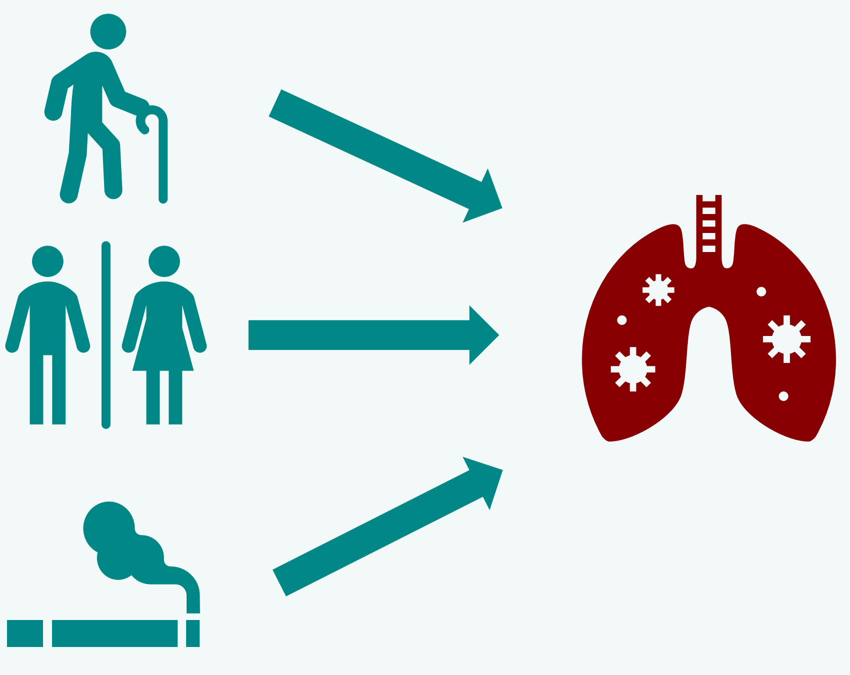
Logistic Regression

Playbook

1. Theory
2. Example
3. Interpretation

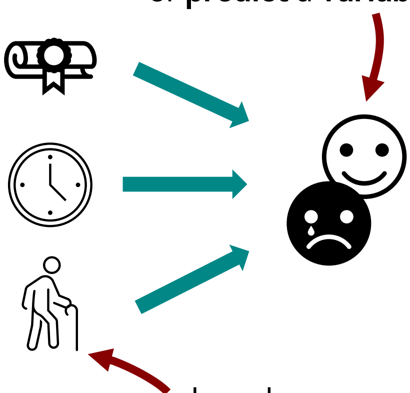


Author: Dr. Mathias Jesussek

What is a **regression**?

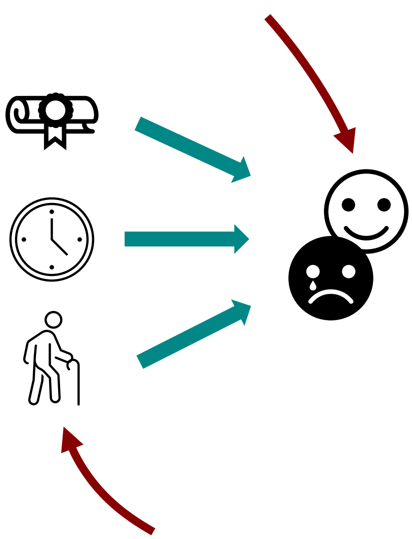
A regression analysis is a method for **modeling relationships** between **variables**.

It makes it possible to **infer** or **predict** a **variable**



based **on one or more other variables**.

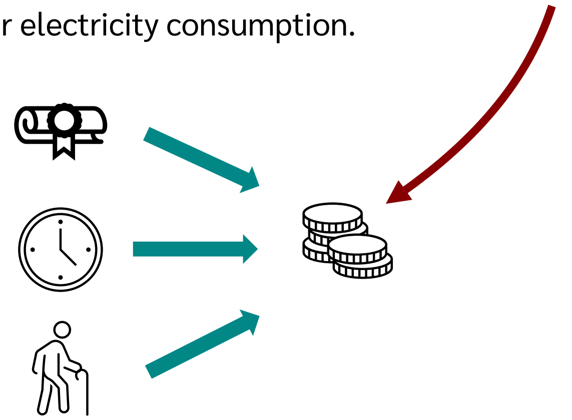
The variable we want to infer or predict is called the **dependent variable** or **criterion**.



