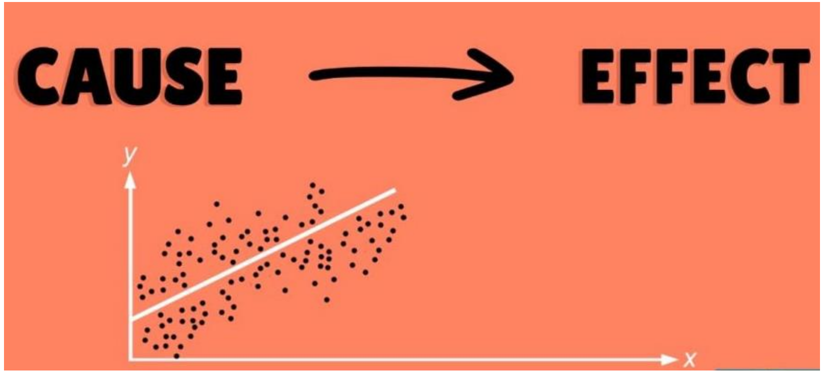
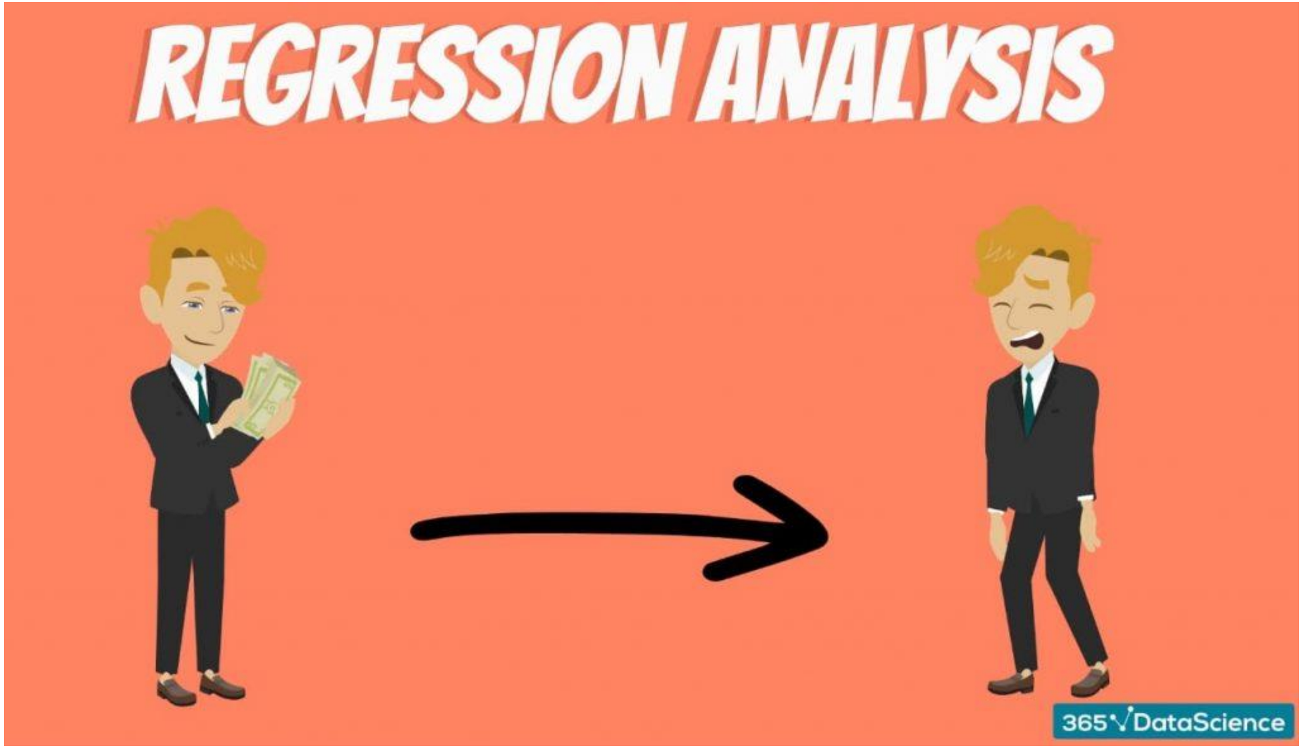
Regression

Regression Analysis

- **Regression analysis** is one of the most widely used methods for prediction.
- It is applied whenever we have a causal relationship between variables





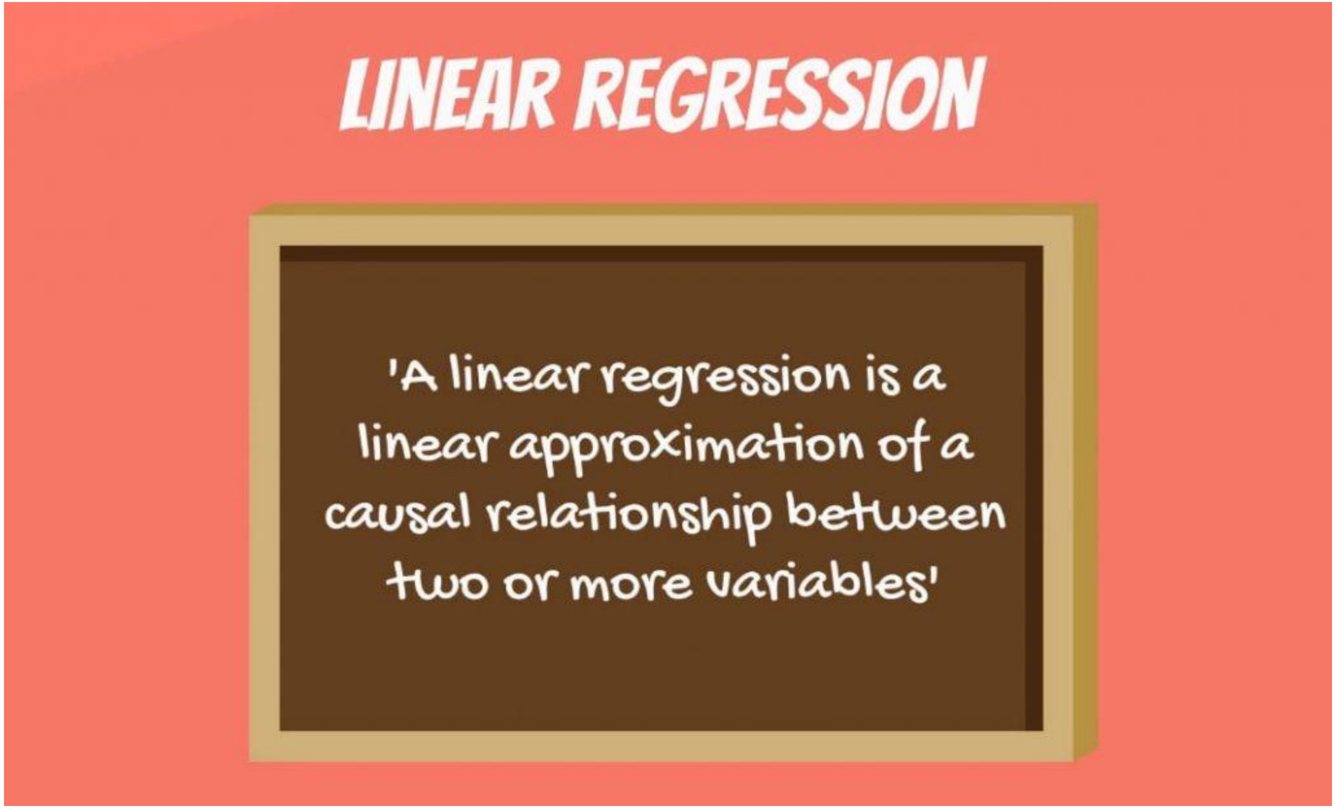
- "The amount of money you spend depends on the amount of money you earn."

Regression

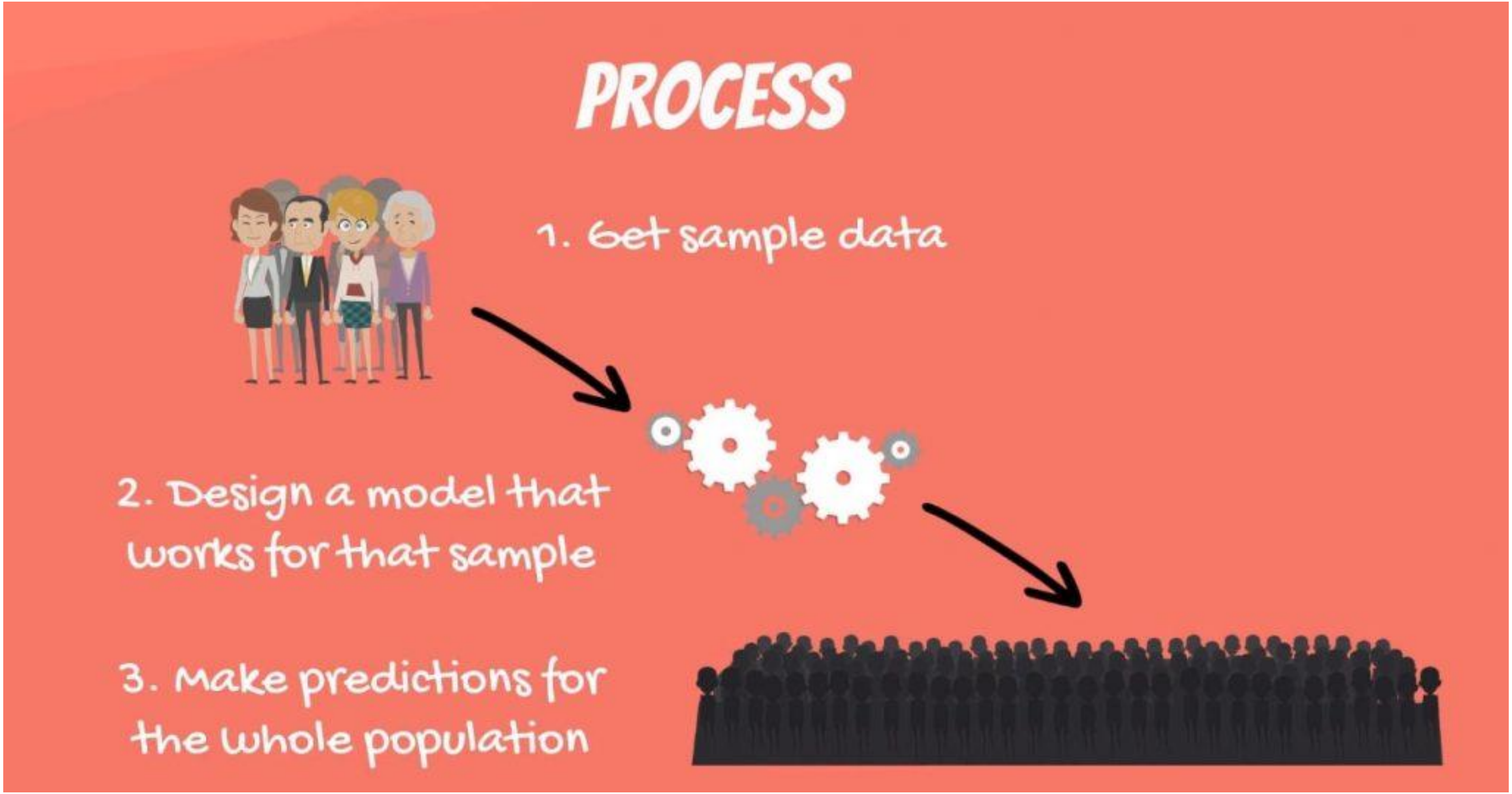
- Regression is a statistical technique that relates a dependent variable to one or more independent variables.
- A regression model is able to show whether changes observed in the dependent variable are associated with changes in one or more of the independent variables.
- Example: we can say that age and height can be described using a linear regression model.
- Since a person's height increases as age increases, they have a linear relationship.

