Simple Linear Regression

Presented and prepared by: Prof. Yasmina Guechari

1. What is a linear regression?

- **Linear Regression** is a form of statistical approach;
- Linear Regression is useful to examine and establish a relationship between two separate variables – independent and dependent variables.
- Linear Regression devided into two categories –
- # **Simple Linear Regression**: The model includes one independent variable
- # **Multiple Linear Regression:** This model includes more than one independent variables

1. What is a linear regression?

- **Dependent Variable:** This is the outcome or the effect that you are trying to measure or explain. It "depends" on other factors, which are the independent variables. A dependent variable Also known as the **predicted variable, explained variable**.
- **Independent variable:** This is the factor that you think might influence or cause changes in the dependent variable. It's what you observe to see if it has an effect. also called the **explanatory variable, exogenous variable, predicting variable.**
- **For example,** if you're studying how studying time affects exam scores, the exam score is the dependent variable.

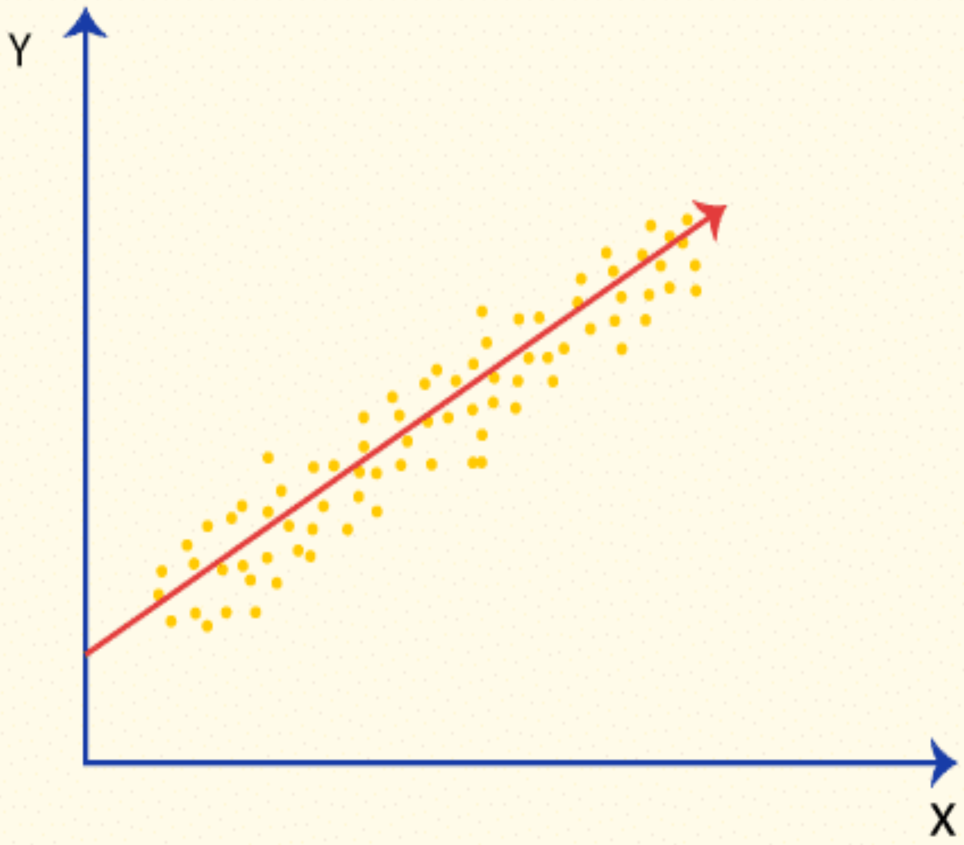
2. The concept of simple linear regression

