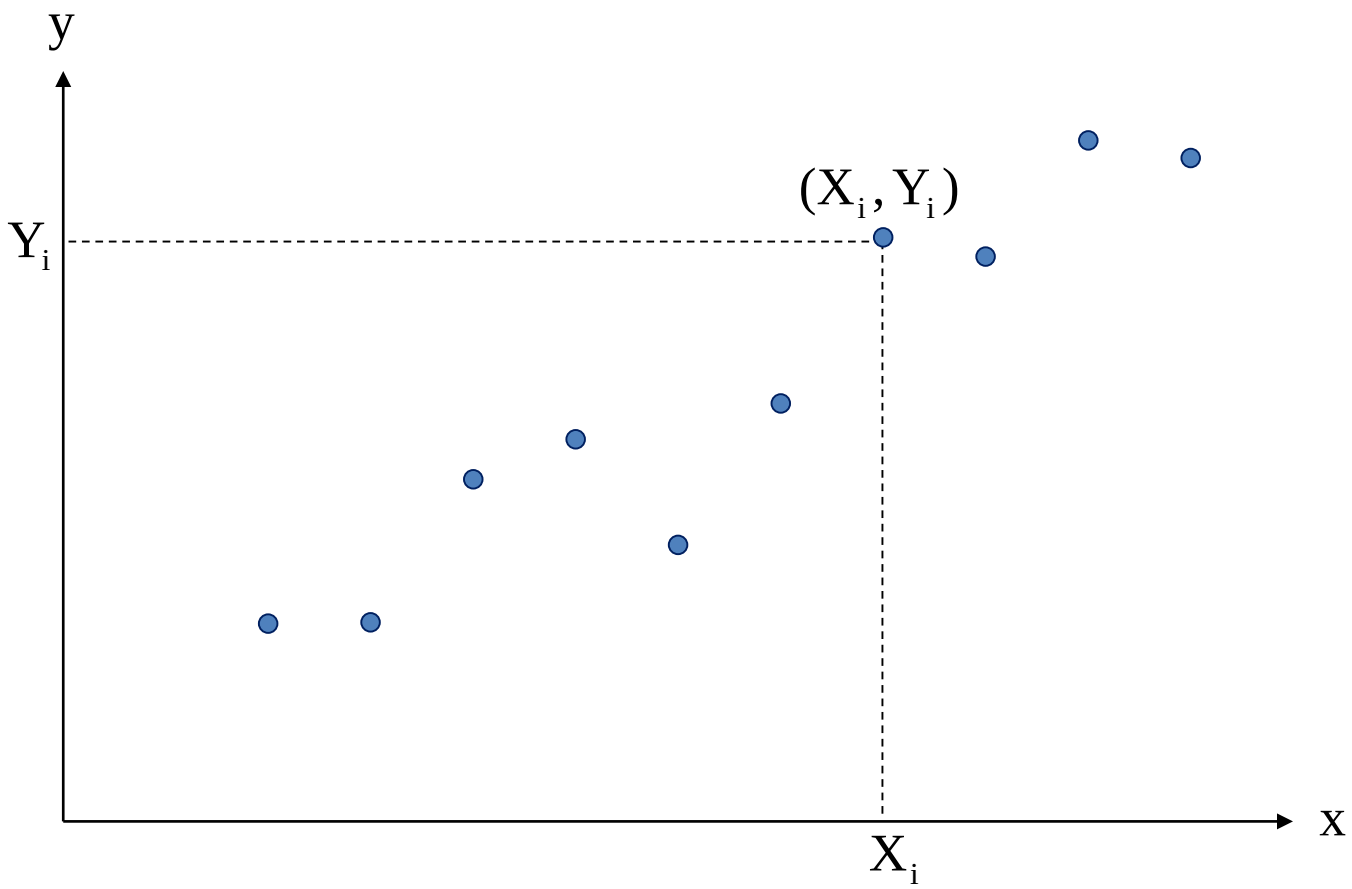
Simple Linear Regression

Simple Linear Regression

Y is the r.v. we are interested in studying and preticting.

