Model Significance Test
- Individual Coefficient Significance Test

3.3 Cross-Validation and Overfitting

- Cross-Validation Techniques
- Overfitting Prevention

Part 4: Diagnostics and Visualization

4.1 Scatterplot

- Using Scatterplots to Visualize the Relationship
- Identifying Outliers

4.2 Residual Analysis

- Histograms and QQ-plots of Residuals
- Checking for Normality

4.3 Influence and Leverage

- Identifying Influential Observations
- Impact of Leverage on Regression

KEYWORDS (NEW)

Exogénéité

Combinaison linéaire

Estimateur des moindres carrés ordinaires (EMCO)

Coefficient de détermination (R^2)

Analyse de régression

Hétéroscédasticité

Combinaison linéaire de coefficients

Régression linéaire

Régression linéaire généralisée

Régression simple

Résidus de déviance

Résidus de Pearson

Estimation par régression linéaire

Résidus (modèle de régression)

Use Case with python

Use Case: Predicting Student Performance in Exams

In the context of the course on Simple Linear Regression, which covers topics related to understanding linear regression, estimation methods, model evaluation, diagnostics, and visualization, let's explore a use case related to predicting student performance in exams based on study hours.

Description:

In this use case, we will apply simple linear regression techniques to predict a student's exam score based on the number of hours they studied. Understanding the relationship between study hours and exam scores is important for educators and students to optimize study strategies.

Key Components:

Introduction to Linear Regression: Understanding the fundamentals of linear regression, its definition, and common applications in statistics.
Dependent and Independent Variables: Identifying variables in the context of linear regression, such as the dependent variable (exam score) and independent variable (study hours).
Simple Linear Regression Model: Formulating the simple linear regression equation to model the relationship between study hours and exam scores. Interpreting the coefficients, including the slope and intercept.
Assumptions of Linear Regression: Understanding the assumptions of linear regression, including linearity, independence, homoscedasticity, and residual normality. Diagnosing violations of these assumptions.
Ordinary Least Squares (OLS) Method: Applying the OLS method to estimate regression coefficients by minimizing the sum of squared residuals. Calculating and interpreting regression coefficients.
Standard Error of the Estimate (SEE): Computing the SEE to assess model fit and quantify the dispersion of data points around the regression line.
Coefficient of Determination (R-squared): Calculating and interpreting R-squared as a measure of the proportion of variation in exam scores explained by study hours.
Hypotheses and Statistical Tests: Performing statistical tests to assess the significance of the regression model as a whole and the individual coefficients.
Cross-Validation and Overfitting: Using cross-validation techniques to evaluate the model's performance and prevent overfitting.
Diagnostics and Visualization: Visualizing the relationship between study hours and exam scores using scatterplots. Analyzing residuals through histograms and QQ-plots to check for normality. Identifying influential observations and assessing leverage.

Python Code Example (Student Performance Prediction):

```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error, r2_score

# Load student performance dataset (features: StudyHours; target: ExamScore)
data = pd.read_csv('student_performance.csv')

# Split the dataset into training and testing sets
X = data[['StudyHours']]
y = data['ExamScore']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=42)

# Fit a simple linear regression model
model = LinearRegression()
model.fit(X_train, y_train)

# Predict exam scores on the test set
y_pred = model.predict(X_test)

# Evaluate the model
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)

print(f'Mean Squared Error: {mse:.2f}')
print(f'R-squared: {r2:.2f}')

# Visualize the regression line and scatterplot
plt.scatter(X_test, y_test, label='Actual Data')
plt.plot(X_test, y_pred, color='red', linewidth=2, label='Regression Line')
plt.xlabel('Study Hours')
plt.ylabel('Exam Score')
plt.legend()
plt.show()
```

REFERENCES V2

TOP 20 REFERENCES

- Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. Wiley.
- Seber, G. A., & Lee, A. J. (2012). Linear Regression Analysis. Wiley.
- Hocking, R. R. (2003). Methods and Applications of Linear Models: Regression and the Analysis of Variance. Wiley.
- Agresti, A., & Finlay, B. (2009). Statistical Methods for the Social Sciences. Pearson.
- Fox, J., & Monette, G. (1992). Generalized Collinearity Diagnostics. Journal of the American Statistical Association.
- Ryan, T. P. (1997). Modern Regression Methods. Wiley.
- Myers, R. H., Montgomery, D. C., & Anderson-Cook, C. M. (2016). Response Surface Methodology: Process and Product Optimization Using Designed Experiments. Wiley.
- Wu, C. F. J., & Hamada, M. (2009). Experiments: Planning, Analysis, and Optimization. Wiley.
- Snee, R. D. (1977). Validation of Regression Models: Methods and Examples. Technometrics.
- Christensen, R. (2011). Plane Answers to Complex Questions: The Theory of Linear Models. Springer.
- Allen, M. P. (1997). Understanding Regression Analysis. Springer.
- Judge, G. G., Hill, R. C., Griffiths, W. E., Lütkepohl, H., & Lee, T. (1988). Introduction to the Theory and Practice of Econometrics. Wiley.
- Berry, W. D., & Feldman, S. (1985). Multiple Regression in Practice. Sage Publications.
- Kleinbaum, D. G., & Klein, M. (2010). Logistic Regression: A Self-Learning Text. Springer.
- Mason, R. L., Gunst, R. F., & Hess, J. L. (2003). Statistical Design and Analysis of Experiments. Wiley.
- Atkinson, A. C., & Donev, A. N. (1992). Optimum Experimental Designs. Oxford University Press.
- Box, G. E. P., & Draper, N. R. (1987). Empirical Model-Building and Response Surfaces. Wiley.
- Maronna, R. A., Martin, R. D., & Yohai, V. J. (2006). Robust Statistics: Theory and Methods. Wiley.
- Chatterjee, S., & Price, B. (1991). Regression Analysis by Example. Wiley.
- Greene, W. H. (2003). Econometric Analysis. Prentice Hall.