The variables we use for prediction are called **independent variables** or **predictors**.

What is the **difference** between a **linear regression** and a **logistic regression?**

In a **linear regression**, the dependent variable is a **metric variable**, e.g. salary or electricity consumption.



In a **logistic regression**, the dependent variable is a **dichotomous variable**.

What is a dichotomous variable?

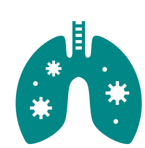
Dichotomous variables are variables with only **two values.**

For example:

Whether a person **buys** or does **not buy** a particular product

or

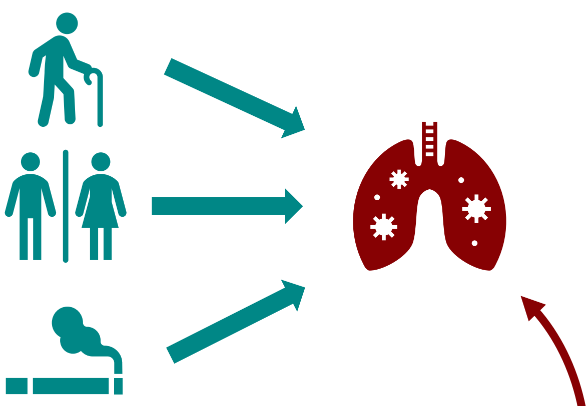
whether a disease is **present** or **not**



DATAtab

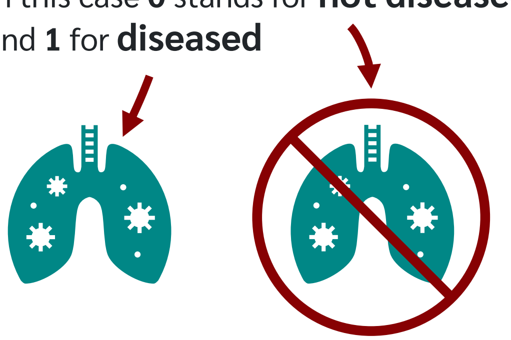
How can logistic regression be used ?

With the help of **logistic regression**, we can determine what has an influence on whether a certain **disease is present or not**.



We could study the influence of **age**, **gender** and **smoking status** on that particular **disease**.

In this case **0** stands for **not diseased** and **1** for **diseased**



and the **probability** for the **occurrence of the characteristic 1 (=characteristic present)** is estimated.

Our **data set** might look like this:

Here we have the **independent variables**

Age	Gender	Smoker status	Disease
22	female	Non-smoker	1
25	female	Smoker	1
18	male	Smoker	0
45	male	Non-smoker	0
12	female	Smoker	0
43	male	Smoker	1
23	male	Smoker	0
33	male	Smoker	1
...	...	...	...

and here the **dependent variable** with 0 and 1.

We could now investigate what influence the **independent variables** have on the disease.

If there is an influence, then we can **predict** how **likely** a person is to have a certain disease.

Now, of course, the question arises:

Why do we need **logistic regression** in this case?

Why can't we just use **linear regression?**

A quick recap:

In **linear regression**, this is our **regression equation**:

$$\hat{y} = b_1 \cdot x_1 + b_2 \cdot x_2 + \ldots + b_k \cdot x_k + a$$

We have the dependent variable

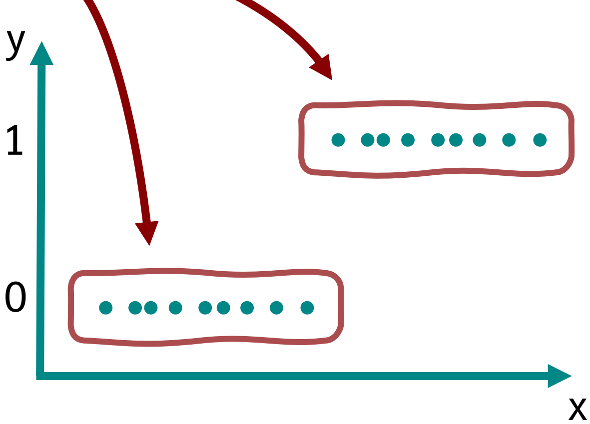
the independent variables

$$\boxed{\hat{y}} = \boxed{b_1} \cdot \boxed{x_1} + \boxed{b_2} \cdot \boxed{x_2} + \ldots + \boxed{b_k} \cdot \boxed{x_k} + \boxed{a}$$

and the **regression coefficients.**

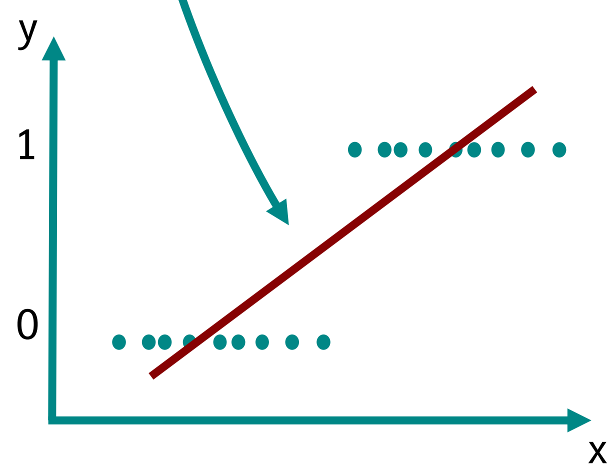
$$\boxed{\hat{y}} = b_1 \cdot x_1 + b_2 \cdot x_2 + \ldots + b_k \cdot x_k + a$$

However, we now have a **dependent variable** that is either **0** or **1**.

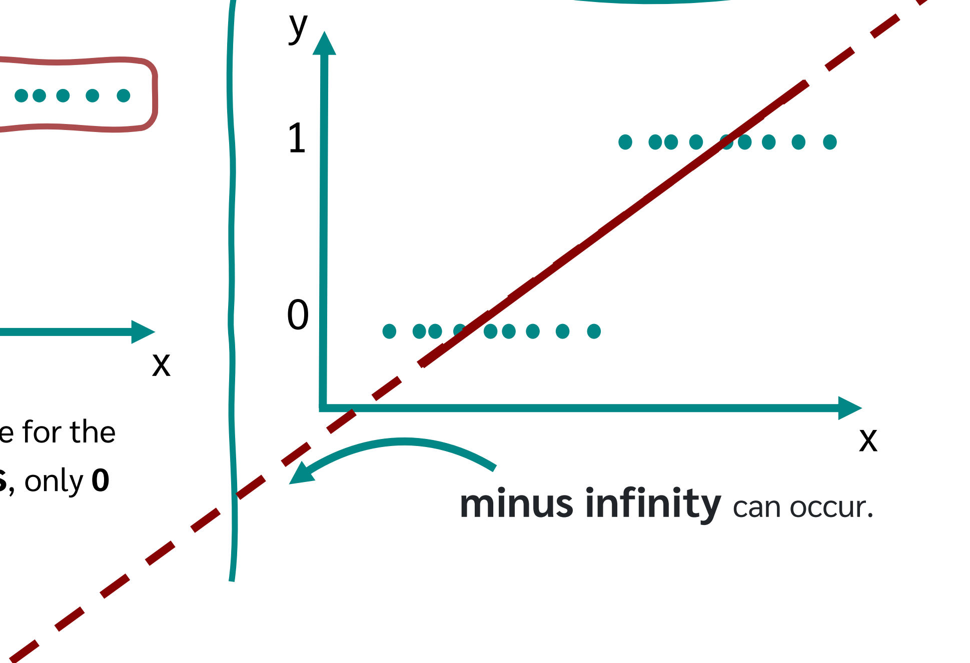


No matter which value we have for the **independent variables**, only **0** or **1** results.

A **linear regression** would now simply put a **straight line** through the **points**.