Sample of n observed pairs

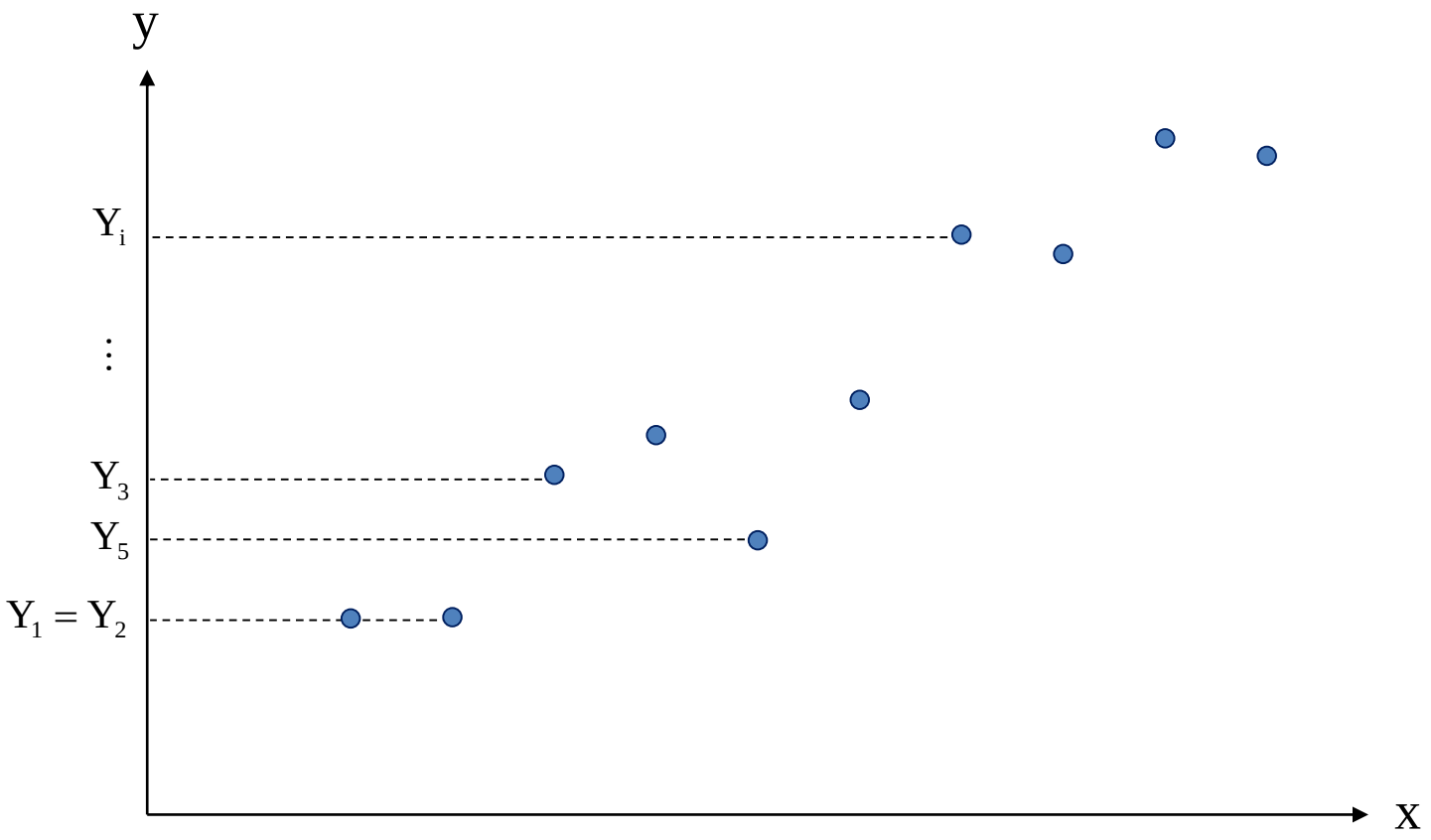
(X_1, Y_1)
(X_2, Y_2)
$\vdots$
(X_i, Y_i)
$\vdots$
(X_n, Y_n)



Simple Linear Regression

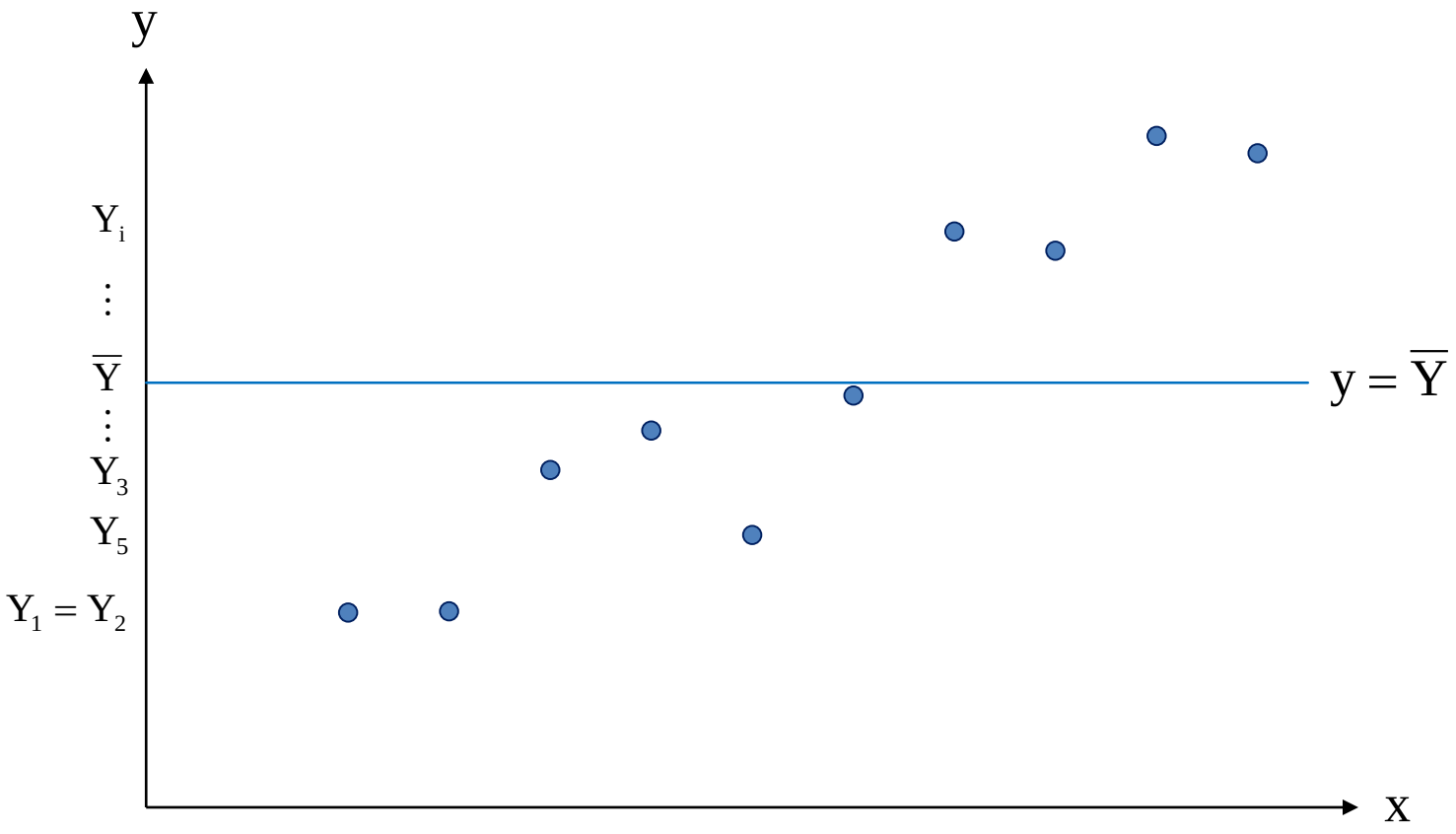
It is of the most importance to estimate the expected value of Y, μ, as well as its variance, σ^2, a measure of the variability around μ.