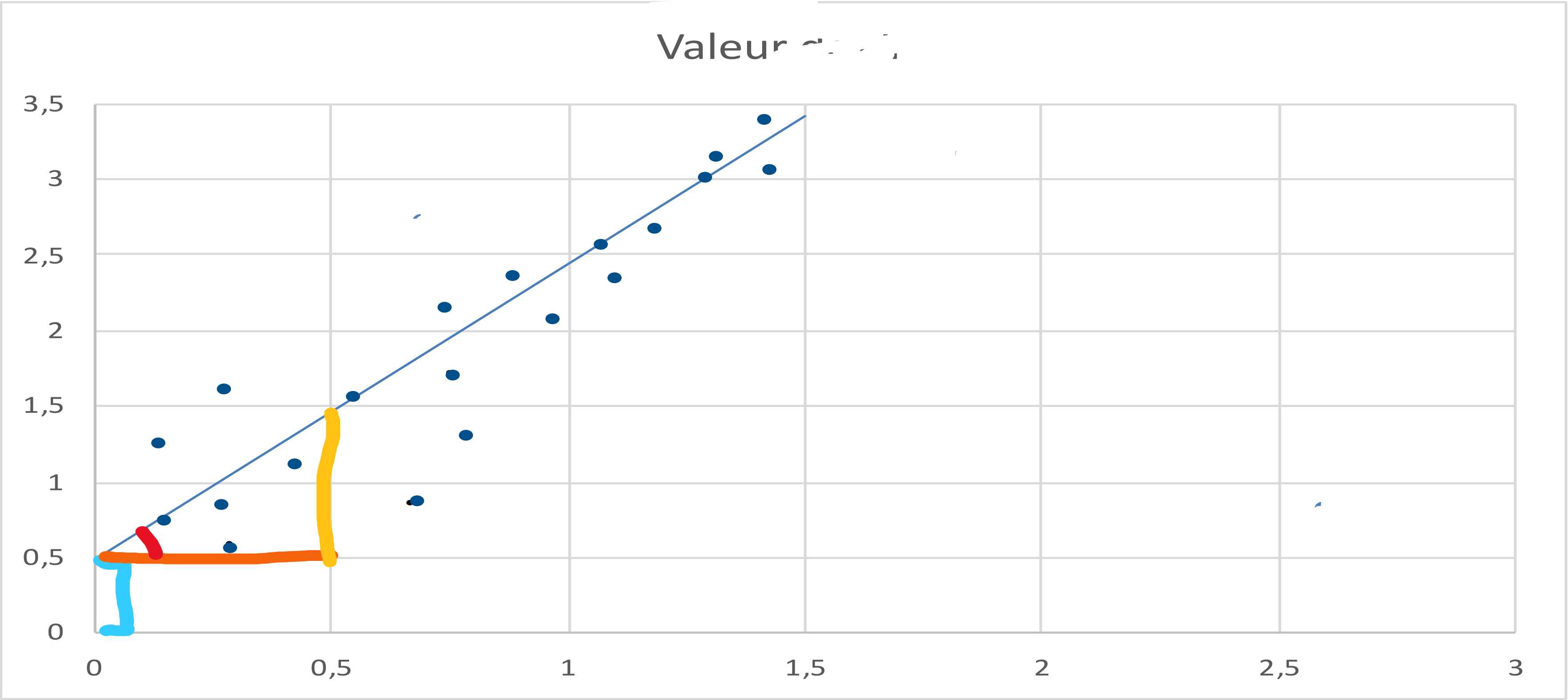
- ***Simple linear regression*** is a regression model that estimates the relationship between one independent variable and one dependent variable using a straight line.
- You can use simple linear regression when you want to know:

1. How strong the relationship is between two variables
2. The value of the dependent variable at a certain value of the independent variable -This helps in the forecasting or anticipating process.

Linear

Regression





Valeur d...
3,5
3
2,5
2
1,5
1
0,5
0
0
0,5
1
1,5
2
2,5
3

3. How to perform a simple linear regression

- **Simple linear regression formula:** **t**he formula for a simple linear regression is:

$$\boldsymbol{Y = \beta_0 + \beta_1 X + \varepsilon}$$

y is the predicted value of the dependent variable (**y**) for any given value of the independent variable (**x**).

$\boldsymbol{\beta_0}$ is the **intercept**, the predicted value of **y** when the **x** is 0.

$\boldsymbol{\beta_1}$ is the regression coefficient – how much we expect **y** to change as **x** increases or decrease.

x is the independent variable (the variable we expect is influencing **y**).

$\boldsymbol{\varepsilon}$ is the **error** of the estimate, or how much variation there is in our estimate of the regression coefficient.

3. How to perform a simple linear regression

- To find the linear equation by hand, you need to get the value of "β_0" and "β_1".
- Then substitute the resulting value in the slope formula and that gives you your linear regression equation.
- We will take the following example to understand how it is done.

3. How to perform a simple linear regression

- Lets take the following dataset as an example:

$\bar{X}$=$\frac{\sum_i^N x_i}{N}$ =16500/5= 3300, $\bar{Y}$=$\frac{\sum_i^N y_i}{N}$=12000/5=2400

Month	Income (Xi)	Money spent (Yi)	$(X_i - \bar{X})$	$(Y_i - \bar{Y})$	$(X_i - \bar{X})(Y_i - \bar{Y})$	$(X_i - \bar{X})^2$
1	2000	2000	-1300	-400	520000	1690000
2	3000	2200	-300	-200	60000	90000
3	4000	3000	700	600	420000	490000
4	2500	1200	-800	-1200	960000	640000
5	5000	3600	1700	1200	2040000	2890000
Sum	16500	12000			4000000	5800000

3. How to perform a simple linear regression

The least square estimator for the slope β_1 and intercept β_0 and error term ε in the linear regressin model are:

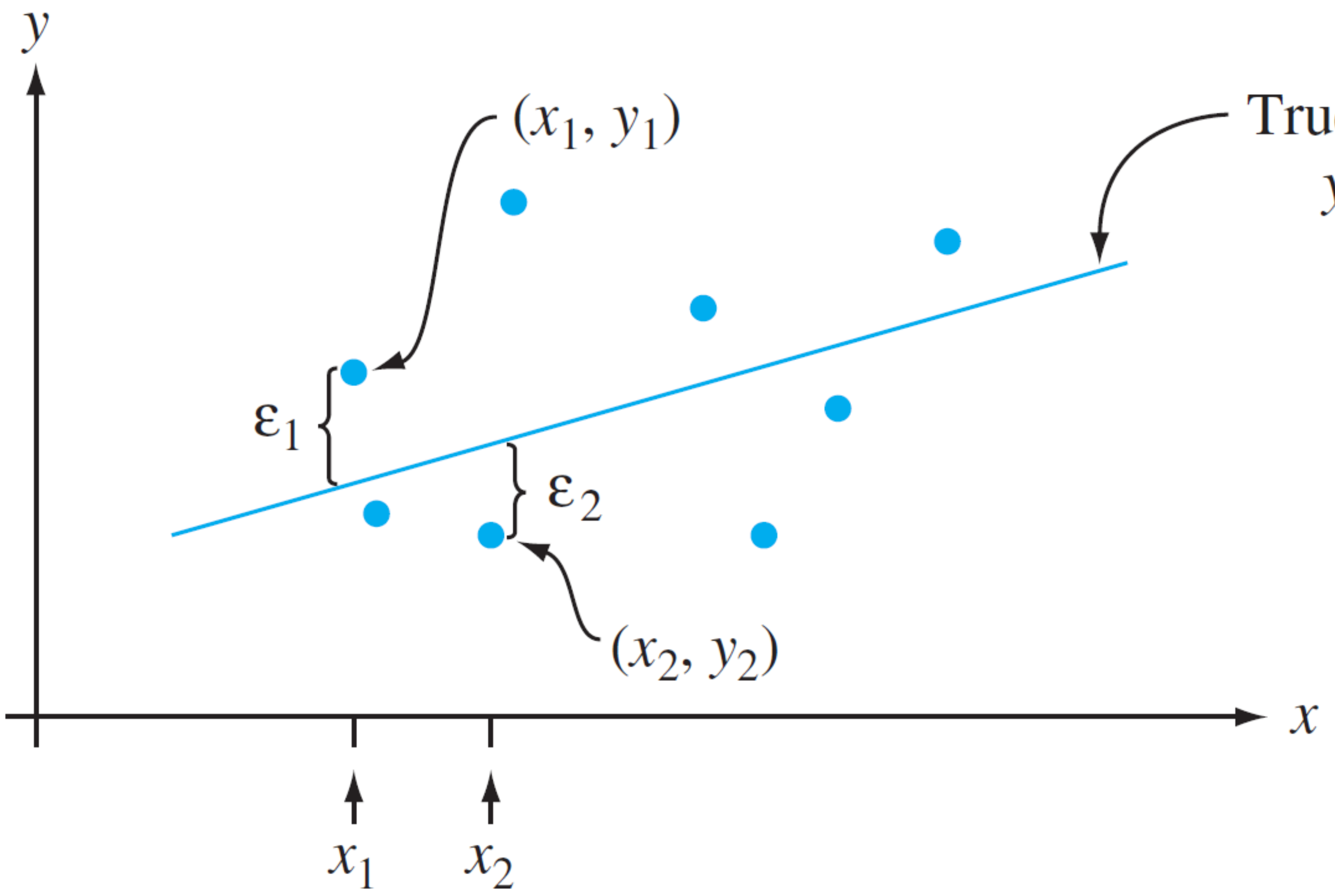
- $\hat{\beta}_1 = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{N}(X_i - \bar{X})^2}$..............................(1)
- $\hat{\beta}_0 = \overline{Y_i} - \hat{\beta}_1 \overline{X_i}$ ....(2)
- $\widehat{Y_i} = \hat{\beta}_0 + \hat{\beta}_1 X_i$ ....(3)
- $\widehat{\varepsilon_i} = Y_i - \widehat{Y_i}$...(4)
- **So, if we continue with the example in above table, the value of β_1 and β_0 can be as follow:**
- $\hat{\beta}_1 = \frac{4000000}{5800000} = 0.689$,$\hat{\beta}_0 = 2400 - 0.689 * 3300 = 126.3$
- **Our regression model is:**
- $\hat{Y} = 126.3 + 0.689X + \varepsilon$

3. How to perform a simple linear regression

- *If we want to make a prediction about the value of Y for a given value of X let's say 8000*
- $\widehat{Y_i}$=126.3+0.689*8000=5638.3
- To find the value of error we calculate $\widehat{Y_i}$ then substract it from the actual or the true value of *Y*
- $\widehat{Y_i}$ = 126.3+0.689*5000=3571.3
- The value of error is
- $\widehat{\varepsilon_i} = Y_i - \widehat{Y_i}$, $\widehat{\varepsilon_i}$=3600-3571.3=28.7