Purpose of Regression

- In statistical analysis, regression is used to identify the associations between variables occurring in some data.
- It can show the magnitude of such an association and determine its statistical significance.
- Regression is a powerful tool for statistical inference and has been used to try to predict future outcomes based on past observations.



- **Regression** models are highly valuable, as they are one of the most common ways to make inferences and predictions.


PROCESS
1. Get sample data
2. Design a model that works for that sample
3. Make predictions for the whole population

Linear Regression

- Linear regression is a statistical practice of calculating a straight line that specifies a mathematical relationship between two variables.
- Linear regression is defined as an algorithm that provides a linear relationship between an independent variable and a dependent variable to predict the outcome of future events.

Linear Regression

- The independent variable is also the predictor or explanatory variable that remains unchanged due to the change in other variables.
- However, the dependent variable changes with fluctuations in the independent variable.
- The regression model predicts the value of the dependent variable, which is the response or outcome variable being analyzed or studied.

Linear Regression

- Thus, linear regression is a supervised learning algorithm that simulates a mathematical relationship between variables and makes predictions for continuous or numeric variables such as sales, salary, age, product price, etc.

Linear Regression

- There is a dependent variable, labeled *Y*, being predicted, and independent variables, labeled *x1*, *x2*, and so forth.
- These are the predictors.
- *Y* is a function of the *X* variables, and the **regression model** is a linear approximation of this function.

DEPENDENT

/predicted/

INDEPENDENT

/predictors/

$$Y = F(x_1, x_2, \ldots, x_k)$$

The dependent variable Y is a function of the independent variables x1 to xk

SIMPLE LINEAR REGRESSION MODEL

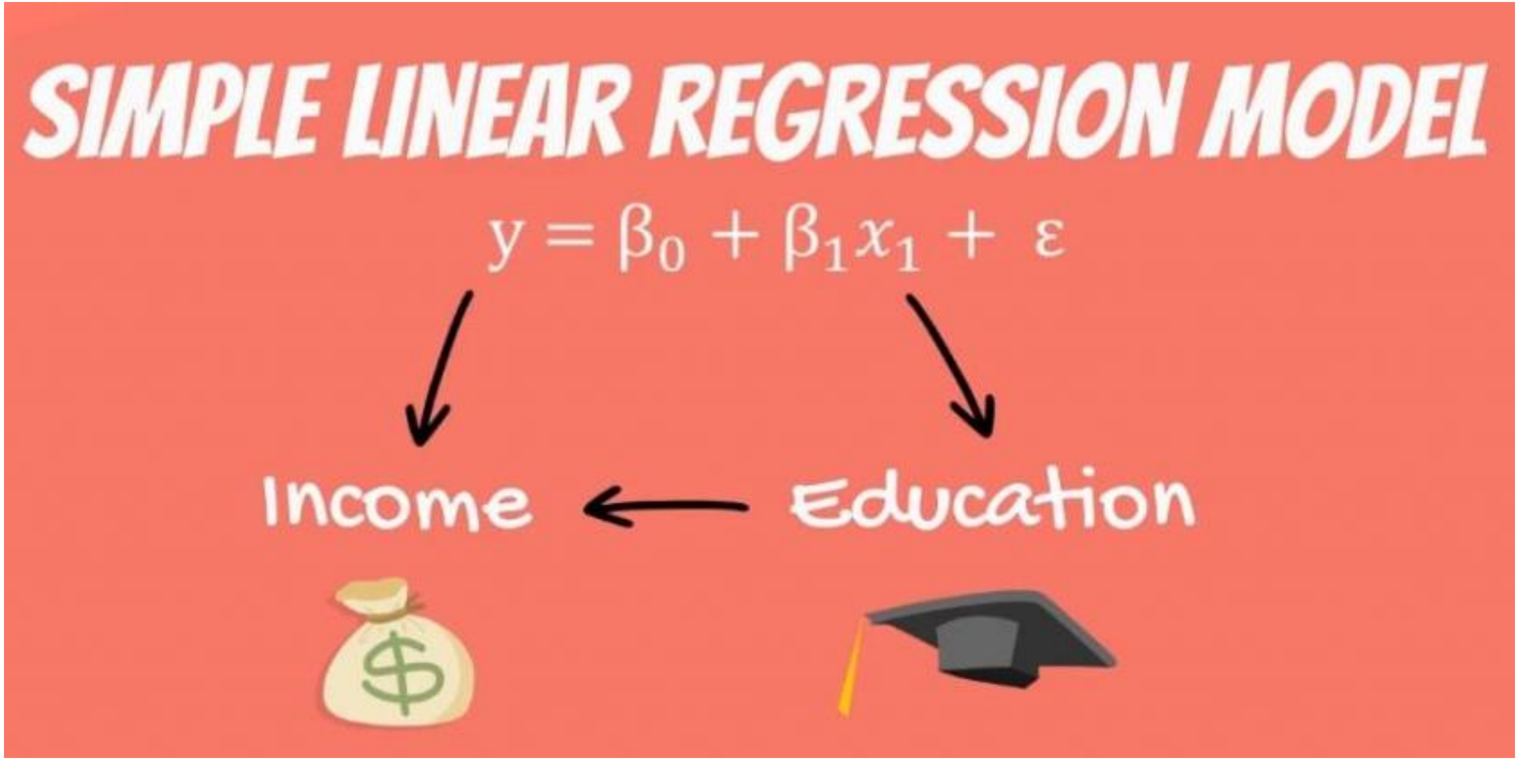
$$y = \beta_0 + \beta_1 x_1 + \varepsilon$$

Dependent variable

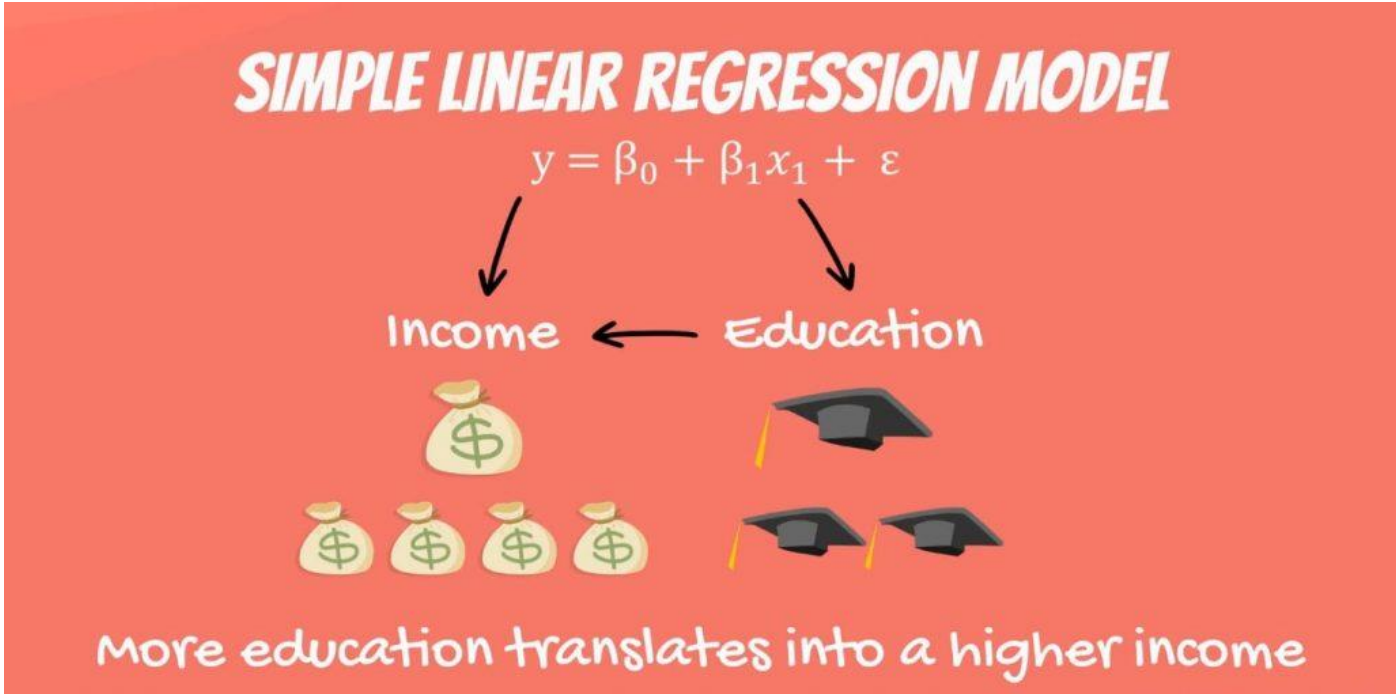
Independent variable

Simple Linear Regression

- When using **regression analysis**, we want to predict the value of Y, provided we have the value of X.
- But to have a **regression**, Y must depend on X in some way. Whenever there is a change in X, such change must translate to a change in Y.