Let's think, by now, about the Y's observations only.



Simple Linear Regression

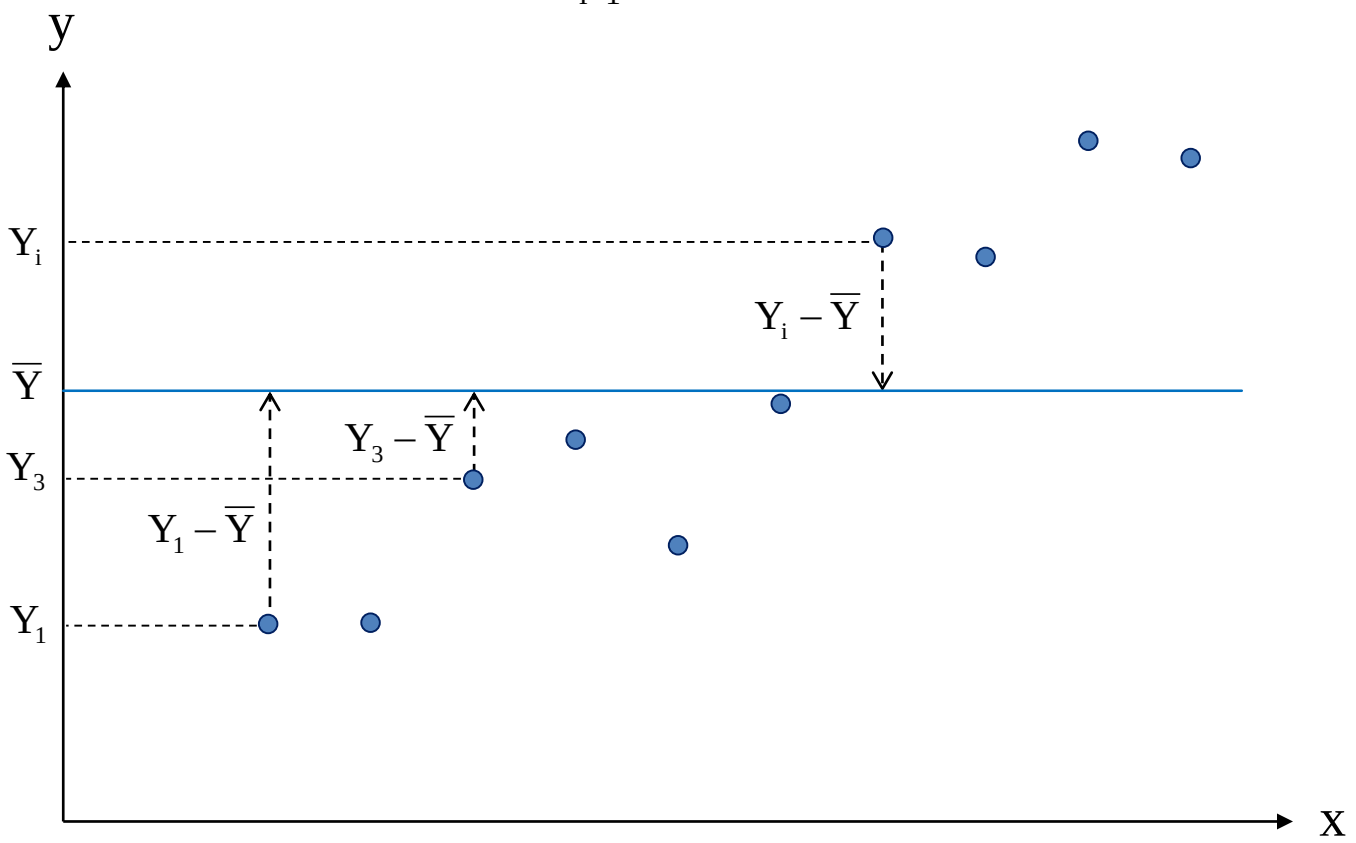
As already known, the best estimate of μ is the sample mean $\overline{Y}$