4. Estimating the deviation

- The value $Y_i - \widehat{Y}_i$, is a positive when the point lies above the regression line and a negative number when it lies below the line.
- The error (or residual) can be thought of as a measure of deviation and we can summarize the notation in the following way: $\widehat{\varepsilon}_i = Y_i - \widehat{Y}_i$
- The following example show the error;



y
(x_1, y_1)
True regression line
$y = \beta_0 + \beta_1 x$
ε_1
ε_2
(x_2, y_2)
x
x_1
x_2

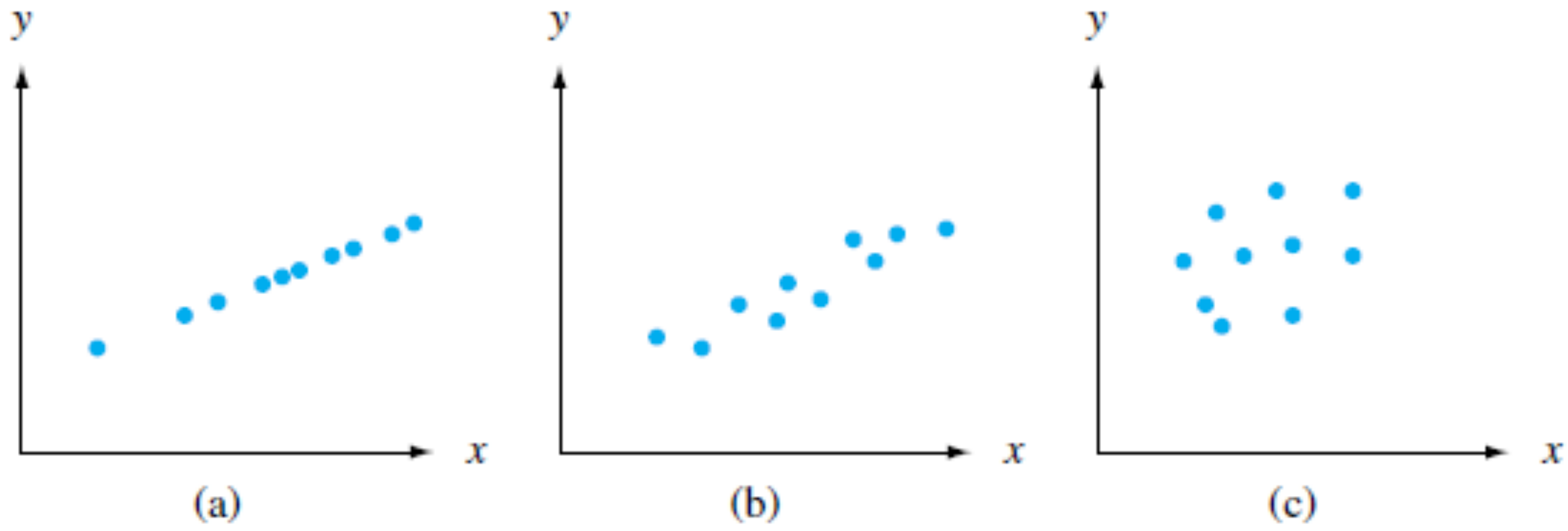
4. Estimating the deviation

- **The Sum of Squares Error denoted by SSE**, is defined as the variation of the dependent variable ***unexplained*** by the independent variable. SSE is given by the sum of the squared differences of the ***actual*** y-value Y_i and the ***predicted*** y-values $\hat{Y}_i$
- SSE$= \sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{N}\left[Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)\right]^2$

Month	X_i	Y_i	$\hat{Y}_i$	Y_i - $\hat{Y}_i$	$(Y_i - \hat{Y}_i)^2$	
1	2000	2000	1504.3	495.7	420292.89	$\hat{Y}_i = \widehat{\beta_0} - \widehat{\beta_1} X_i$
2	3000	2200	2193.3	6.7	44.89	$\widehat{\varepsilon_i} = Y_i - \widehat{Y_i}$
3	4000	3000	2882.3	117.7	13853.29	$sse = \sum_{i=1}^{N} (Y_i - \widehat{Y_i})^2$
4	2500	1200	1848.3	-648.3	420292.89	$\widehat{\sigma^2} = \frac{SSE}{N-2} = \frac{\sum_{i=1}^{N} (Y - \hat{Y}_i)^2}{N-2}$
5	5000	3600	3571.3	28.7	823.69	So, $\widehat{\sigma^2}$=680733.25/3
SSE					680733.25	$\widehat{\sigma^2}$=226911.08

Roughly speaking, 226911.08 is the *magnitude of a typical deviation from the estimated regression line*—some points are closer to the line than this and others are further away.

5. Model Validity



5. Model validity

- The *error sum of squares* SSE can be interpreted as a measure of how much variation in *y* is left *unexplained by the model*—
- In the first plot (a) SSE = 0, There is no unexplained variation, or all variation is explained. all the points are fall exactly on a straight line. In this case, *all (100%) of the variation in y can be attributed to the* variation in *x*.
- In the second plot (b) unexplained variation is small, means most variation is explained.
- In the third plot (c), *the simple linear regression model fails to explain variation in y by relating y to x*.