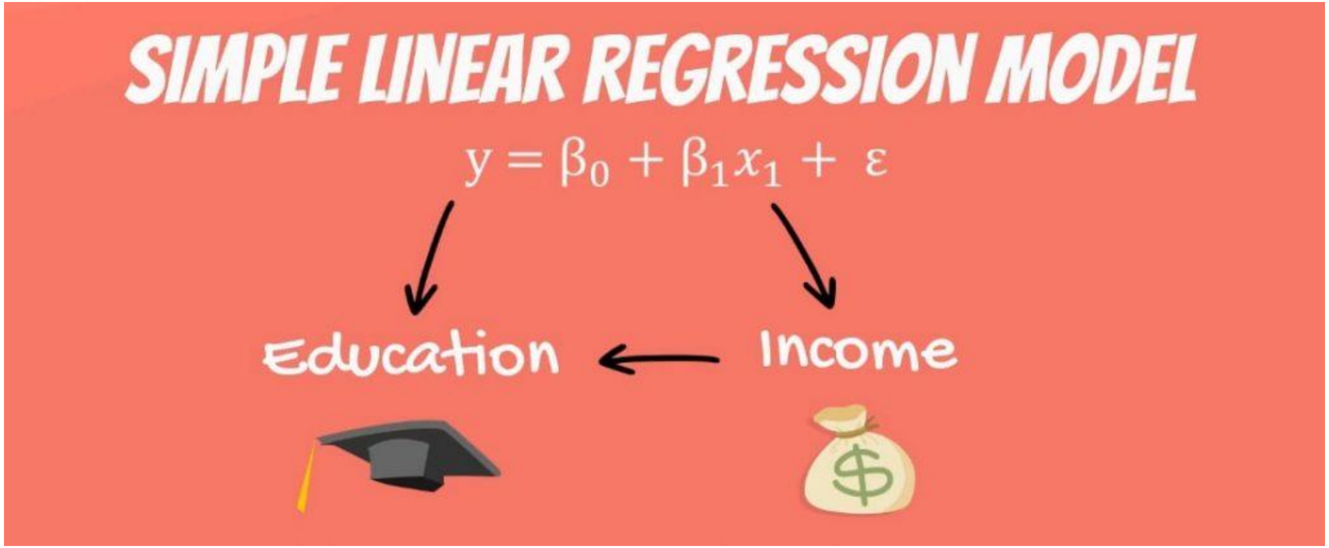


- the income a person receives depends on the number of years of education that person has received.
- The *dependent variable* is income, while the *independent variable* is years of education.

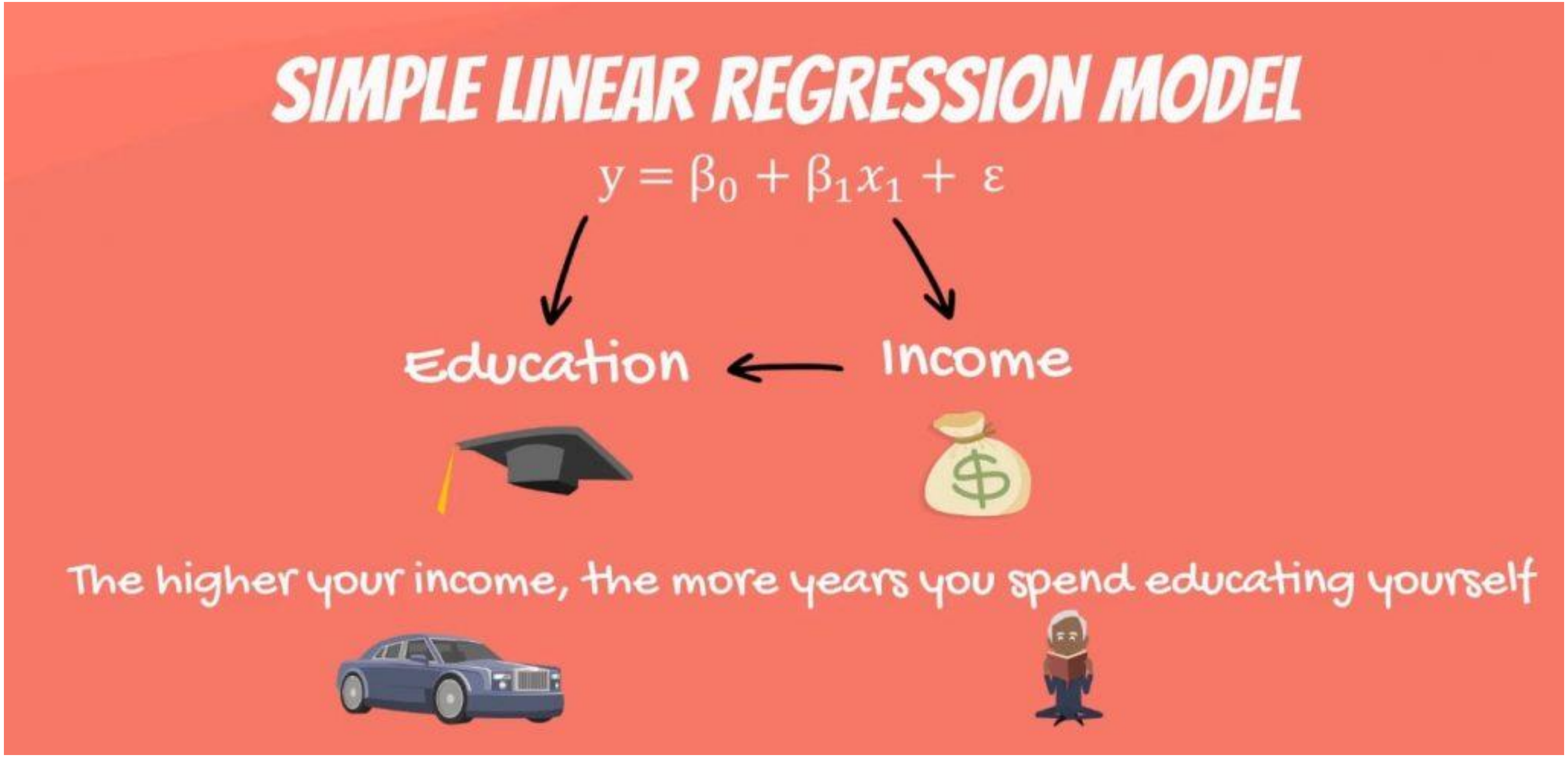


- There is a causal relationship between the two.
- The more education you get, the higher the income you are likely to receive.

Is the Reverse Relationship Possible?

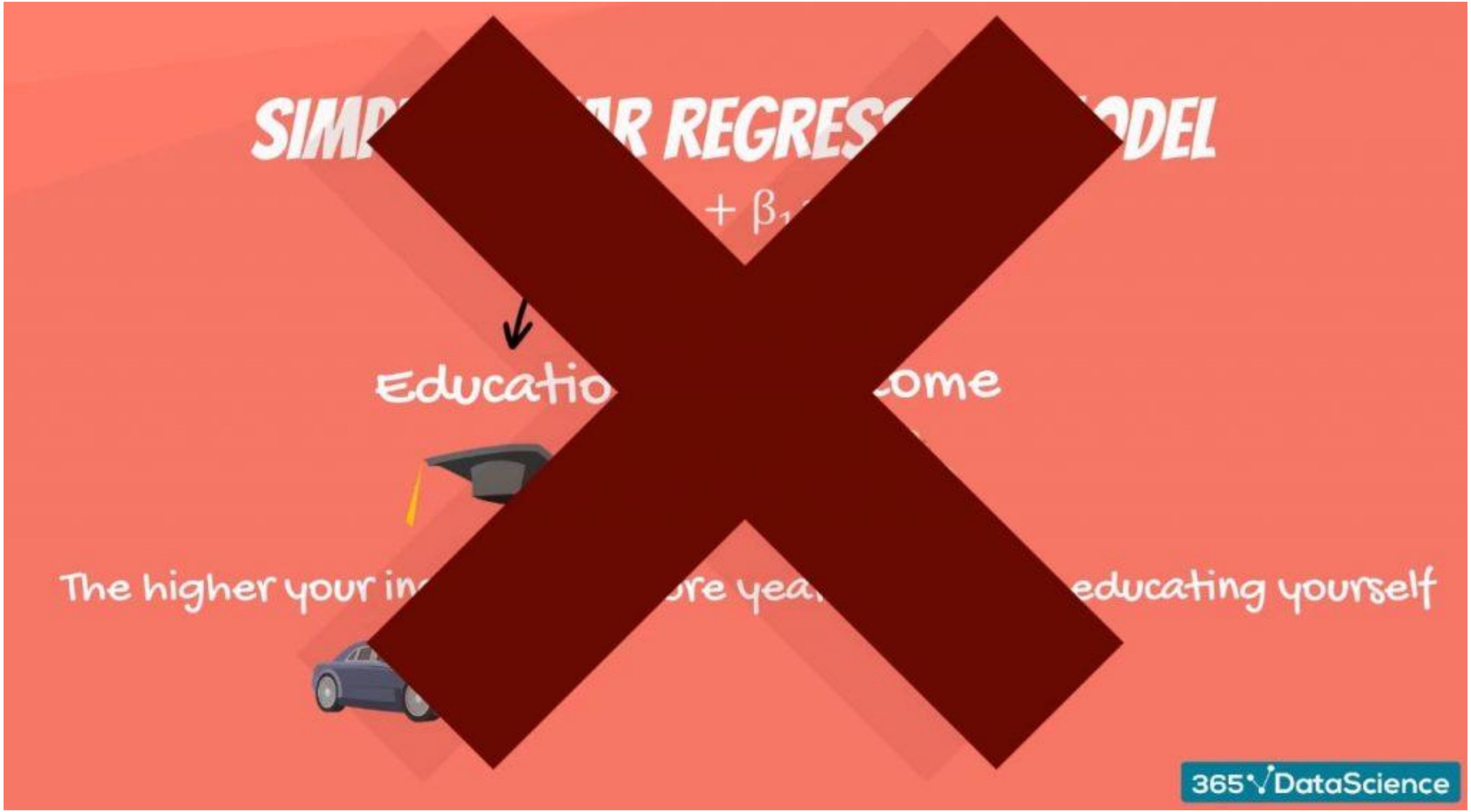


- What if education depends on income.