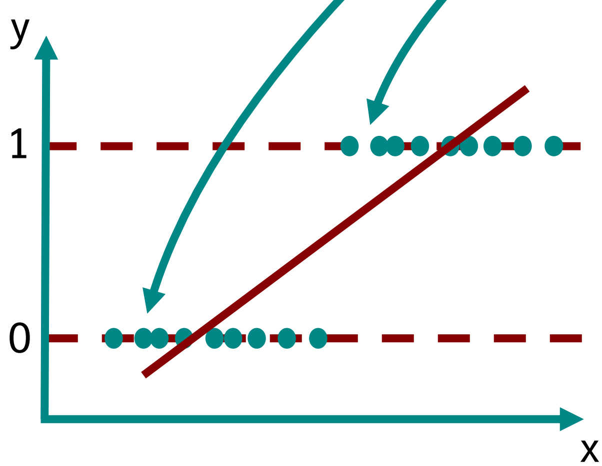
We can now see, that in the case of **linear regression**, values between **plus and**



minus infinity can occur.

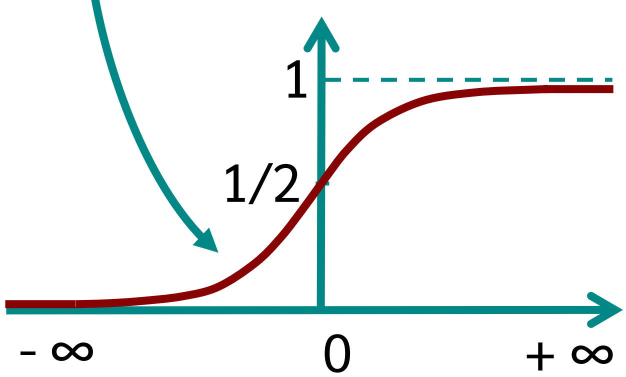
However, the **goal** of **logistic regression** is to estimate the **probability** of occurrence.

The value range for the prediction should therefore be between **0** and **1**.

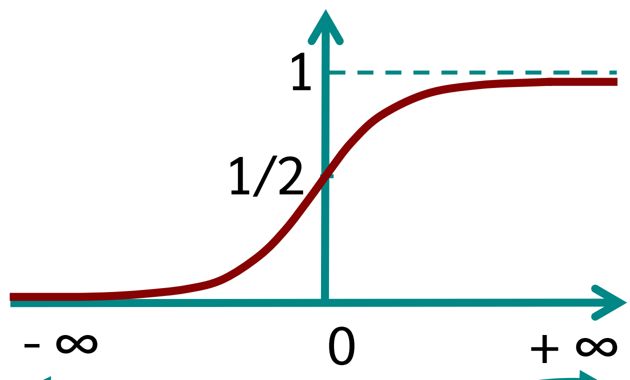


So we need a **function** that only takes values between **0** and **1**!

And that is exactly what the **logistic function** does.



No matter where we are on the **x-axis,**



between **minus** and **plus infinity** only values between **0** and **1** result.

And that is exactly what we want!

The **equation** for the **logistic function** looks like this:

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

The **logistic function** is now used by the logistic regression.

For **z**, the equation of the **linear regression** is now simply inserted.

$$\underbrace{\hat{y} = b_1 \cdot x_1 + b_2 \cdot x_2 + \ldots + b_k \cdot x_k + a}$$

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

This gives us this **equation**:

$$f(z) = \frac{1}{1 + e^{-(b_1 \cdot x_1 + \ldots + b_k \cdot x_k + a)}}$$

Thus, the **probability** that the dependent variable is **1** is given by:

$$P(y = 1 | x_1, \ldots, x_n) = \frac{1}{1 + e^{-(b_1 \cdot x_1 + \ldots + b_k \cdot x_k + a)}}$$

What does this look like for our **example** ?

In our example,
the **probability** of having a **certain disease**

$$P(is\ diseased) = \frac{1}{1 + e^{-(b_1 \cdot Age + b_2 \cdot Male + b_3 \cdot Smoker + a)}}$$

is a function of **age**, **gender** and **smoking status**.