6. Model validity

- **Model Validity:** defined as "the process of checking that the model is a good representation of the target".

- **To examine the validity of the model**, we follow the following steps:

- Measuring the **explanatory power** and **correlational strength** of the model: For this purpose, we calculate the **The coefficient correlation and the coefficient of determination R^2**.

6. Model validity

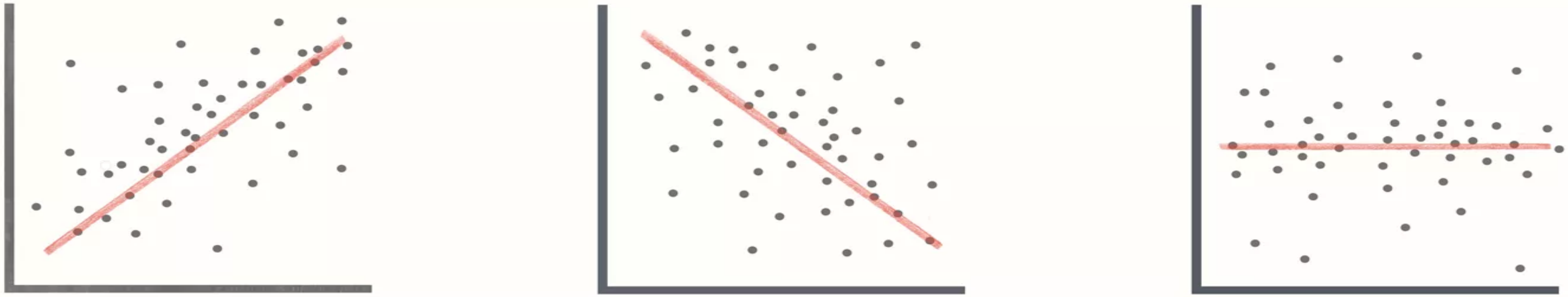
First- Correlation coefficient is the value that determine the strength of associations between data variables.

- The most common, called a Pearson correlation coefficient, measures the strength and the direction of a linear relationship between two variables.
- Values always range from -1 for a perfectly inverse, or negative, relationship to 1 for a perfectly positive correlation. Values at, or close to, zero indicate no linear relationship or a very weak correlation.

6. Model validity

- The further the coefficient is far from zero, whether it is positive or negative, the **better the fit** and **the greater the correlation**.
- The values of -1 (for a negative correlation) and 1 (for a positive correlation) describe perfect fits in which all data points align in a straight line, indicating that the variables are perfectly correlated.
- The closer the correlation coefficient is to zero the weaker the correlation, until at zero no linear relationship exists at all.

Correlation Coefficient



Positive Correlation   Negative Correlation   No Correlation

ThoughtCo.

6. Model validity

- To calculate the Pearson correlation, start by determining each variable's standard deviation as well as the covariance between them. The correlation coefficient is covariance divided by the product of the two variables' standard deviations.
- $\rho_{x,y} = \frac{Cov\ (x,y)}{\sigma_x \sigma_y}$
- $\rho_{x,y}$:pearson correlation coefficient
- $cov(x, y)$: covariance of variables x and y
- σ_x:standard deviation of x
- σ_y: standard deviation of y

6. Model validity

- Standard deviation is a measure of the dispersion of data from its average.
- Covariance shows whether the two variables change together,
- The correlation coefficient measures the strength of that relationship on a normalized scale, from -1 to 1.

6. Validity of model

- **Second- The coefficient of determination:** The coefficient of determination $(R)^2$ measures the proportion of the total variability of the dependent variable that is explained by the independent variable. It is calculated using the formula below:

- $R^2 = \frac{Explained\ variation}{Total\ variation} = \frac{Sum\ of\ Squares\ Regression\ (SSR)}{Sum\ of\ Squares\ Total\ (SST)}$ ……………………(I)

- Before we continue driving or formulating the formula of R^2, let's first define SSR and SST.

1. ***The sum of squares regression*** *is the variation of the dependent variable* ***explained by*** *the independent variable. It is given by the sum of the squared differences of the* ***predicted*** *y-value* $\hat{Y}_i$ and ***mean*** of y-observations$\bar{Y}$.