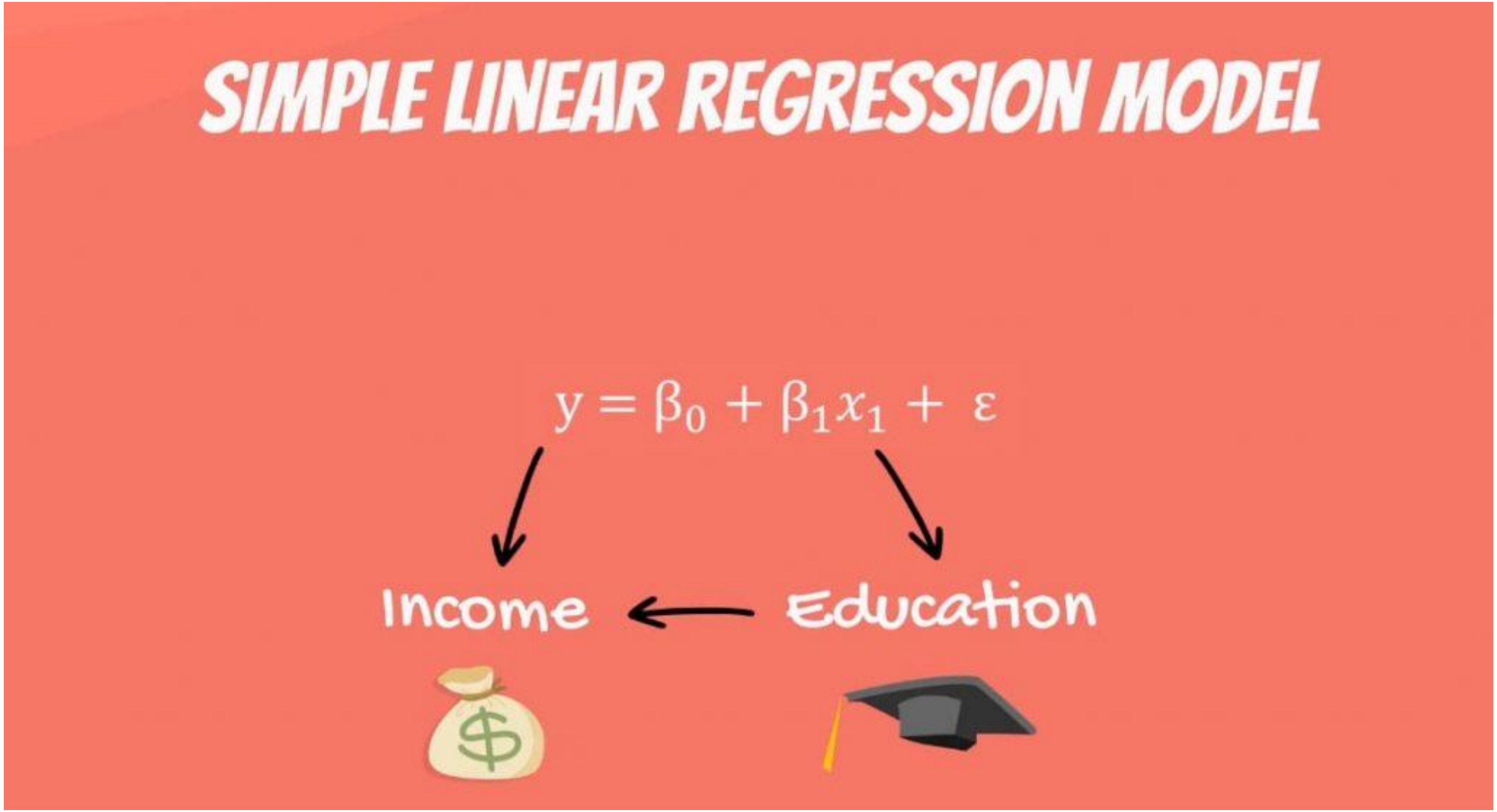


- This would mean the higher your income, the more years you spend educating yourself.

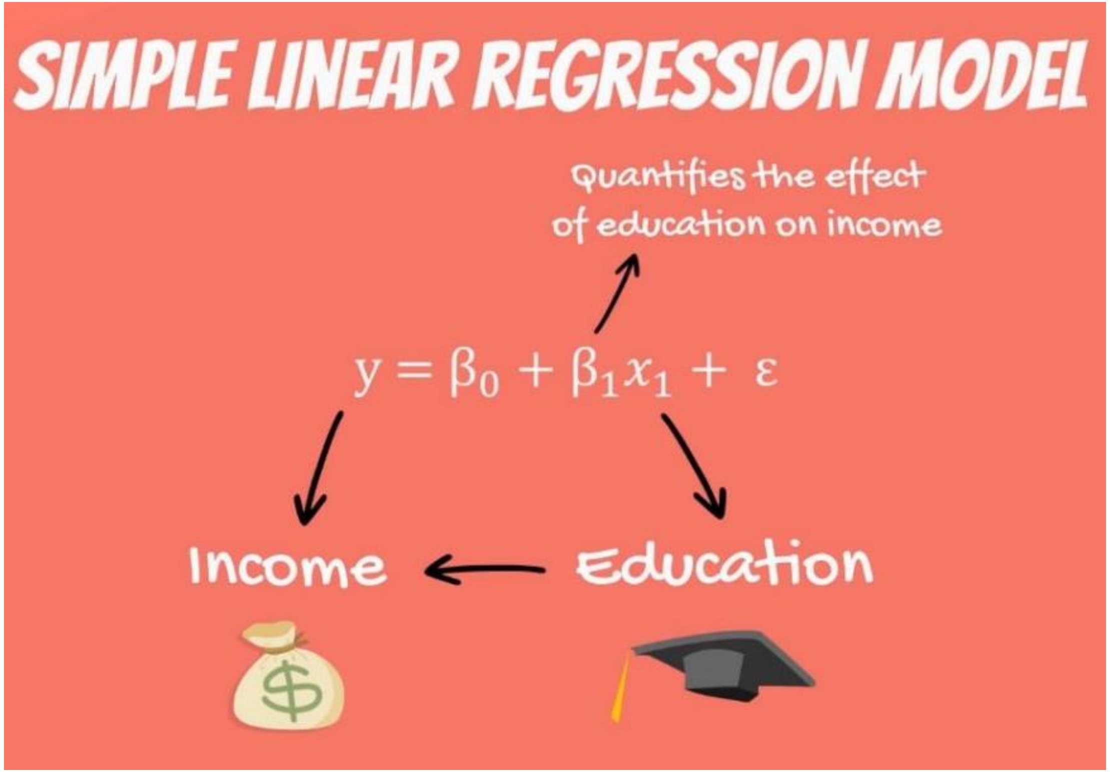
- Putting high tuition fees aside, wealthier individuals don’t spend more **years** in school.
- Moreover, high school and college take the same number of years, no matter your tax bracket.
- Therefore, a causal relationship like this one is faulty, if not plain wrong. Hence, it is unfit for **regression analysis**.


The higher your
educating yourself
365√DataScience

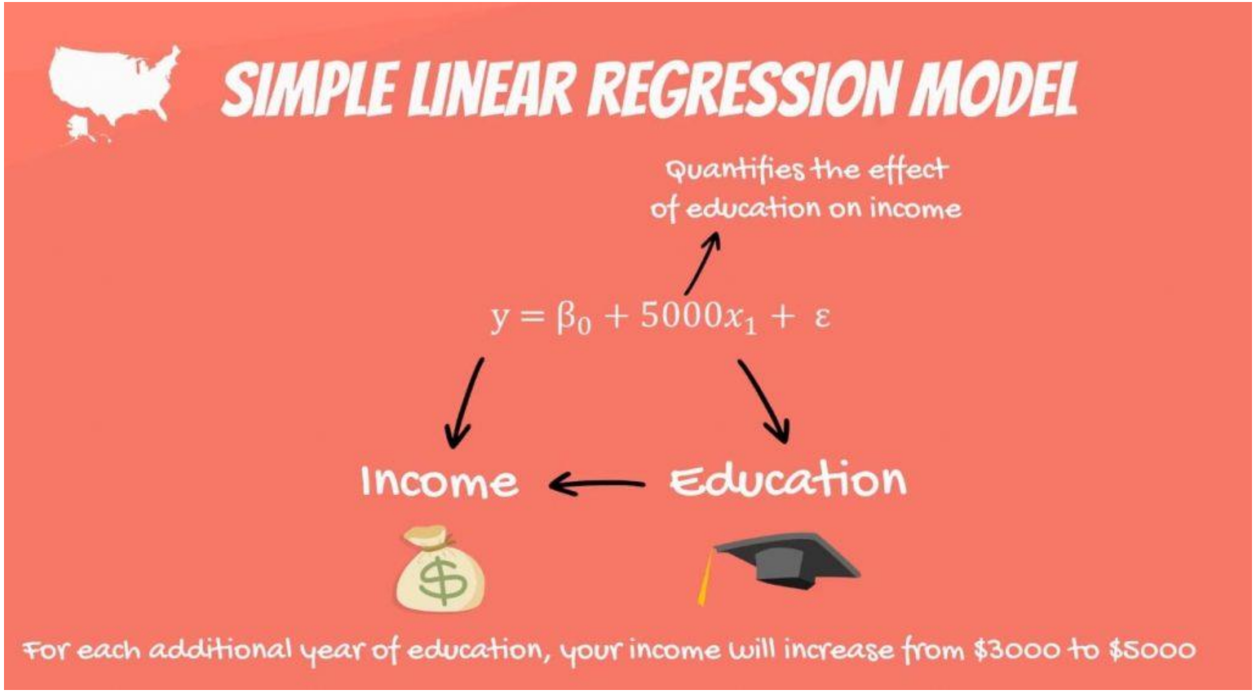
- Let’s go back to the original **linear regression** example.
- Income is a function of education.
- The more years you study, the higher the income you will receive.
- This sounds about right.