For **z,** the equation of the **linear regression** is now simply inserted.

$$\hat{y} = \underbrace{b_1 \cdot x_1 + b_2 \cdot x_2 + \ldots + b_k \cdot x_k + a}$$

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

This gives us this **equation**:

$$f(z) = \frac{1}{1 + e^{-(b_1 \cdot x_1 + \ldots + b_k \cdot x_k + a)}}$$

Thus, the **probability** that the dependent variable is **1** is given by:

$$P(y = 1 | x_1, \ldots, x_n) = \frac{1}{1 + e^{-(b_1 \cdot x_1 + \ldots + b_k \cdot x_k + a)}}$$

What does this look like for our **example** ?

In our example,
the **probability** of having a **certain disease**

$$P(is\ diseased) = \frac{1}{1 + e^{-(b_1 \cdot Age + b_2 \cdot Male + b_3 \cdot Smoker + a)}}$$

is a function of **age**, **gender** and **smoking status**.

$$P(is\ diseased) = \frac{1}{1 + e^{-(\boxed{b_1} \cdot Age + \boxed{b_2} \cdot Male + \boxed{b_3} \cdot Smoker + \boxed{a})}}$$

Now we need to determine the **coefficients** so that our model best represents the given data.

To solve this problem, the so-called **maximum likelihood method** is used.

For this purpose, there are good **numerical methods** that can solve the problem efficiently.

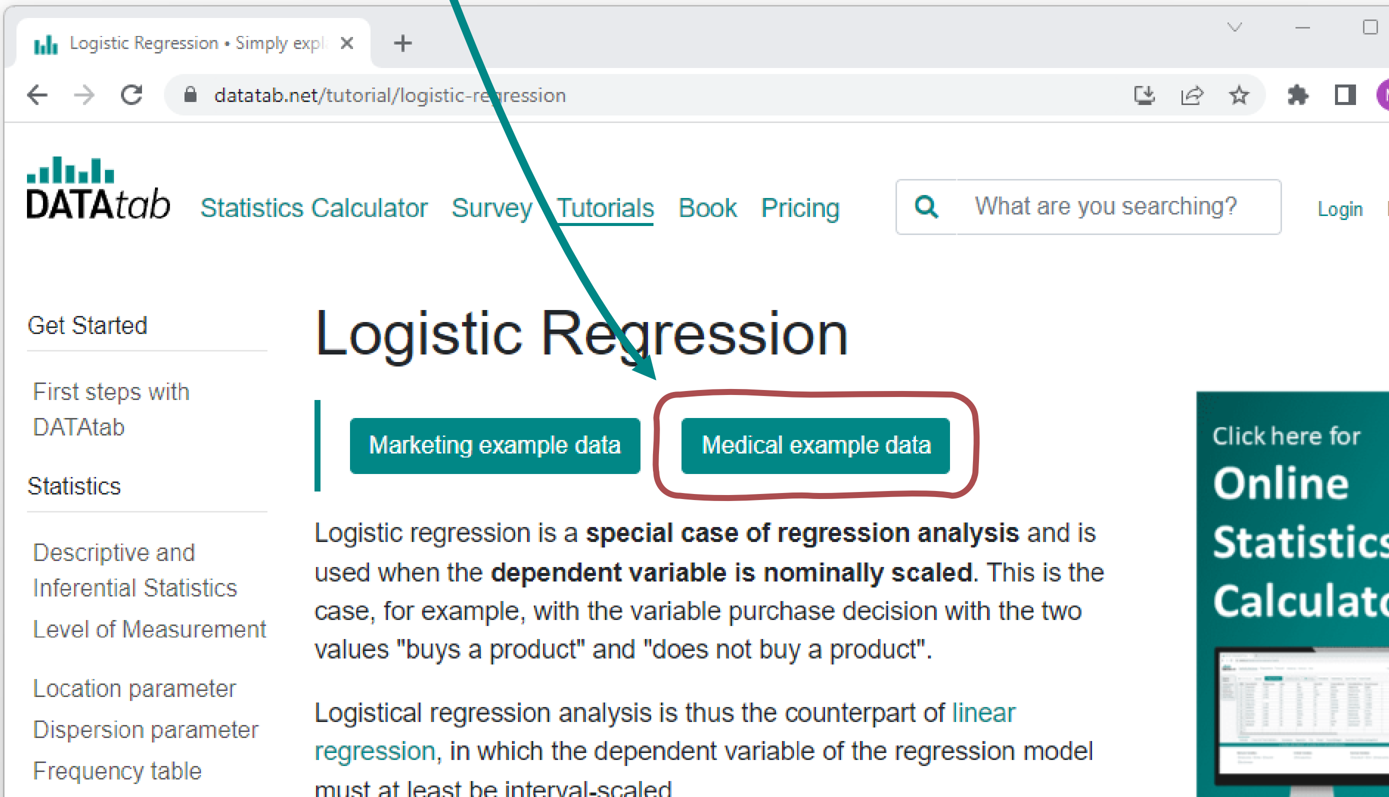
But how do you **interpret** the **results** of a **logistic regression** ?