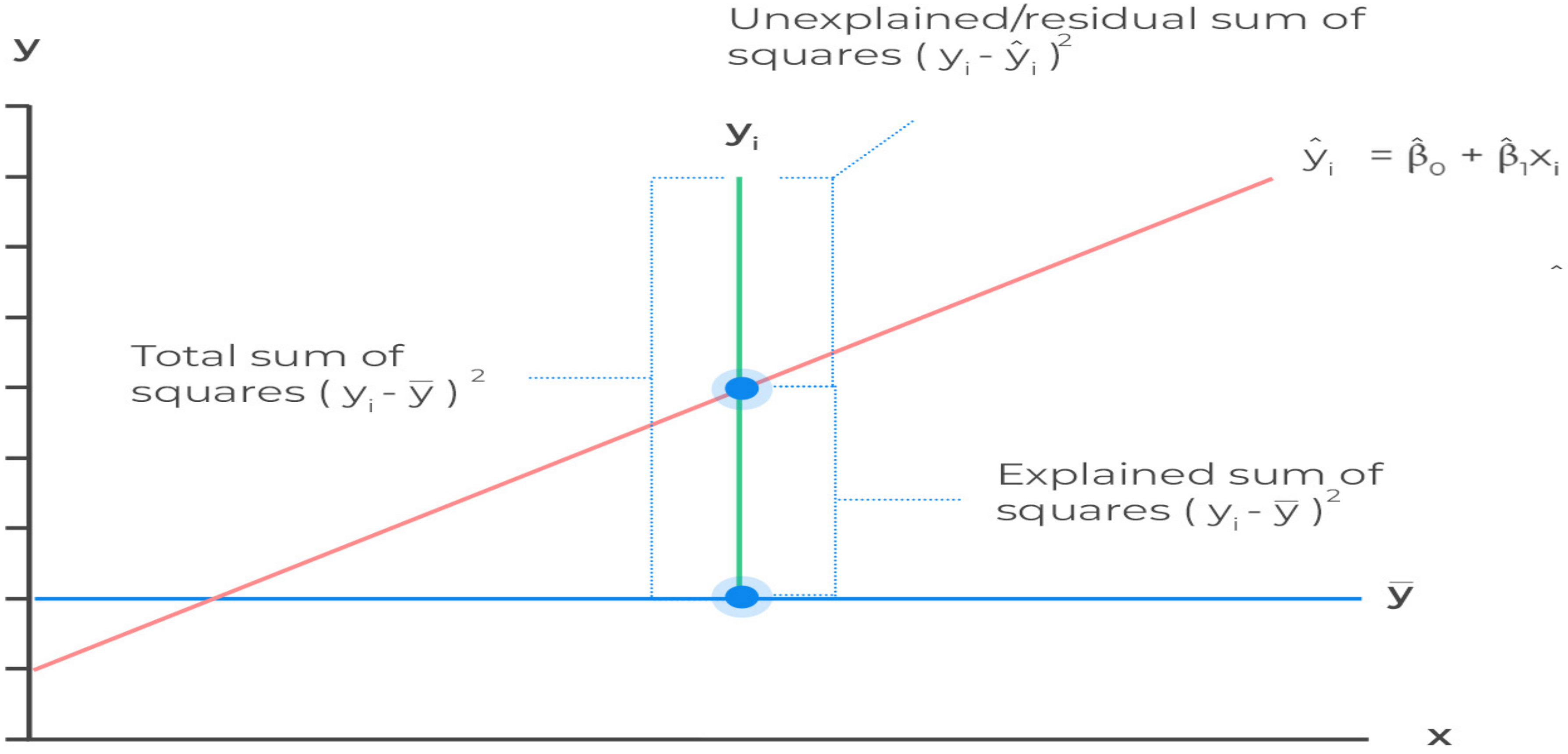
- $SSR = \sum_{i=1}^{N}\left(\hat{Y}_i - \bar{Y}\right)^2$

6. Model validity

2. **Sum of Squares Total (SST):** is a measure of the total variation of the dependent variable. It is the sum of the squared differences of the actual y-value and mean of y-observations.

- $SST = \sum(y_i - \bar{y})^2 = \sum y_i^2 - (\sum y_i)^2 / n$
- The Sum of Squares Total contains two parts:
- Sum of Square Regression (SSR).
- Sum of Squares Error (SSE). We talked about it in previous section
- $SST = SSE + SSR$
- Therefore, the sum of squares total is given by:
- Sum of Squares Total=Explained Variation + Unexplained Variation=SSR+ SSE

SST, SSR & SSE



y
Unexplained/residual sum of squares $(y_i - \hat{y}_i)^2$
y_i
$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$
Total sum of squares $(y_i - \bar{y})^2$
Explained sum of squares $(y_i - \bar{y})^2$
$\bar{y}$
x

6. Model Validity

- Let's continue with formula (I)
- $R^2 = \frac{Total\ variation - unexplained\ variation}{total\ variation} = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST}$ ....(II)
- $R^2 = 1 - SSE/SST = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$ ....(III)
- (a number between 0 and 1) is the proportion of observed *y* variation explained by the model.
- Note that if $SSE = 0$. as in case (a), then R^2=1 .
- It is interpreted as the proportion of observed y variation that can be explained by the simple linear regression model. The higher the value of R^2, the more successful is the simple linear regression model in explaining y variation.

6. Validity of model

Month	Y_i	$\bar{y}$	$y_i - \bar{y}$	$(y_i - \bar{y})^2$	
1	2000	2400	-400	160000	SST=3440000
2	2200	2400	-200	40000	SSE=680733.25
3	3000	2400	600	360000	$r^2 = 1 - \frac{SSE}{SST} = 1 - \frac{680733.25}{3440000}$
4	1200	2400	-1200	1440000	$R^2 = 1 - 0.1979 =$ 0.802 = 80.2%
5	3600	2400	1200	1440000	
Sum				3440000	

That is, 80.2% of the observed variation in Money spent can be explained by the simple linear regression relationship between money spent and Income value.