Simple Linear Regression

The usual way of estimating Y's variability is, then, based on:

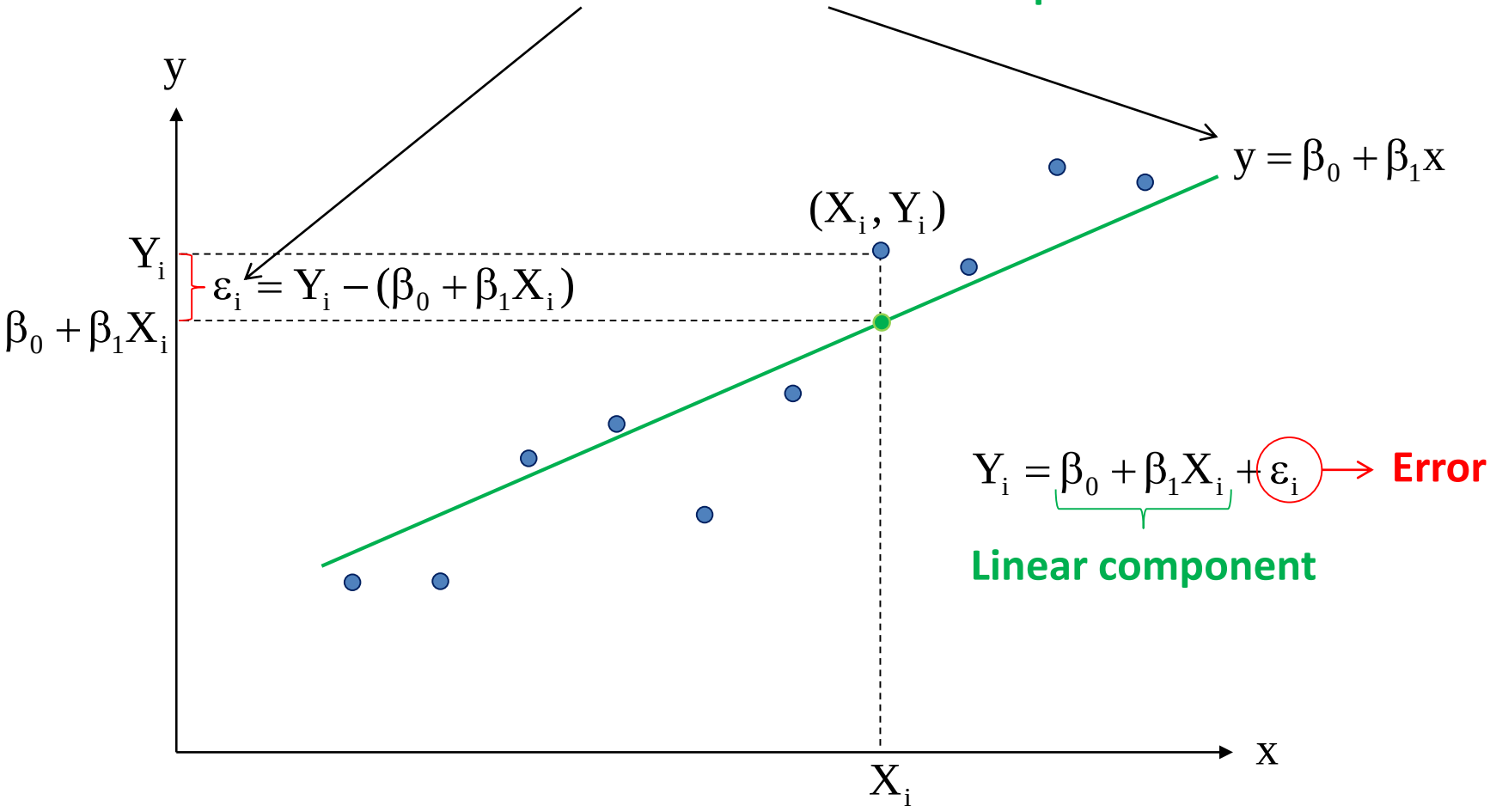
$$\sum_{i=1}^{n} (Y_i - \overline{Y})^2$$



Simple Linear Regression

Let's look again at the scatter plot.

It looks like that exists a **more or less linear relationship** between Y and X.



Simple Linear Regression Model

We can infer that exists a structural relationship between Y and X of the form

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

where:

Y is the dependent (on X), or response, random variable;

X is the independent (of Y and ε), or explanatory, controlled variable;

ε is the random error variable;

β_0 is a parameter, the intercept of the linear relationship component;

β_1 is a parameter, the slope of the linear relationship component.