SIMPLE LINEAR REGRESSION MODEL
$y = \beta_0 + \beta_1 x_1 + \varepsilon$
Income
Education

The Coefficients



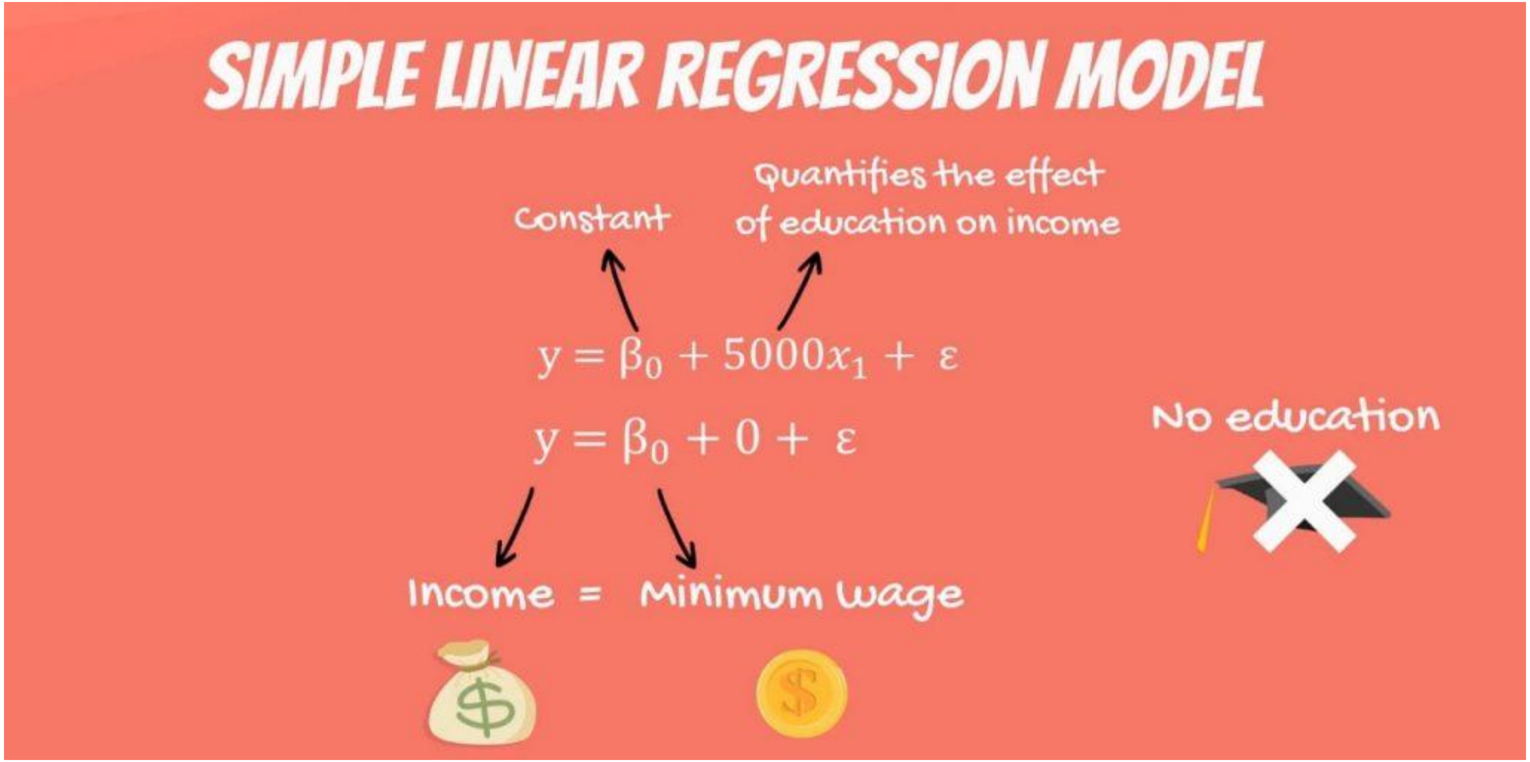
- β_1 is the coefficient that stands before the independent variable.
- It quantifies the effect of education on income.



- If β_1 is 50, then for each additional year of education, your income would grow by $50.
- In the USA, the number is much bigger, somewhere around 3 to 5 thousand dollars.

The Constant

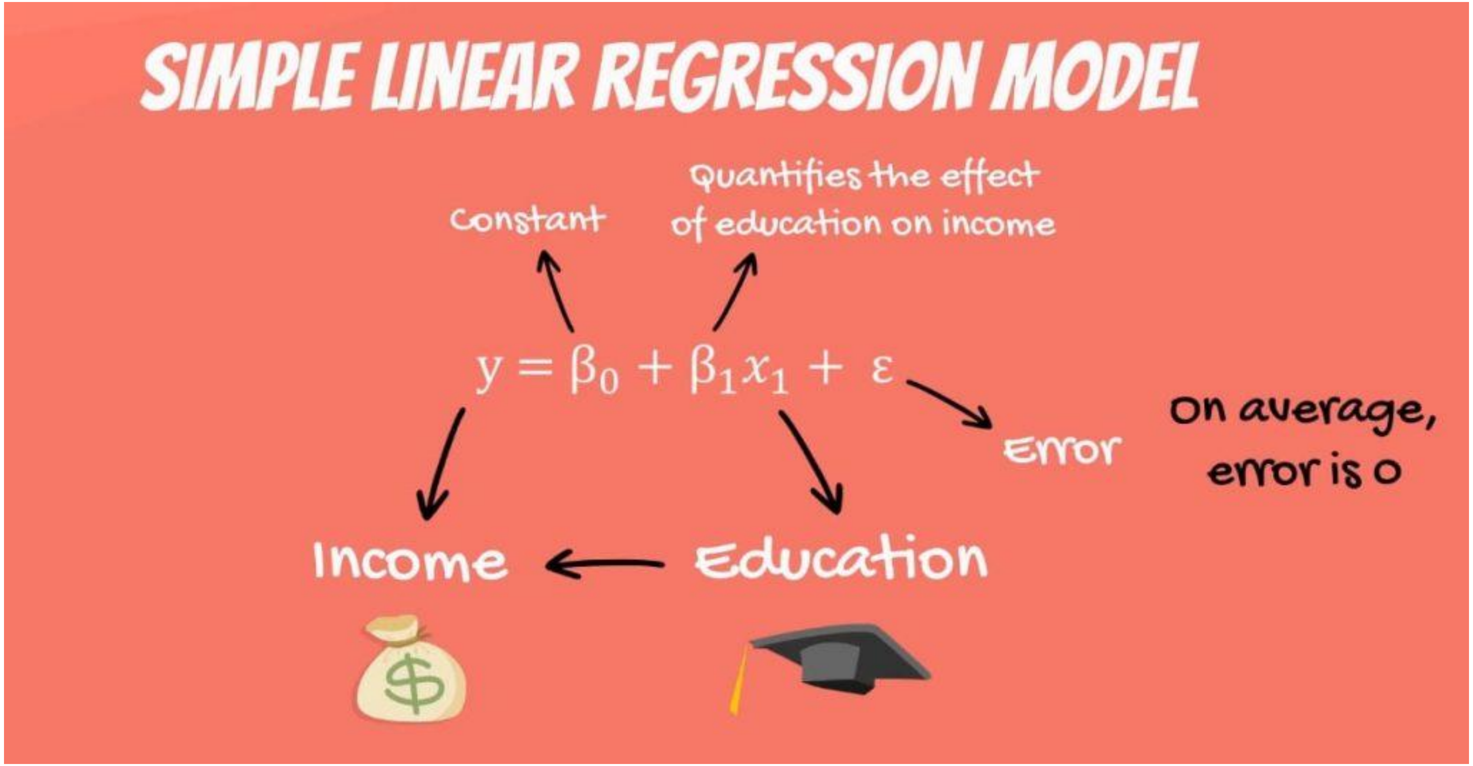
- The other two components are the constant β_0 and the error – epsilon(ε).
- In this **linear regression** example, you can think of the constant β_0 as the minimum wage.
- No matter your education, if you have a job, you will get the minimum wage.
- This is a guaranteed amount.



- So, if you never went to school and plug an education value of 0 years in the formula, what could possibly happen?
- Logically, the **regression** will predict that your income will be the minimum wage.

Epsilon

- The last term is the epsilon(ε).
- This represents the error of estimation.
- The error is the actual difference between the observed income and the income the **regression** predicted.
- On average, across all observations, the error is 0.



- If you earn more than what the **regression** has predicted, then someone earns less than what **regression** predicted.
- Everything evens out.

The Linear Regression Equation

- The original formula was written with Greek letters.
- This tells us that it was the population formula.
- But don't forget that statistics (and data science) is all about sample data.
- In practice, we tend to use the **linear regression *equation***.
- It is simply $\hat{y} = \beta_0 + \beta_1 * x$.

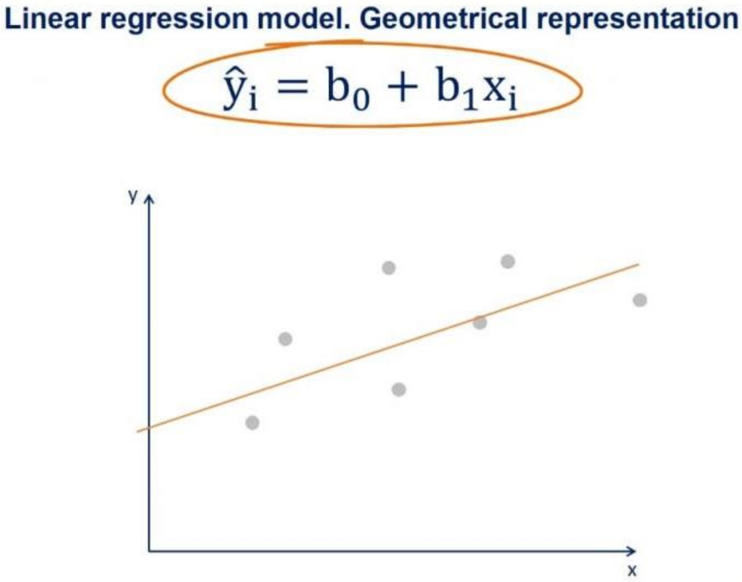
SIMPLE LINEAR REGRESSION EQUATION

$$\hat{y} = b_0 + b_1 x_1$$

- The $\hat{y}$ here is referred to as *y hat*.
- Whenever we have a hat symbol, it is an estimated or predicted value.
- B_0 is the estimate of the **regression** constant β_0. Whereas, b_1 is the estimate of β_1, and x is the sample data for the *independent variable*.