6. Model validity

- **Features of Coefficient of Determination R^2**

$i.$ R^2 lies between 0 and 1.

ii. A high R^2 explains variability better than a low R^2.

iii. If R^2 =0.01, only 1% of the total variability in Y can be explained. On the other hand, if R^2 =0.90, over 90% of the total variability in Y can be explained.

iv. The higher the R^2, the higher the explanatory power of the model.

v. For models with one independent variable, R^2 is calculated by squaring the correlation coefficient between the dependent and the independent variables:

- $R^2 = \left(\frac{Cov(x,y)}{\sigma_y \sigma_x}\right)^2$ ....(IV)

7. Test of overall significance in regression

- Once the estimation of a regression model is complete, we would like to:

➢ check the statistical significance of a regression model, this requires us to calculate the F-statistic.

➢ The F-statistic confirms whether the slope (denoted by β_i) in a regression model is equal to zero.

➢ In a typical simple linear regression hypothesis, the null hypothesis is formulated as: $H_0: \beta_1 = 0$ against the alternative hypothesis, $H_0: \beta_1 \neq 0$.

7. Test of overall significance in regression

- The Sum of Squares Regression (SSR) and Sum of Squares Error (SSE) are employed to calculate the F-statistic. In the calculation, both the Sum of Squares Regression (SSR) and Sum of Squares Error (SSE) are adjusted for the degrees of freedom.
- The **Sum of Squares Regression is divided by the number of independent variables**, k, to get the Mean Square Regression (MSR).That is:
- $MSR = \frac{SSR}{k} = \frac{\sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2}{k} = \sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2$
- Therefore, in Simple Linear Regression Model, MSR = SSR.

7. Test of overall significance in regression

- Also, the Sum of Squares Error (SSE) is divided by degrees of freedom given by n−k−1 (this translates to n−2 for simple linear regression) to arrive at Mean Square Error (MSE). That is,
- $MSE = \frac{SSE}{n-k-1} = \frac{\sum_{i=1}^{N}(y_i-\hat{y}_i)^2}{n-k-1} = \frac{\sum_{i=1}^{N}(y_i-\hat{y}_i)^2}{n-2}$ since *K=1*
- Finally, to calculate the F-statistic for the linear regression, we find the ratio of MSR to MSE. That is,
- $F-statistic = \frac{MSR}{MSE} = \frac{\frac{SSR}{k}}{\frac{SSE}{n-k-1}}$ since *k=1* $F-statistic = \frac{SSR}{\frac{SSE}{n-2}}$
- $F-statistic = \frac{SSR}{\frac{SSE}{n-2}} = \frac{\sum_{i=1}^{N}(\hat{y}_i-\bar{y})^2}{\frac{\sum_{i=1}^{N}(y_i-\hat{y}_i)^2}{n-2}}$

7. Test of overall significance in regression

- A large F-statistic value proves that the regression model is effective in its explanation of the variation in the dependent variable and vice versa.
- An F-statistic of 0 indicates that the independent variable does not explain the variation in the dependent variable.

Month	y_i	$\hat{y}_i$	$y_i - \hat{y}_i$	$\hat{y}_i - \bar{y}$	$(y_i - \hat{y}_i)^2$	$(\hat{y}_i - \bar{y})^2$
1	2000	1504.3	495.7	-895.7	420292.89	802227.49
2	2200	2193.3	6.7	-206.7	44.89	42724.89
3	3000	2882.3	117.7	482.3	13853.29	232613.29
4	1200	1848.3	-648.3	-551.7	420292.89	304372.89
5	3600	3571.3	28.7	1171.3	823.69	1371943.69
Sum	12000				680733.25	2753882.25

$\bar{y}$ = 2400

$$F - statistic = \frac{SSR}{\frac{SSE}{n-2}} = \frac{\sum_{i=1}^{N} (\hat{y}_i - \bar{y})^2}{\frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{n-2}} = \frac{2753882.25}{\frac{680733.25}{3}} = 12.14$$

7. Test of overall significance in regression

- When you fit a regression model to a dataset, you will receive a regression table as output, which will tell you the F-statistic along with the corresponding p-value for that F-statistic.
- If the p-value is less than the significance level (*common level are .01, .05, and .10*), then you have sufficient evidence to conclude that your regression model fits the data better than the intercept-only model.
- For example if we find in our regression the value:
- **F-statistic:** 12.14
- **P-value:** 0.0332
- Since the p-value is less than the significance level, we can conclude that our regression model fits the data better than the intercept-only model.