Let's take a look at this fictitious **example**.

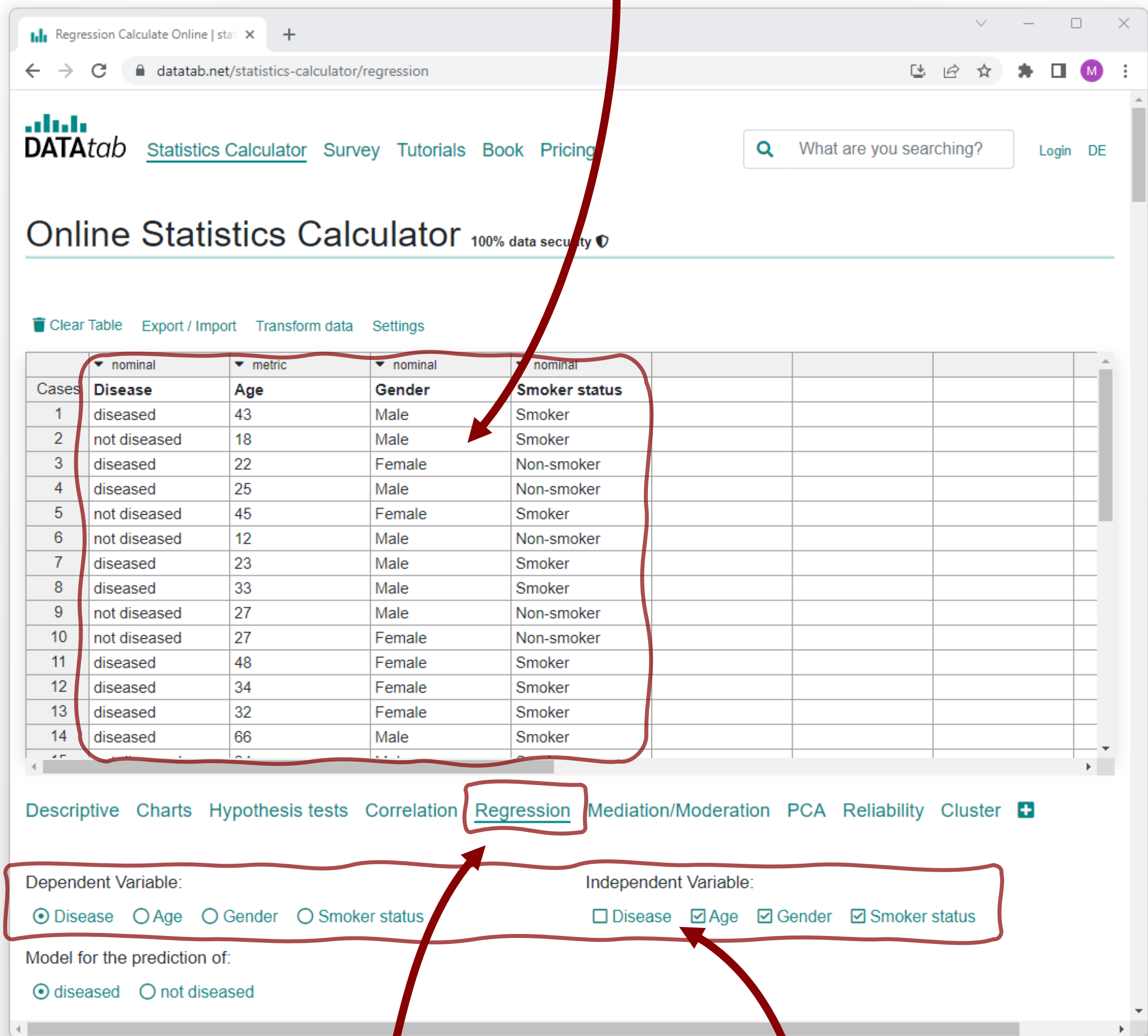
Age	Gender	Smoker status	Disease
22	female	Non-smoker	1
25	female	Smoker	1
18	male	Smoker	0
45	male	Non-smoker	0
12	female	Smoker	0
43	male	Smoker	1
23	male	Smoker	0
33	male	Smoker	1
...	...	...	...