$(X_1, Y_1), \ldots, (X_n, Y_n)$ are pairs of observations of the model – a sample of n pairs,

thus, all the pairs verify the equation: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, \ldots, n$

Simple Linear Regression Model – Assumptions

Since the r.v.'s ε_1, ε_2,…, ε_n, represent **errors**, it is natural that their **expected values** are all equal to **zero** and, thus, we **assume** that

(1) $$E(\varepsilon_i) = 0, \quad i = 1, \ldots, n$$

It is also natural the **errors** have all the **same variance** and, thus, we **assume** that

(2) $$Var(\varepsilon_i) = \sigma^2, \quad i = 1, \ldots, n$$

Finally, the **errors** must be **statistically not related** among each other (otherwise the linear component would "suffer" from lack of fitness in predicting the dependent variable), what leads to the **assumption** that

(3) $$Cov(\varepsilon_i, \varepsilon_j) = 0, \quad i, j = 1, \ldots, n; i \neq j$$

Simple Linear Regression Model – Moments

In face of the previous assumptions, the model is such that, when X is known:

$$E(Y) = E(\underbrace{\beta_0 + \beta_1 X}_{\text{CONSTANT}} + \varepsilon) = \beta_0 + \beta_1 X + E(\varepsilon) = \beta_0 + \beta_1 X = \mu$$

$$Var(Y) = Var(\underbrace{\beta_0 + \beta_1 X}_{\text{CONSTANT}} + \varepsilon) = Var(\varepsilon) = \sigma^2$$

The expected value of Y is a linear function of X (μ is totally described by X).

$\mu(x) = y = \beta_0 + \beta_1 x$ ⟶ The straight line fitting the points in the scatter plot corresponds to the expected value of Y

The variance of Y is exactly the same as the variance of the error.

$Var(Y) = Var(\varepsilon) = \sigma^2$

Simple Linear Regression Model – Estimation

If the linear model hold, we know that the expected value of Y, μ, varies according to the corresponding value of X, and if we knew the true slope, β_1, and intercept, β_0, of the straight line relating Y and X we could immediately find μ. Also, we need to know σ^2 in order to make good predictions for Y.

But… the **real β_0, β_1 and σ^2 are unknown**!...

What can be done?

Use our sample of data pairs to obtain the best estimates of β_0 and β_1 and σ^2.

How?

Simple Linear Regression Model – Estimation

Suppose that b_0 and b_1 are **estimates** of β_0 and β_1, respectively.

Using b_0 and b_1 we can estimate/predict the expected value of each Y_i, also called **predicted value** of Y_i, based on the correspondent X_i (using the straight line fitted to the set of pairs):

$$\hat{Y}_i = b_0 + b_1 X_i \quad , i = 1, \ldots, n$$

Based on each predicted value Y_i we can, then, predict the correspondent error computing the so called **residuals**:

$$e_i = \hat{\varepsilon}_i = Y_i - \hat{Y}_i = Y_i - (b_0 + b_1 X_i) \quad , i = 1, \ldots, n$$

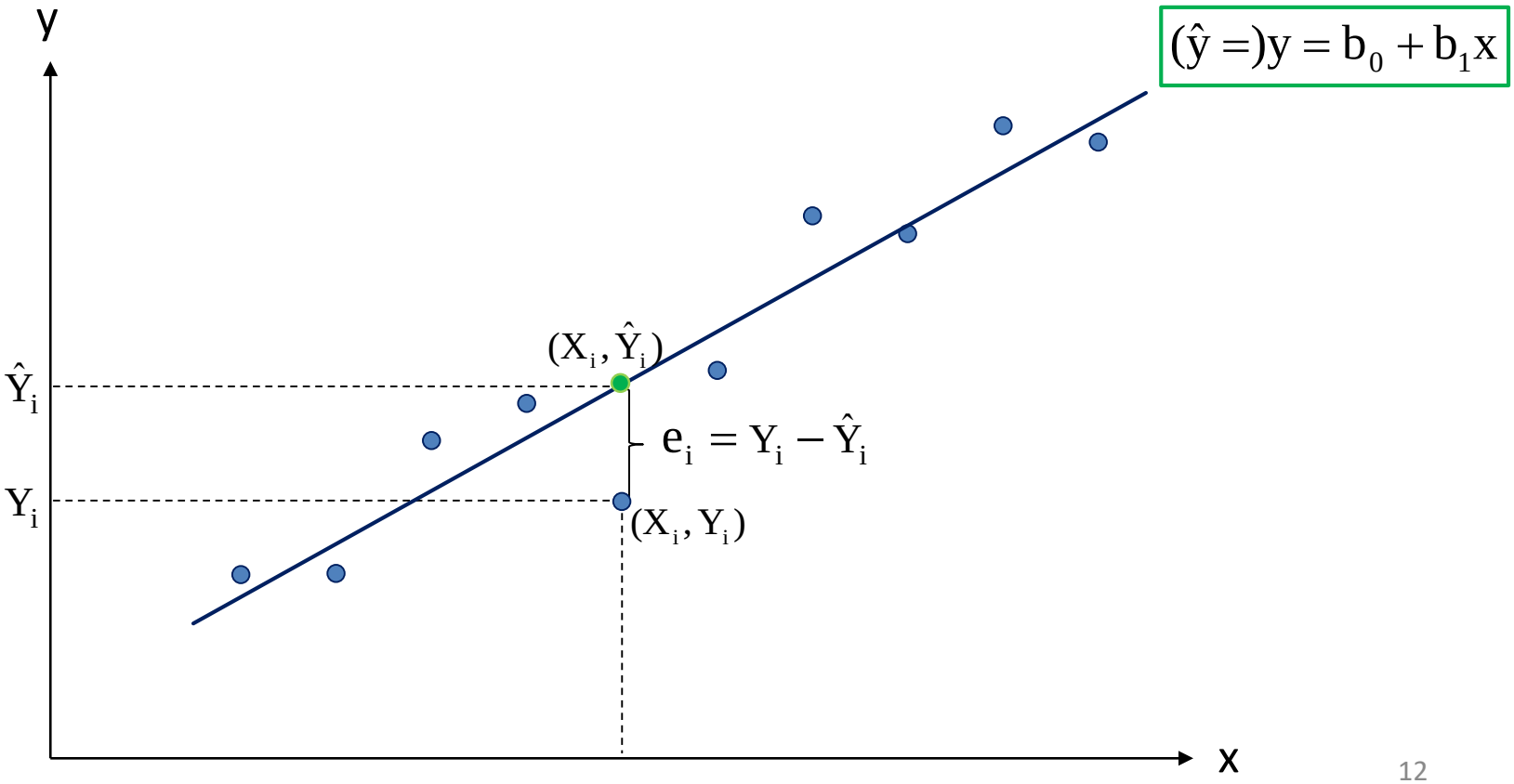
Recalling that $E(\varepsilon)=0$ (and then $Var(\varepsilon)=E(\varepsilon^2)$) we can use the set of n residuals to estimate σ^2, the common variance of ε and Y, based (as usual) on the **Residual Sum of Squares** (also called **Sum of Squares due to Error**):