The Regression Line

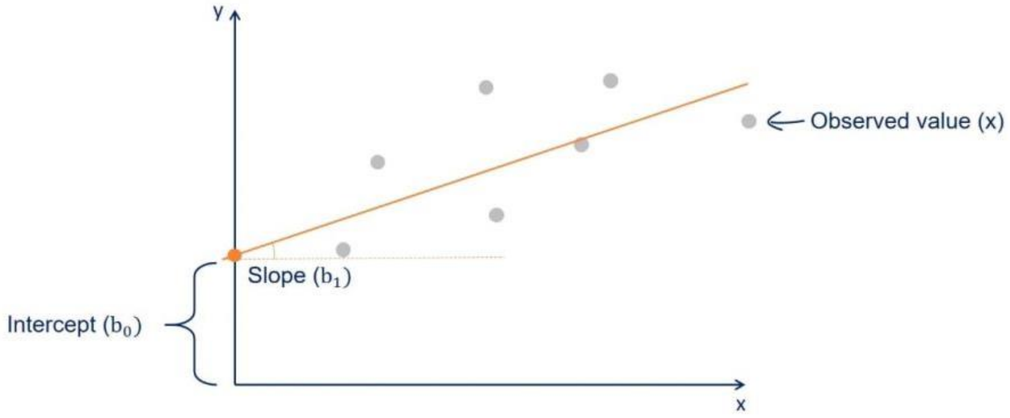
- When we plot the data points on an *x-y* plane, the **regression line** is the best-fitting line through the data points.
- We plot the line based on the **regression equation**.



- The grey points that are scattered are the observed values.
- B_0, as we said earlier, is a *constant* and is the intercept of the **regression line** with the y-axis.
- B_1 is the slope of the **regression line**.
- It shows how much *y* changes for each unit change of *x*.

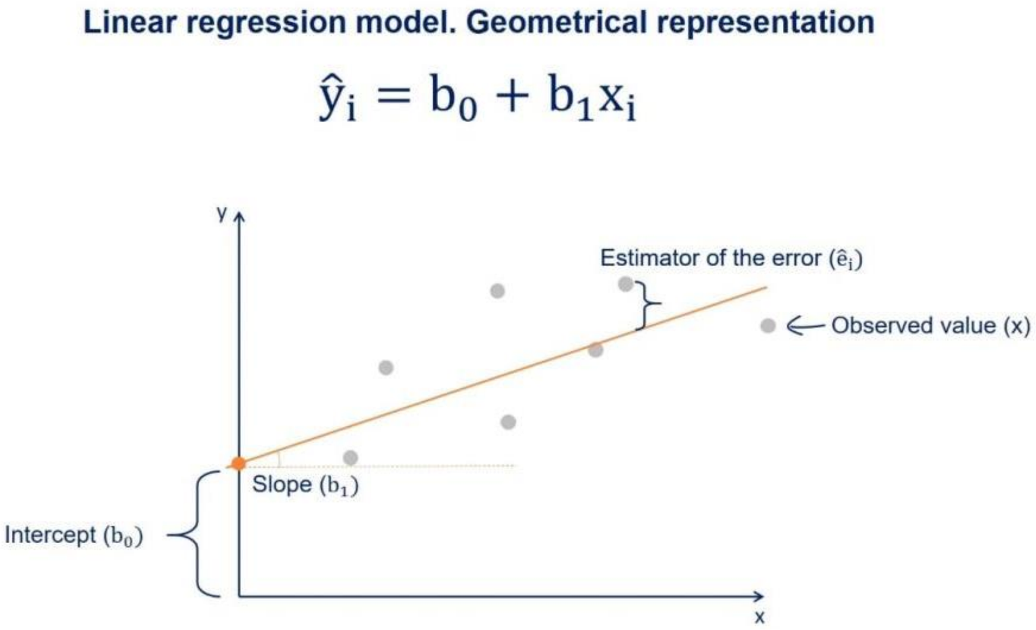
Linear regression model. Geometrical representation

$$\hat{y}_i = b_0 + b_1 x_i$$

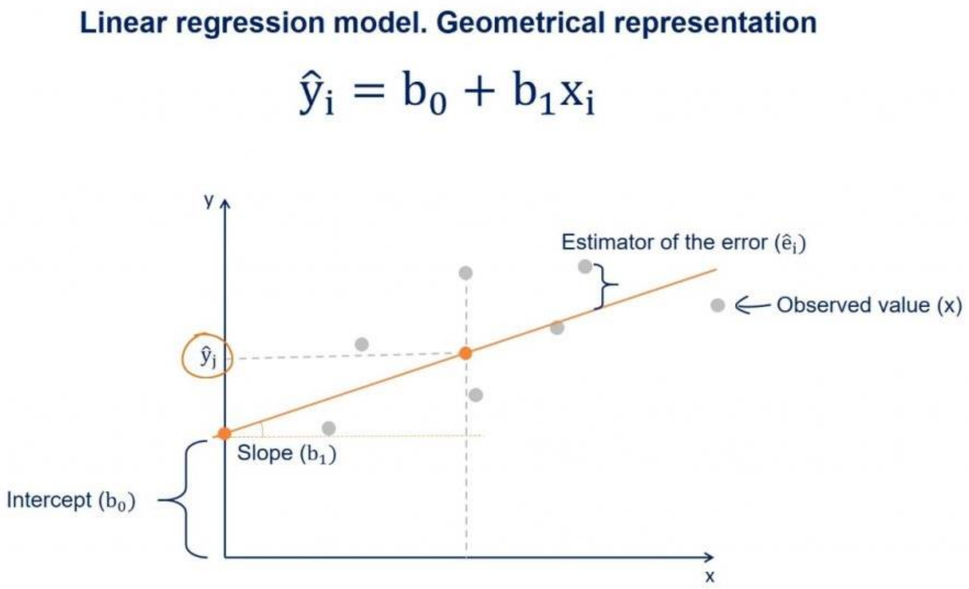


The Estimator of the Error

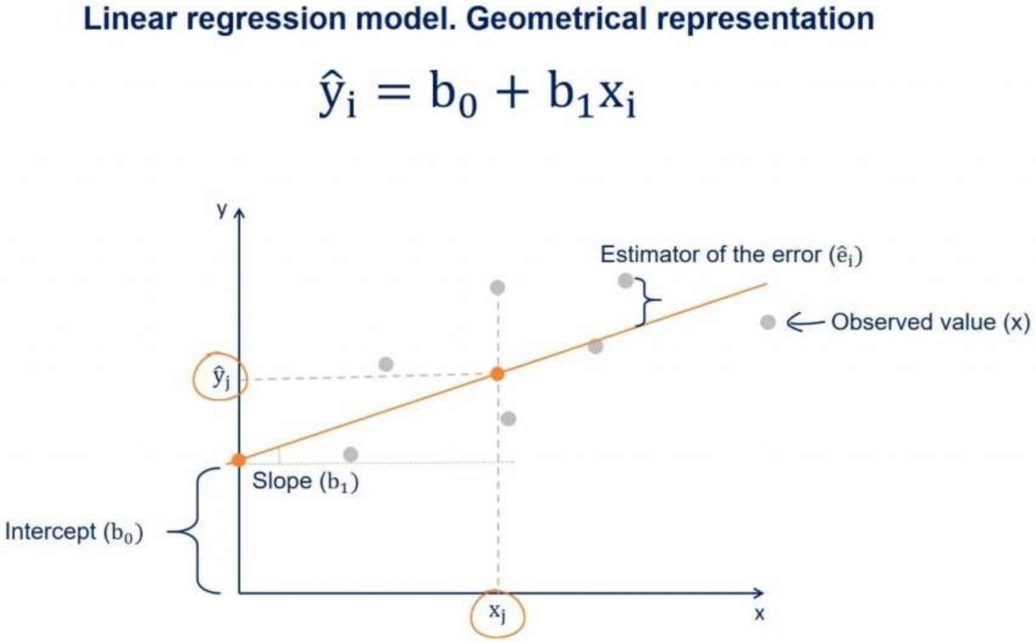
- The distance between the observed values and the **regression line** is the *estimator of the error term epsilon*.
- Its **point estimate** is called residual.



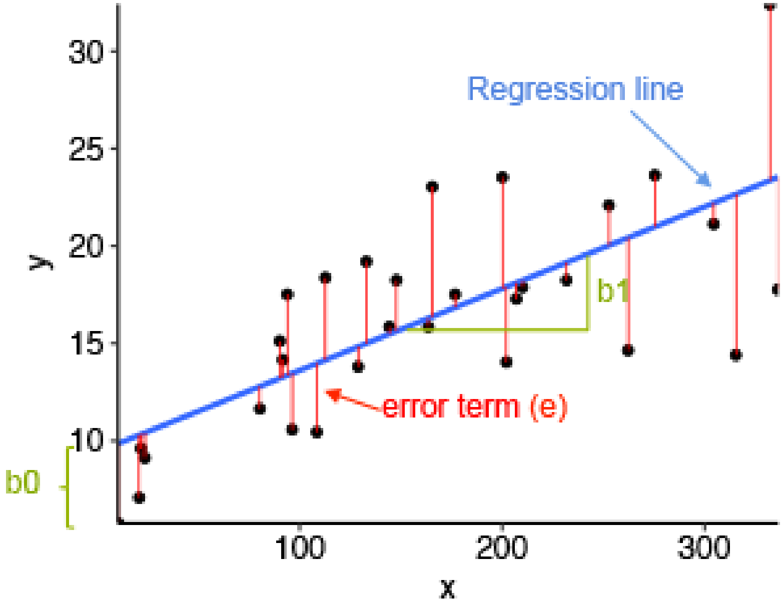
- Now, suppose we draw a perpendicular from an observed point to the **regression line**.
- The intercept between that perpendicular and the **regression line** will be a point with a *y* value equal to $\hat{y}$.



- As we said earlier, given an x, $\hat{y}$ is the value predicted by the **regression line**.