If you like, you can download the **example dataset** for free and **follow the steps** in parallel. Please just use this **link**.

Or load it from the **logistic Regression tutorial**



When you **use** the **link**, the data is **automatically loaded**.



We want to **calculate a logistic regression**, so we just **click on regression**.

When we **copy** our **data** in here, the **variables show up** down here.

Depending on how your **dependent variable is scaled**, DATAtab will calculate either a **logistic** or a **linear regression** under the tab Regression.

We choose **disease** as the **dependent** variable and **age**, **gender**, and **smoking status** as the independent variables. Datatab now calculates a logistic regression for us.

DATAtab

If you don't know how to interpret the results, you can click on

Summary in words

We will now go through all the tables slowly and understandably. Let's start at the top.

Logistic Regression

Summary in words

Result

Copy Word Copy Excel

Total number of cases	Correct assignments	In percent
36	26	72.22 %

Classification table

Copy Word Copy Excel

		Predicted		
		not diseased	diseased	Correct
Observed	not diseased	11	5	68.75 %
	diseased	5	15	75 %
	Total			72.22 %

Chi-Squared Test

Copy Word Copy Excel

Chi2	df	p
8.79	3	.032

Model Summary

Copy Word Copy Excel

-2 Log-Likelihood	Cox & Snell R^2	Nagelkerke R^2	McFadden's R^2
40.67	0.22	0.29	0.18

Model

Copy Word Copy Excel

	Coefficient B	Standard error	z	p	Odds Ratio	95% conf. interval
Age	0.04	0.03	1.68	.092	1.04	0.99 - 1.1
Male	0.87	0.8	1.08	.28	2.39	0.49 - 11.55
Smoker	1.34	0.79	1.7	.089	3.81	0.82 - 17.76
Constant	-2.73	1.26	2.16	.03		

Prediction for your data

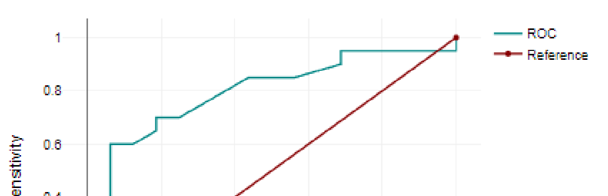
Copy Word Copy Excel

	Prediction
Age	
Male	
Smoker	
Probability	

ROC-Curve

Download png Download svg Settings

ROC Curve (AUC: 0.778)



DATAtab

Let's Start

The first thing that is displayed is the **results table**. In the **results table** you can see that a total of **36 people** were examined.

Result

Copy Word Copy Excel

Total number of cases	Correct assignments	In percent
36	26	72.22 %

With the help of the calculated **regression model**, **26 of 36 persons** could be correctly assigned. That is **72.22%**!

Then comes the **classification table**.

Here you can see how often the categories **not diseased** and **diseased** were **observed** and how often they were **predicted**.

Classification table

Copy Word Copy Excel

		Predicted		
		not diseased	diseased	Correct
Observed	not diseased	11	5	68.75 %
	diseased	5	15	75 %
	Total			72.22 %

In total, **"not diseased"** was observed **16** times.