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2$$

Simple Linear Regression Model – Least Squares Estimation

The smaller the variance of Y, the more accurate the prediction of Y based only on its expected value. As we already know that the best way of estimating σ^2 is based on SSE, we will say that:

The best estimates of β_0 and β_1 are those b_0 and b_1, respectively, that minimize the residual sum of squares, SSE – the **Least Squares (LS) estimates**.



Expressions for b_0 and b_1

$$b_0 = \overline{Y} - b_1\overline{X}$$

$$b_1 = \frac{S_{XY}}{S_{XX}} = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sum (X_i - \overline{X})^2} = \frac{\sum X_i Y_i - n\,\overline{X}\,\overline{Y}}{\sum X_i^2 - n\,\overline{X}^2} = \frac{\sum X_i Y_i - \frac{\left(\sum X_i\right)\left(\sum Y_i\right)}{n}}{\sum X_i^2 - \frac{\left(\sum X_i\right)^2}{n}}$$

Three important properties

$$\overline{\hat{Y}} = \frac{1}{n}\sum_{i=1}^{n} \hat{Y}_i = \overline{Y}$$

The mean of the predicted values is equal to the mean of the observed values.

$$\overline{e} = \frac{1}{n}\sum_{i=1}^{n} e_i = 0$$

The mean of the residuals is equal to zero.

$$(\overline{X}, \overline{Y})$$

The pair of sample means belongs to the LS straight line.

Analyzing the Model – Measures of Variation

Usually, a r.v. variability around its unknown expected value is expressed in terms of the sample variance. In what concerns de dependent variable Y, we have:

$$S_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 \quad \left(= \frac{S_{YY}}{n-1}\right)$$

S_Y^2 is expressed in terms of what is called the **Total Sum of Squares – STT**:

$$SST = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 \quad (= S_{YY})$$

It turns out that SST can be factorized as:

$$\underbrace{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}_{SST} = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2 + \underbrace{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}_{SSE}$$

Analyzing the Model – Measures of Variation

Thus, the total variability of the Y observations around its mean is the sum of two factors, one corresponding to the SSE and another which expression is

$$SSR = \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2$$

Since the mean of the predicted values is equal to the mean of the observed values, SSR measures the variability of the predicted values around its mean. On the other hand, the predicted values are points on the regression line and, thus, SSR is also referred to as the **Sum of Squares due to Regression**. Also, if all the predicted values were exactly equal to the correspondent observed values, then SSR would measure the variability of the Y observations around its mean. For this reason it is said that SSR measures the part of the total variability of the Y observations that is due to the regression.

$$\underbrace{\sum_{i=1}^{n} (Y_i - \overline{Y})^2}_{SST} = \underbrace{\sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2}_{SSR} + \underbrace{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}_{SSE}$$

Analyzing the Model – Measures of Variation

In conclusion, the Total Sum of Squares (of deviations of observed values from its mean) is decomposed into one factor measuring the sum of squares due to the regression (squares of deviations of predicted values from its mean) and another factor measuring the residuals sum of squares (of deviations between observed and predicted values). This is called the **partition of the total sum of squares**.

$$\underbrace{\sum_{i=1}^{n}(Y_i-\overline{Y})^2}_{SST}=\underbrace{\sum_{i=1}^{n}(\hat{Y}_i-\overline{Y})^2}_{SSR}+\underbrace{\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2}_{SSE}$$

The closer SSR is to SST, or equivalently the smaller SSE is, the best is the regression line in explaining the dispersion of the observed pairs distributed around it.

Testing the Significance of the Model – ANOVA table

The partition of the total sum of squares is very useful, starting with the fact that it allows us to test the significance of the model. To do this, it is usual to construct an **Analysis of Variance** (**ANOVA**) table as follows:

ANOVA

Source	*df*	*SS*	*MS*	*F*
Regression	1	$SSR = \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2$	$MSR = \frac{SSR}{1}$	$\frac{MSR}{MSE}$
Residual	$n-2$	$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$	$MSE = \frac{SSE}{n-2}$	
Total	$n-1$	$SST = \sum_{i=1}^{n} (Y_i - \overline{Y})^2$		

Note: *df* stands for “degrees of freedom”.

