An Introduction to Linear Regression Models

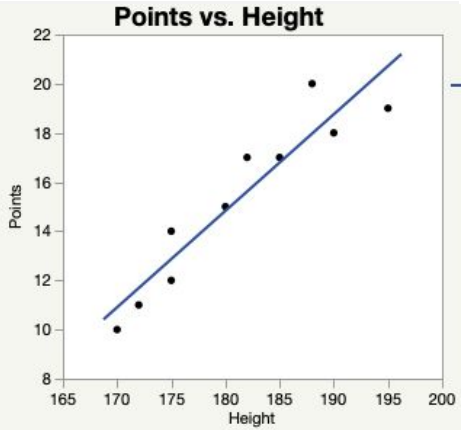
Please read through this introductory tutorial carefully, and pay attention to the concepts highlighted.

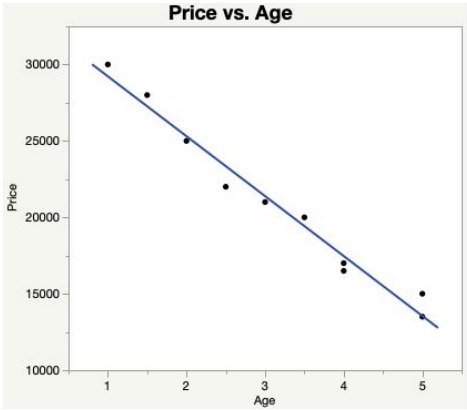
You will be asked to apply these concepts to complete the study tasks.

Feel free to ask us any questions if you find something confusing or unclear.

Linear regression is commonly used to find relationships between variables

For example, consider a **basketball** coach selecting players.
Here, ***height*** (input variable) is commonly considered to *linearly* influence ***number of points*** (output variable) scored during games.





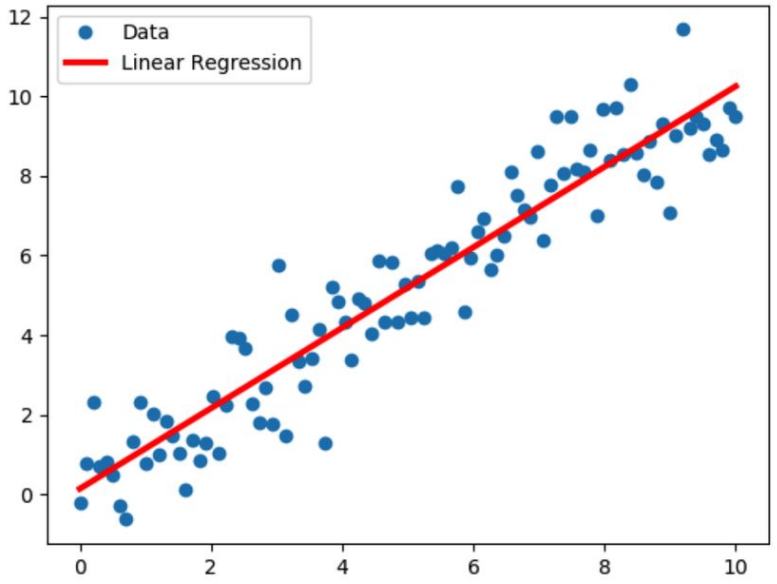
As a second example, consider a **second-hand car dealership.**
Here, the ***age*** (input) of the car typically has a *negative linear* effect on the **selling price** (output)**.**

Input Variables in Linear Regression

Quite often, there can be **multiple input variables** (X1, X2,...,Xn) that have an influence on an output variable (Y).
For example, both ***age*** and ***engine capacity*** can determine the ***selling price*** of a car.

In **linear regression**, this relationship is a linear combination:

$$Y = a1\ X1 + a2\ X2 + \ldots + an\ Xn$$



Here: Y = a X + b

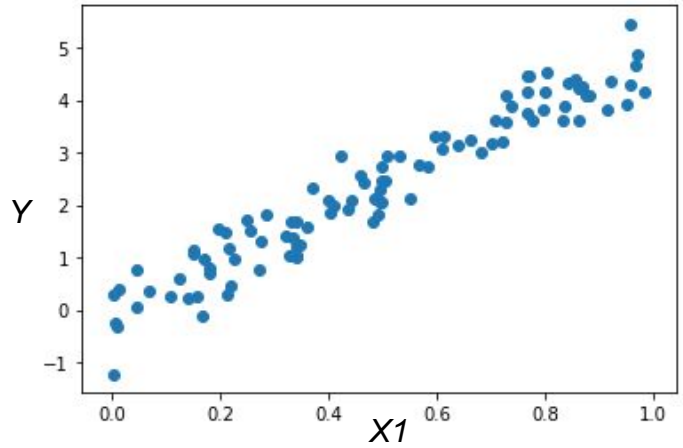
Including and Excluding Variables

Given several **input variables** (X1, X2,...,Xn), quite often **not all** input variables are related to the output variable (Y).
For example, the **colour** of a car might not determine the **selling price.**