- The figure below illustrates the linear regression model, where:
- the best-fit regression line is in blue
- the intercept (b0) and the slope (b1) are shown in green
- the error terms (e) are represented by vertical red lines



Simple Linear Regression

- From the scatter plot above, it can be seen that not all the data points fall exactly on the fitted regression line.
- Some of the points are above the blue curve and some are below it; overall, the residual errors (e) have approximately mean zero.
- The sum of the squares of the residual errors are called the **Residual Sum of Squares** or **RSS**.
- The average variation of points around the fitted regression line is called the **Residual Standard Error (RSE)**. This is one the metrics used to evaluate the overall quality of the fitted regression model. The lower the RSE, the better it is.

Simple Linear Regression

- Since the mean error term is zero, the outcome variable y can be approximately estimated as follow:
- y ~ b0 + b1*x
- Mathematically, the beta coefficients (b0 and b1) are determined so that the RSS is as minimal as possible.
- This method of determining the beta coefficients is technically called **least squares** regression or **ordinary least squares** (OLS) regression.
- Once, the beta coefficients are calculated, a t-test is performed to check whether or not these coefficients are significantly different from zero.
- A non-zero beta coefficients means that there is a significant relationship between the predictors (x) and the outcome variable (y).

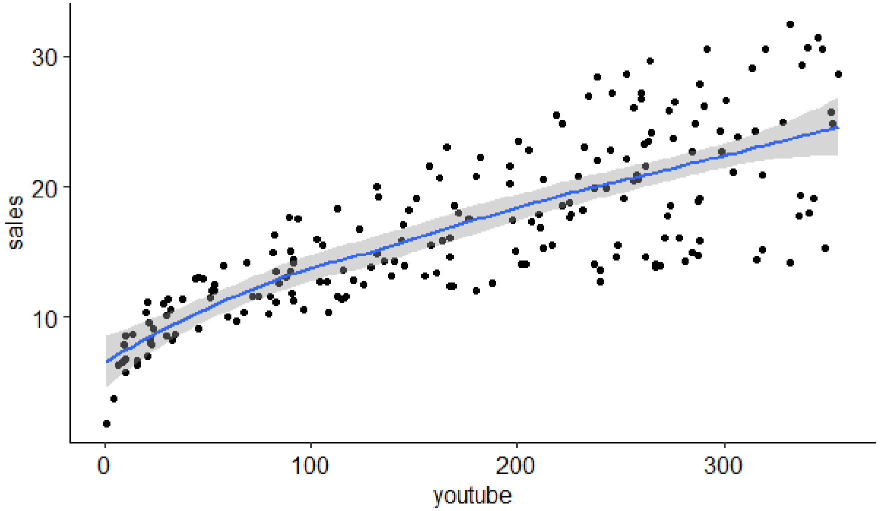
Simple Linear Regression in R

- Consider that, we want to evaluate the impact of advertising budgets of three medias (youtube, facebook and newspaper) on future sales.
- We’ll use the marketing data set [datarium package]. It contains the impact of three advertising medias (youtube, facebook and newspaper) on sales.
- Data are the advertising budget in thousands of dollars along with the sales.
- The advertising experiment has been repeated 200 times with different budgets and the observed sales have been recorded.

SLR using R

- **#loading packages**
- **library**(tidyverse)
- **library**(ggpubr)
- theme_set(theme_pubr())

- The graph above suggests a linearly increasing relationship between the sales and the youtube variables.
- This is a good thing, because, one important assumption of the linear regression is that the relationship between the outcome and predictor variables is linear and additive.
- It's also possible to compute the correlation coefficient between the two variables using the R function cor():



- The correlation coefficient measures the level of the association between two variables x and y.
- Its value ranges between -1 (perfect negative correlation: when x increases, y decreases) and +1 (perfect positive correlation: when x increases, y increases).
- A value closer to 0 suggests a weak relationship between the variables.
- A low correlation (-0.2 < x < 0.2) probably suggests that much of variation of the outcome variable (y) is not explained by the predictor (x).
- In such case, we should probably look for better predictor variables.
- In our example, the correlation coefficient is large enough, so we can continue by building a linear model of y as a function of x.

Computation

- The simple linear regression tries to find the best line to predict sales on the basis of youtube advertising budget.
- The linear model equation can be written as follow: sales = b0 + b1 * youtube
- The R function lm() can be used to determine the beta coefficients of the linear model:
- model <- lm(sales ~ youtube, data = marketing)
- model

Interpretation

- From the output above:
- the estimated regression line equation can be written as follow: sales = 8.44 + 0.048*youtube
- the intercept (b0) is 8.44. It can be interpreted as the predicted sales unit for a zero youtube advertising budget. Recall that, we are operating in units of thousand dollars. This means that, for a youtube advertising budget equal zero, we can expect a sale of 8.44 *1000 = 8440 dollars.
- the regression beta coefficient for the variable youtube (b1), also known as the slope, is 0.048. This means that, for a youtube advertising budget equal to 1000 dollars, we can expect an increase of 48 units (0.048*1000) in sales. That is, sales = 8.44 + 0.048*1000 = 56.44 units. As we are operating in units of thousand dollars, this represents a sale of 56440 dollars.

Regression line

- To add the regression line onto the scatter plot, you can use the function stat_smooth() [ggplot2].
- By default, the fitted line is presented with confidence interval around it.
- The confidence bands reflect the uncertainty about the line.
- If you don't want to display it, specify the option se = FALSE in the function stat_smooth().

- ggplot(marketing, aes(youtube, sales)) + geom_point() + stat_smooth(method = lm)

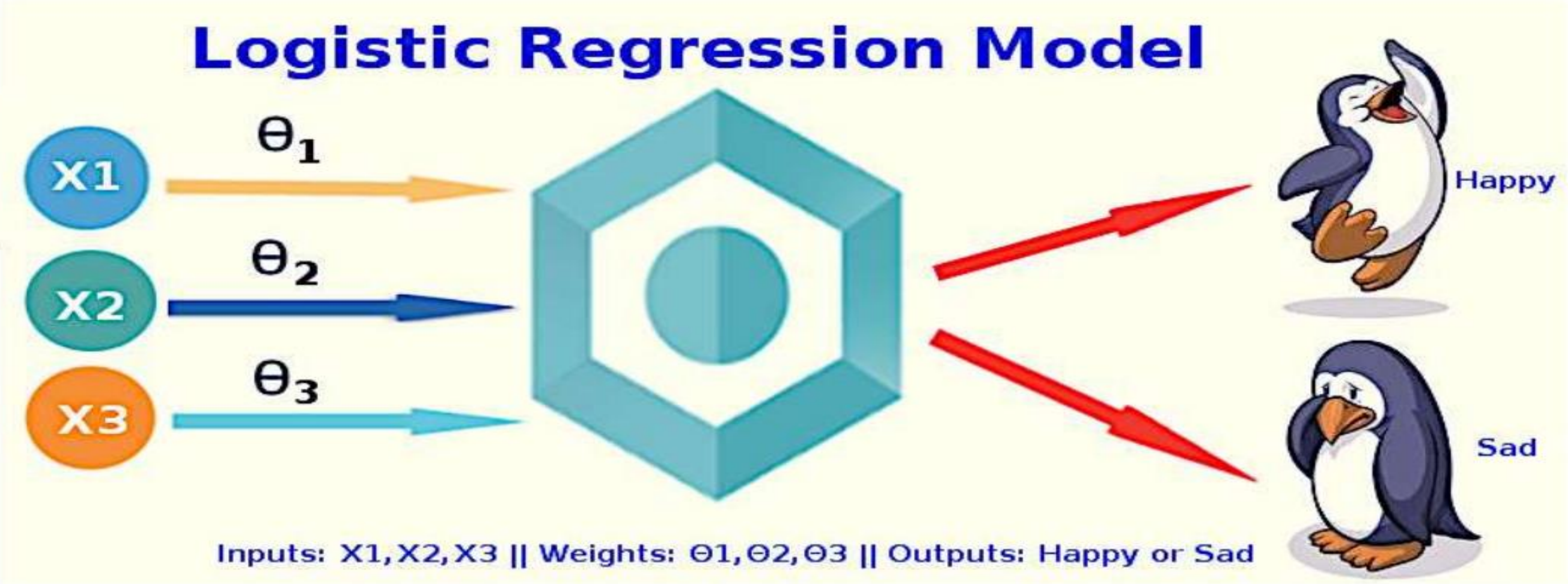
Model assessment

- In the previous section, we built a linear model of sales as a function of youtube advertising budget: sales = 8.44 + 0.048*youtube.
- Before using this formula to predict future sales, you should make sure that this model is statistically significant, that is:
- there is a statistically significant relationship between the predictor and the outcome variables
- the model that we built fits very well the data in our hand.
- In this section, we'll describe how to check the quality of a linear regression model.

Model summary

- The summary outputs shows 6 components, including:
- **Call**. Shows the function call used to compute the regression model.
- **Residuals**. Provide a quick view of the distribution of the residuals, which by definition have a mean zero. Therefore, the median should not be far from zero, and the minimum and maximum should be roughly equal in absolute value.
- **Coefficients**. Shows the regression beta coefficients and their statistical significance. Predictor variables, that are significantly associated to the outcome variable, are marked by stars.
- **Residual standard error** (RSE), **R-squared** (R2) and the **F-statistic** are metrics that are used to check how well the model fits to our data.

Logistic Regression



- Logistics Regression is used when the dependent variable is categorical.
- The values are strictly in the range of 0 and 1.
- It is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio- level independent variables.

Logistic Regression

- Linear regression predicts the numerical response but is not suitable for predicting the categorical variables.
- When categorical variables are involved, it is called classification problem.
- Logistic regression is suitable for binary classification problem.
- Logistic Regression is a supervised classification model.

Logistic Regression

- Classification is a task where we assign a label to an input based on certain features.
- Is the mail spam or not? The answer is Yes or No. Thus, categorical dependent variable is a binary response of Yes or No.
- If the student should be admitted or not is based on entrance examination marks.

Logistic Regression

- Logistic Regression is a supervised learning algorithm used for binary classification tasks, meaning we're trying to categorize instances into one of two classes.
- Despite its name, Logistic Regression is used for classification, not regression.
- The term "logistic" refers to the logistic function, also known as the sigmoid function, which is a key component of this algorithm.