Testing the Significance of the Model – ANOVA table

Sometimes is useful to construct the ANOVA table in terms of S_{XX}, S_{YY} and b_1:

ANOVA

Source	*df*	*SS*	*MS*	*F*
Regression	1	$b_1^2 S_{XX} = (n-1) b_1^2 S_X^2$	$MSR = \frac{SSR}{1}$	$\frac{MSR}{MSE}$
Residual	$n-2$	$S_{YY} - b_1^2 S_{XX} = = (n-1)(S_Y^2 - b_1^2 S_X^2)$	$MSE = \frac{SSE}{n-2}$	
Total	$n-1$	$S_{YY} = (n-1) S_Y^2$		

Testing the Significance of the Model – ANOVA table

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Testing the significance of the model is equivalent to test if the slope, β_1, is significantly different from zero, that is, testing the hypothesis:

$$H_0 : \beta_1 = 0 \quad vs \quad H_1 : \beta_1 \neq 0$$

The previous hypothesis are tested comparing the value of the **F statistic** with the critical value of the F distribution with 1 *df* in the numerator and n–2 *df* in the denominator, $F_{0.05;1,n-2}$. Large values of F signify that MSR is larger than MSE and, thus, the model shall be significant.

Decision:

$F > F_{0.05;1,n-2} \Rightarrow$ Reject H_0: **model is SIGNIFICANT**

$F < F_{0.05;1,n-2} \Rightarrow$ Do not reject H_0: **model is NOT SIGNIFICANT**

Significant model – How "good" is the model?

Even if a linear model is significant, its "quality" may be poor, in the sense that SSR may "fail" to explain the biggest part of SST (or, equivalently, SSE may be too large in comparison with SSR.

One measure of the quality of the model is the **coefficient of determination**, R^2:

$$R^2 = \frac{SSR}{SST} = \frac{\sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2}{\sum_{i=1}^{n} (Y_i - \overline{Y})^2} \quad \left(= b_1^2 \frac{S_{XX}}{S_{YY}} = b_1^2 \frac{S_X^2}{S_Y^2} \right)$$

R^2 indicates the proportion of the variance in the dependent (response) variable that is predicted by the independent (explanatory) variable.

$0 \le R^2 \le 1$

- The model is as "weak" (bad) as closer to 0 is R^2 .
- The model is as "strong" (good) as closer to 1 is R^2.

Coefficient of determination *vs* Sample Correlation Coefficient

Recall that the correlation coefficient measures the "strength" (along with the direction) of the linear relation between two r.v.'s X and Y:

$$\rho_{XY} = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\mathrm{var}(Y)}}$$

The **sample correlation coefficient** for data sets $(X_1, \ldots, X_n)$ and $(Y_1, \ldots, Y_n)$, is obtained by substituting the sample covariance and variances into the formula above, to get:

$$r_{XY} = \frac{C_{XY}}{\sqrt{S_X^2 S_y^2}} = \frac{\frac{1}{n-1}\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\frac{1}{n-1}\sum (X_i - \overline{X})^2 \frac{1}{n-1}\sum (Y_i - \overline{Y})^2}} =$$

$$= \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum (X_i - \overline{X})^2 \sum (Y_i - \overline{Y})^2}} = \frac{S_{XY}}{\sqrt{S_{XX} S_{YY}}} = b_1 \sqrt{\frac{S_{XX}}{S_{YY}}} = b_1 \frac{S_X}{S_Y}$$

Coefficient of determination *vs* Sample Correlation Coefficient

The interpretation of r_{XY} is absolutely analogous to that for ρ_{XY}:

- The closer to 1 (resp. -1) r_{XY} is, the stronger the linear relation between the two data sets, and in the same (resp. opposite) direction.

- The closer to 0 r_{XY} is, the weaker the linear relation between the two data sets.

Therefore, about the **linear relation between** the two data sets:

- $\mathbf{r_{XY}}$ directly gives the **strength** and **direction** of the relation, but not the proportion of the dependent variable variability explained by the independent variable.

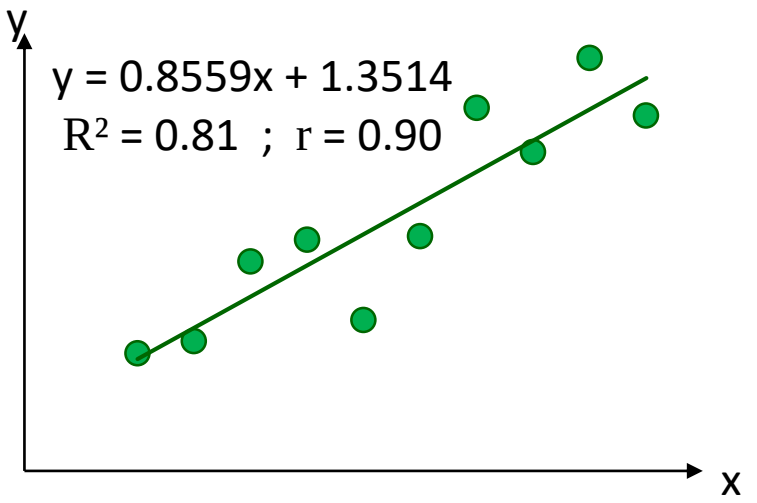
- $\mathbf{R^2}$ directly gives **the proportion of the dependent variable variability explained by the independent variable**, but not the strength and direction of the relation.