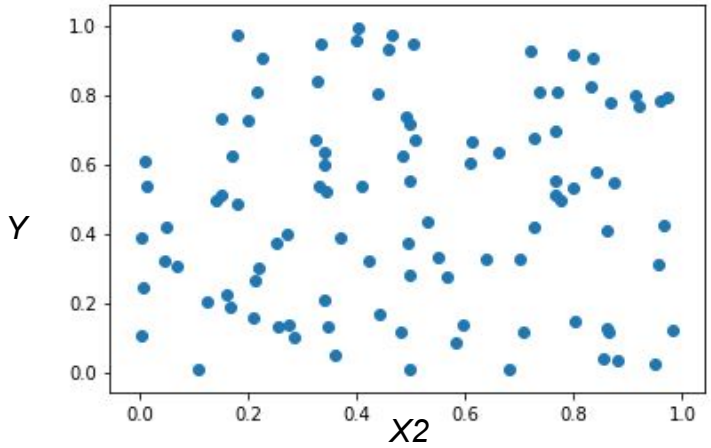
Here, there is **no correlation** between colour and selling price. In a linear regression model, input variables that are not correlated to the output variable should **not be included** in the model.

Example: Including and Excluding Variables

For instance in this example:



X1: correlated to Y



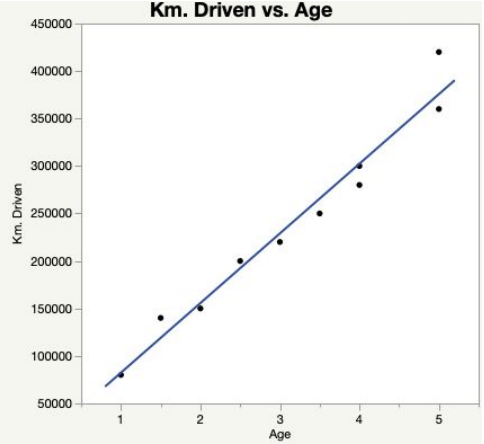
X2: not correlated to Y

It is clear that **X2 is not correlated** to Y and **should not be included** into the regression model. However, **X1 is correlated and should be included**.

Collinearity in Input Variables

Quite often, **two input variables** (Xi, Xj) can be **correlated** to each other.

For example, the ***age*** of a car may be directly correlated to the ***number of kilometres driven***.



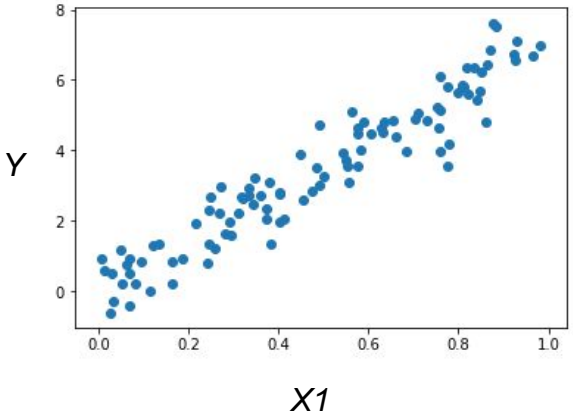
Models with Correlated Input Variables

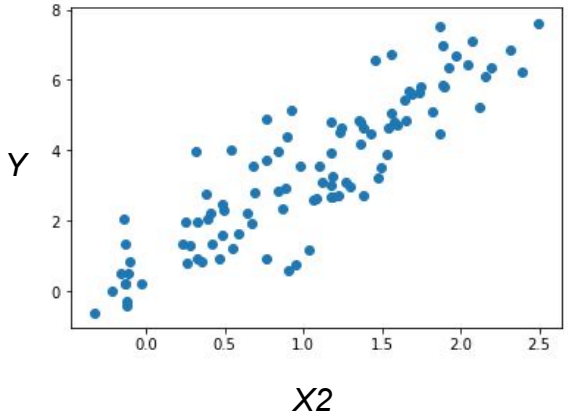
If all input variables are independent (uncorrelated) of each other, then the task is simple: all input variables correlated to the output Y have to be included in the regression model.

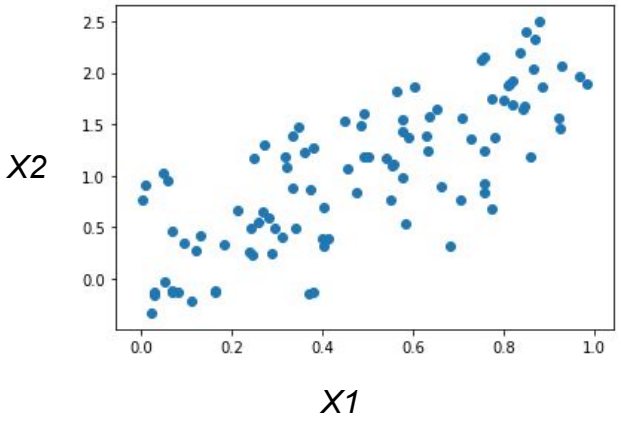
However, **if two input variables Xi and Xj are correlated** to each other, but also to the output Y, it is **better to include only one of the two** into the linear regression model. *(Indeed, the necessary information is already included, and having both might weaken the system.)*

An Example Case

For instance, in this example:







X1 and X2 are correlated to the output Y, so both should be included?

In fact **no**: X1 and X2 are also correlated, so it is better to include only one!

Your task

During the study, you will complete a series of tasks where **you will construct linear regression models.**

With the help of the assistant, your task will be to choose which variables to include to the linear regression. The **assistant will display relevant information** about input and output variables (as graphs), and will ask you to either **include or exclude input variables from the model**.

Before we start the task, let’s take some time for a **short preliminary test**.