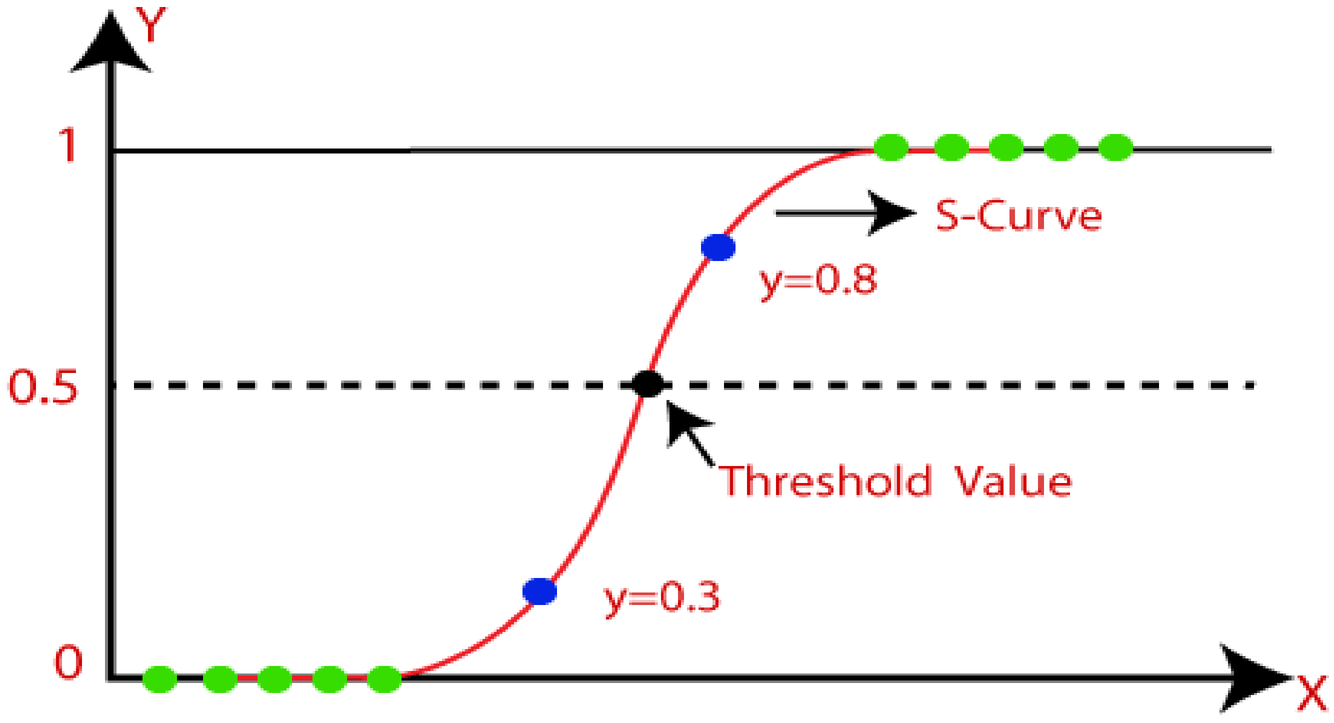
Sigmoid Function

- The sigmoid function takes any input and transforms it into a value between 0 and 1.
- This makes it perfect for transforming the output of a linear equation into a probability score.
- The sigmoid function is defined as:

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

- Here, **z** is the linear combination of input features and their respective weights.

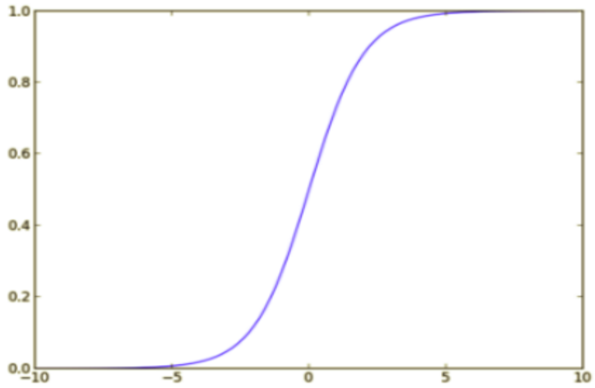
Logistic Regression



Sigmoid Function

Sigmoid Function

- The Sigmoid Function also called Logistics Function gives S shape that can take any real value and map into a value between 0 and 1.
- The range of the values is between 0 and 1.
- If the output of the sigmoid function is more than 0.5, we classify the outcome as 1 or Yes.
- If the output of the sigmoid function is less than 0.5, we classify the outcome as 0 or No.



$$f(x) = \frac{1}{1+e^{-(x)}}$$

Working

- Now, what if the organization wants to know whether an employee would get a promotion or not based on their performance?
- Linear graphs won’t be suitable in this case.
- We clip the line at zero or one and convert it into a sigmoid curve(s-curve)
- Based on the threshold values, the organization can decide whether an employee will get a salary increase or not.

Classification Model Evaluation

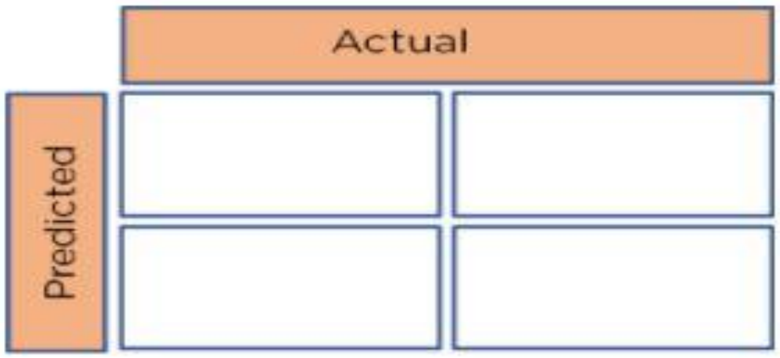
- Confusion Matrix
  - Accuracy Score
  - Precision Score
  - Recall Score
  - F1 Score
- Confusion Matrix is the tabular representation of Actual vs Predicted Values. It helps us find the accuracy of the model.
- Confusion Matrix is a Square Matrix
- E.g. 2 X 2 ; 3 X 3

Confusion Matrix

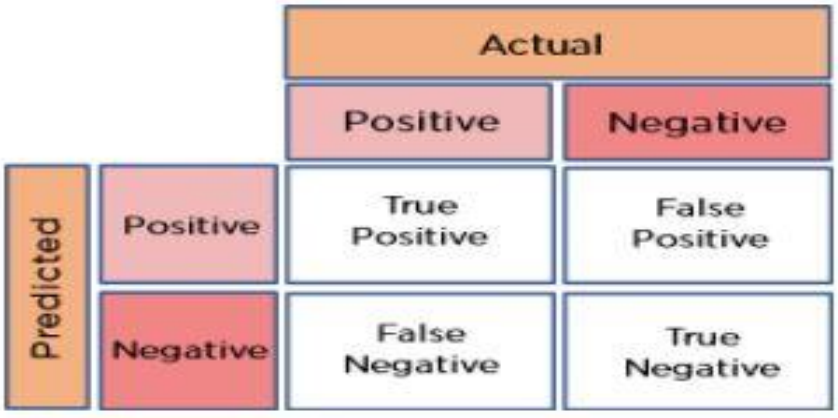
- The confusion matrix is a matrix used to determine the performance of the classification models for a given set of test data.
- It can only be determined if the true values for test data are known.
- Since it shows the errors in the model performance in the form of a matrix, hence also known as an **error matrix**.

Confusion Matrix

- A confusion matrix presents a table layout of the different outcomes of the prediction and results of a classification problem and helps visualize its outcomes.
- It plots a table of all th values of a classifier.



- We can obtain four different combinations from the predicted and actual values of a classifie



Features of Confusion Matrix

- For the 2 prediction classes of classifiers, the matrix is of 2*2 table, for 3 classes, it is 3*3 table, and so on.
- The matrix is divided into two dimensions, that are **predicted values** and **actual values** along with the total number of predictions.
- Predicted values are those values, which are predicted by the model, and actual values are the true values for the given observations.

Confusion Matrix Structure

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Confusion Matrix - Combinations

- True Positive: The number of times our actual **positive values are equal to the predicted positive**. You predicted a positive value, and it is correct.
- False Positive: The number of times our model **wrongly predicts positive values as negatives**. You predicted a negative value, and it is actually positive.
- True Negative: The number of times our **actual negative values are equal to predicted negative values**. You predicted a negative value, and it is actually negative.
- False Negative: The number of times our model **wrongly predicts negative values as positives**. You predicted a negative value, and it is actually positive.

Confusion Matrix Metrics

- Consider a confusion matrix made for a classifier that classifies people based on whether they speak English or Spanish.
- From the diagram, we can see that:
- True Positives (TP) = 86
- True Negatives (TN) = 79
- False Positives (FP) = 12
- False Negatives (FN) = 10

	English Speaker	Spanish Speaker
English Speaker	86	12
Spanish Speaker	10	79

- Just from looking at the matrix, the performance of our model is not very clear.
- To find how accurate our model is, we use the following metrics:
  - Accuracy
  - Precision
  - Recall
  - F1-Score

- Accuracy: The accuracy is used to find the portion of correctly classified values. It tells us how often our classifier is right.
- It is the sum of all true values divided by total values.
- In this case:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

- Accuracy = (86 +79) / (86 + 79 + 12 + 10) = 0.8823 = 88.23%

- Precision: Precision is used to calculate the model's ability to classify positive values correctly.
- It is the true positives divided by the total number of predicted positive values.

- In this case,

$$\text{Precision} = \frac{TP}{TP + FP}$$

- Precision = 86 / (86 + 12) = 0.8775 = 87.75%

- Recall: It is used to calculate the model's ability to predict positive values.
- "How often does the model predict the correct positive values?".
- It is the true positives divided by the total number of actual positive values.

- In this case,
- Recall = 86 / (86 + 10) = 0.8983

$$\text{Recall} = \frac{TP}{TP + FN}$$

- F1-Score: It is the harmonic mean of Recall and Precision.
- t is useful when you need to take both Precision and Recall into account.

$$\text{F1-Score} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

- In this case,
- F1-Score = (2* 0.8775 * 0.8983) / (0.8775 + 0.8983) = 0.8877 = 88.77%

References

- https://365datascience.com/tutorials/python-tutorials/linear-regression/
- http://www.sthda.com/english/articles/40-regression-analysis/167-simple-linear-regression-in-r/
- https://statsandr.com/blog/multiple-linear-regression-made-simple/#simple-linear-regression-reminder
- https://statsandr.com/blog/binary-logistic-regression-in-r/