TEXTE DE DESCRIPTION DU COURS

Embark on a fascinating journey into the heart of Simple Linear Regression, a foundational statistical concept that forms the bedrock of data analysis. This journey begins with a captivating narrative that invites you to explore the world of data-driven decision-making. Imagine you are a data enthusiast, eager to unravel the mysteries hidden within datasets, seeking the key to making informed predictions and extracting meaningful insights. Simple Linear Regression promises to be your guiding light in this quest. It's not just about crunching numbers; it's about understanding the intricacies of relationships between variables, discovering patterns, and harnessing the power of data to drive informed choices.

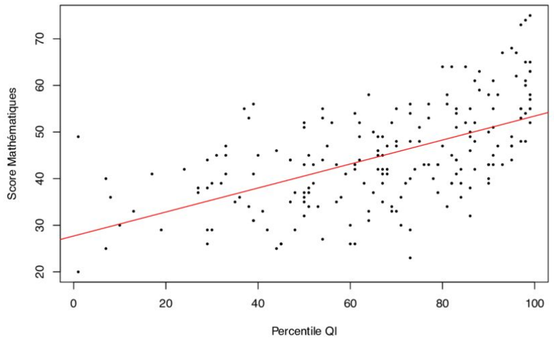
To fully appreciate the significance of Simple Linear Regression, it's essential to delve into its historical context and the influential figures who shaped its development. Picture renowned statisticians like Sir Francis Galton and Karl Pearson, whose pioneering work laid the groundwork for this statistical technique. Their contributions have endured through time, underscoring the enduring relevance of Simple Linear Regression. As you embark on your journey, you'll stand on the shoulders of these giants, drawing inspiration from their groundbreaking insights to navigate the complexities of data analysis.

The practicality of Simple Linear Regression extends across a wide spectrum of industries and professions. Consider the retail executive eager to forecast product sales based on advertising expenditure or the scientist exploring the correlation between environmental factors and species abundance. Business analysts leverage it to optimize marketing budgets, while researchers employ it to examine cause-and-effect relationships in diverse fields. Whether you're a data scientist, economist, marketer, or simply someone with a curiosity for understanding the world through data, this course offers invaluable skills that transcend disciplinary boundaries.

In an era where data is the lifeblood of organizations and informs critical decision-making, mastering Simple Linear Regression is not just advantageous; it's essential. As you progress through this course, you'll gain a deep understanding of how data can be harnessed to make predictions, identify trends, and inform strategies. The course empowers you to transform raw data into actionable insights, equipping you with the tools to make data-driven decisions that can steer your career towards new heights. By mastering Simple Linear Regression, you become a proficient data storyteller, capable of translating complex information into narratives that drive meaningful change.

Now that you've embarked on this enlightening journey into the world of Simple Linear Regression, it's time to take action. Enroll in our course to unlock the full potential of this statistical tool and elevate your data analysis skills. Whether you're driven by personal curiosity, professional growth, or the desire to make a positive impact in your field, this course offers the knowledge and expertise you need. Join our community of learners and begin your exploration of the data-driven future today. Feel free to reach out to the Weeki team for further guidance or any inquiries you may have. Together, we'll pave the way for a future illuminated by data-driven insights and informed decision-making.

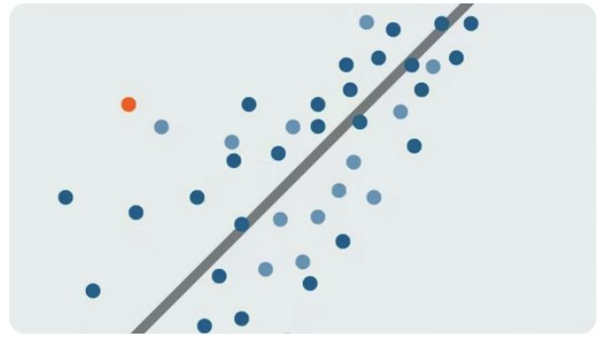
DESIGN DE LA CARD



Author: Baptiste Mokas, Weeki

Course Name: Simple Linear Regression

#RegressionAnalysis #DataModeling #StatisticalInference
#GeneralLinearModel



Duke University

Linear Regression and Modeling

Compétences que vous acquerrez: Probability & Statistics, Regression, Business Analysis, Data Analysis, General Statistics, Statistical Analysis,…

4.8 (1.7k avis)

Débutant · Course · 1 à 4 semaines