Example of how to perform a linear regression in Microsoft Excel

Hour studied (X)	Exam Score (Y)
2	60
3	62
4	66
5	71
6	78
8	80
10	86
12	93

8. How to read the value in Excel table

	Coefficients	Standard error	T-stat	P-value	Lower 95%	Upper 95%
Constant	53,7390	1,5463	34,754069	3,78E-08	49,956138	57,52339408
Hour studied (X)	3,3216	0,2192	15,15160	5,21E-06	2,7852088	3,858065982

ANALYSE DE VARIANCE = ANOVA-test					
	Degree of freedom	Somme des carrés	Moyenne des carrés	F	Valeur critique de F
Regression SSR	1	943,345	943,345	229,5712	5,2135E-06
Residual SSE	6	24,655	4,109		
Total SST	7	968			

Regression Statistics	
Multiple R	0,987182855
Coefficient of determination R^2	0,974529989
Adjusted R^2	0,970284987
Standard error	2,027106754
Observations	8

8. How to read the table

- From the table we see: $\hat{\beta}_1$ = 3.3216 and $\hat{\beta}_0$=53.7390
- Standard error related to $\hat{\beta}_1$ is 0.2192, this value provides an estimate of how much the estimated value of parameter $\hat{\beta}$ or other statistic is likely to vary from the true population parameter.
- SSE=24.655
- SSR=943.345
- SST= 968
- F-statistic=229.5712.
- R^2 = 0.9745

8. Testing the hypothesis

In our example we are interested in determining the relationship between exam scores and the number of hours studied;

- Here, the hypothesis: $\begin{Bmatrix} H_0:\ \beta_1 = 0 \\ H_a:\ \beta_1 \neq 0 \end{Bmatrix}$
- After testing the data we get the the value of $\hat{\beta}_1$ = 3.3216 and t-stat=15.1516, p-value=5,21E-06, lower 95% =2.7852 & Upper95%=3.858

Testing the hypothesis- **using confidence interval**

- **A confidence interval** is a statistical range within which the true population parameter is likely to fall. It provides a range of values rather than a single point estimate and is associated with a certain level of confidence (eg: 95%).

Testing the hypothesis- using confidence interval

- **Example:** In the table we see that the value of Coefficient $\hat{\beta}_1$ is 3.3216 which is beteween the lower limit (2.758) and Upper limit (3.858) for confidence level of 95% . The value of $\hat{\beta}_0$is 53.7390 which is beteween the lower (49.956) and Upper limit (57.523) for confidence level of 95%
- From the tables, the estimated value of the $\hat{\beta}_1$ is statistically significant, as its value belongs to the interval determined by the confidence level.

8. Testing the hypothesis **using t-statistic**

- **T-statistic:** the calculated t-statistic is 15.1516, we compare this value to the critical value from t-distribution in the table below, to do so we need to Know the degree of freedom and confidence level

i. The degree of freedom (*df*) of a statistic is calculated from the sample size (*n*).

- *df* = *n* – 2 ; in our example *df* = 6

ii. The significance level: the significance level 95%, or $\alpha = 0.05$....(1-0.95)

- So, from the table of critical values of t, we can see that the critical value from t-distribution is 1.943 which is less than the calculated t-statistic, so we reject the null hypothesis and accept the alternative hypothesis

Critical values of *t* for one-tailed tests

Significance level (α)

Degrees of freedom (*df*)	.2	.15	.1	.05	.025	.01	.005	.001
1	1.376	1.963	3.078	6.314	12.706	31.821	63.657	318.309
2	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327
3	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215
4	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173
5	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893
6	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208
7	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785
8	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501
9	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297
10	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144
11	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025
12	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930
13	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852
14	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787
15	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733
16	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686
17	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646
18	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610
19	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579
20	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552
21	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527
22	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505
23	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485
24	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467
25	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450
26	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435
27	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421
28	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408
29	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396
30	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385
40	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307
50	0.849	1.047	1.299	1.676	2.009	2.403	2.678	3.261
60	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232
70	0.847	1.044	1.294	1.667	1.994	2.381	2.648	3.211
80	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195
100	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174
1000	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098
Infinite	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090

Scribbr

8. Testing the hypothesis **using P-value**

- **P-value:**is the probability of rejecting the null hypothesis when it is true. It is often set before conducting the test and represents the maximum allowable probability of making error.
- Commonly used significance levels are 0.05, 0.01, and 0.10.
- A significance level of 0.05, for example, means that you are willing to accept a 5% ($\alpha = 0.05$)chance of rejecting the null hypothesis when it is actually true.

8. Testing the hypothesis **using P-value**

- So, if the p-vale is less than or equal to the significance level, you reject the null hypothesis.
- If the p-value is greater than the significance level, you fail to reject the null hypothesis.
- In our example t-statistic p- value related to the coefficient $\hat{\beta}_1$ is very small (5.2135E-06=0.000005213= 0.000521%) is less than 1% , so, we conclude that there is a significant positive relationship between the hour studied and exam score, for each 1% increase in hour studied the exam score move up by 3.321%.

- We can also see that the p-value related to F-statistic is very small (0.000036),so the relationship is highly statistically significant.
- From the R-squared, we can see that the the hour studied alone can explain more than 97% of the observed variations in the exam score.