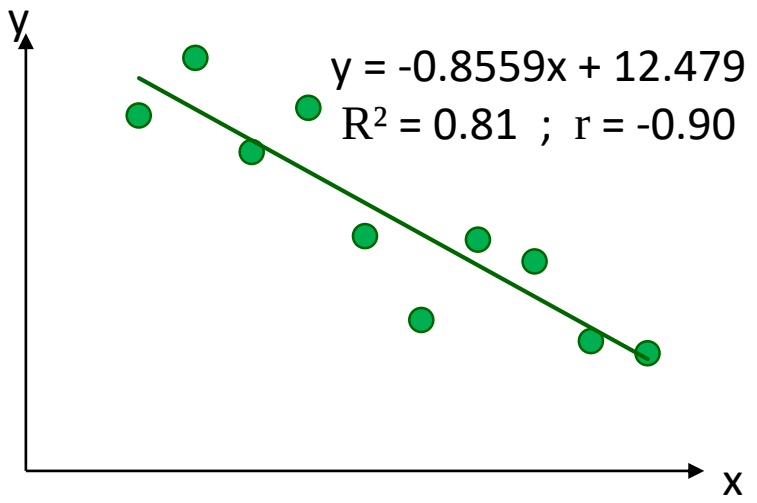
Coefficient of determination *vs* Sample Correlation Coefficient

Nevertheless, since $r_{XY} = b_1 \frac{S_X}{S_Y}$ and $R^2 = b_1^2 \frac{S_X^2}{S_Y^2}$, we have $R^2 = (r_{XY})^2$, and

because r_{xy} has the same sign as b_1:

- Knowing the sign of b_1 and R^2 we can find r_{xy}: $r_{XY} = \text{sign}(b_1) \times \sqrt{R^2}$
- Knowing r_{xy} we can find R^2: $R^2 = (r_{XY})^2$





Coefficient of determination *vs* F Statistic

One more result (a very important one!) relates the coefficient of determination to the F statistic.

Doing some simple calculations, it can be shown that:

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

Therefore, the knowledge about the coefficient of determination (or de sample correlation coefficient) allows to immediately compute the value of the F statistic and make a decision about the statistical significance of the linear regression model.

Significant model – Least Squares Estimates

Expected value of the response variable, Y, given the controlled variable, X:

$$\hat{\mu} = \hat{E}(Y) = b_0 + b_1 X \longrightarrow \textbf{Least squares regression line}$$

Variance of the response variable, Y, given the controlled variable, X:

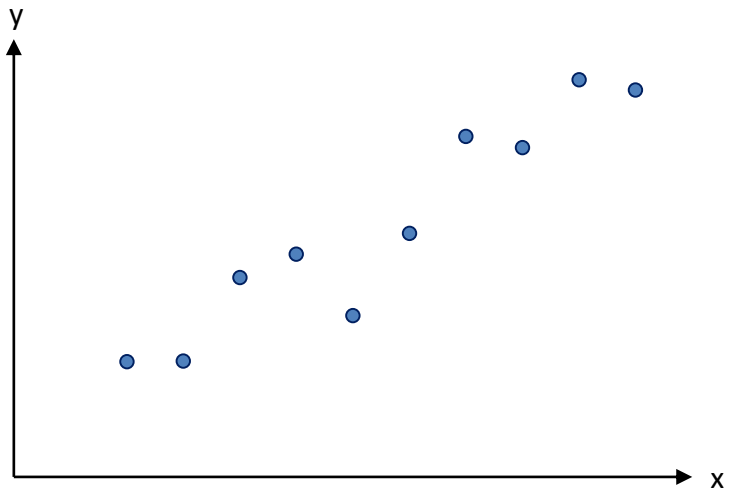
$$\hat{\sigma}^2 = \hat{V}ar(Y) = \hat{V}ar(\varepsilon) = S_e^2 = MSE = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2 \longrightarrow \textbf{Mean squared error}$$

Standard deviation of the response variable, Y, given the controlled variable, X:

$$\hat{\sigma} = S_e = \sqrt{MSE} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n} e_i^2} \longrightarrow \textbf{Standard error}$$

Least Squares Regression – Example

Obs.	X	Y
1	2	3.16
2	3	3.18
3	4	5.47
4	5	6.11
5	6	4.42
6	7	6.68
7	8	9.34
8	9	9.03
9	10	10.89
10	11	10.61
Sum	65	68.89
Sum of squares	505	551.4505
Sum of products		523.17



$$b_0 = \overline{Y} - b_1\overline{X} = \frac{68.89}{10} - 0.91376\,\frac{65}{10} \approx 0.94956$$

$$b_1 = \frac{\sum X_i Y_i - \frac{\left(\sum X_i\right)\left(\sum Y_i\right)}{n}}{\sum X_i^2 - \frac{\left(\sum X_i\right)^2}{n}} = \frac{523.17 - \frac{65 \times 68.89}{10}}{505 - \frac{(65)^2}{10}} \approx 0.91376$$

Least Squares Regression – Example

Obs.	X	Y
1	2	3.16
2	3	3.18
3	4	5.47
4	5	6.11
5	6	4.42
6	7	6.68
7	8	9.34
8	9	9.03
9	10	10.89
10	11	10.61
Sum	65	68.89
Sum of squares	505	551.4505
Sum of products		523.17