Classification table

Copy Word Copy Excel

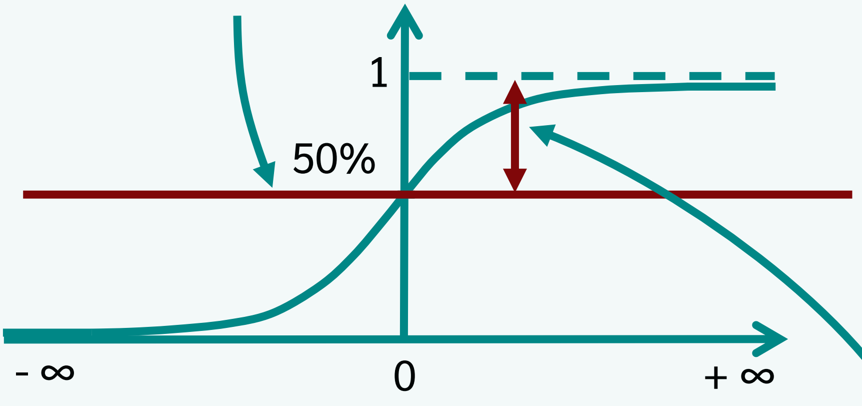
		Predicted		
		not diseased	diseased	Correct
Observed	not diseased	11	5	68.75 %
	diseased	5	15	75 %
	Total			72.22 %

Of these 16 individuals, the regression model **correctly** scored **11** as **not diseased** and **incorrectly** scored **5** as **diseased**.

Of the 20 diseased individuals, **15** were correctly scored as diseased and **5 incorrectly** scored as **diseased**.

To be noted:

For deciding whether a person is **diseased or not** the **threshold** of **50%** is used.



If the **regression model** estimates a value **greater than 50%**, this person is assigned **“diseased”**, otherwise **“not diseased”**.

DATAtab

Now comes the **Chi² test**.

Chi-Squared Test

Copy Word Copy Excel

Chi2	df	p
8.79	3	.032

Here we can read whether the **model** as a whole is **significant or not**.

Two models are compared for this purpose **!**

In one model **all independent variables are used**

$$\frac{1}{1+e^{-(\boxed{b_1 \cdot x_1 + \ldots + b_k \cdot x_k} + a)}} \qquad \frac{1}{1+e^{-(\cancel{b_1 \cdot x_1 + \ldots + b_k \cdot x_k} + a)}}$$

and in the other model the **independent variables are not used**.

With the help of the **Chi² test** we **compare** how good the **prediction** is when the **dependent variables** are **used** and how good it is when the **dependent variables are not used** and the **Chi² test** “tells us” if there is a **significant difference** between these two results.

The **null hypothesis** is that **both models are the same**.

If the **p-value** is less than 0.05, this **null hypothesis** is **rejected**.

Chi-Squared Test

Copy Word Copy Excel

Chi2	df	p
8.79	3	.032

In our example, the **p-value is less than 0.05** and we assume that there is a significant difference between the models. Thus, the model as a whole is **significant**.

Next comes the **model summary**.

In this table we see on the one hand the **-2 log likelihood value** and on the other hand we are given different **coefficients of determination R^2**.

Model Summary

Copy Word Copy Excel

-2 Log-Likelihood	Cox & Snell R^2	Nagelkerke R^2	McFadden's R^2
40.67	0.22	0.29	0.18

R^2 is used to find out how well the regression model explains the dependent variable. In a **linear regression**, the **R^2** indicates the proportion of the variance that can be explained by the independent variables. The more variance can be explained, the better the regression model.

However, in the case of **logistic regression**, the meaning is different and there are different ways to calculate the **R^2**. Unfortunately, there is also **no agreement** yet on which way is the **"best" way**.

DATAtab gives you the **R^2** according to
Cox and Snell, according to **Nagelkerke** and according to **McFadden**.

Model Summary

Copy Word Copy Excel

-2 Log-Likelihood	Cox & Snell R^2	Nagelkerke R^2	McFadden's R^2
40.67	0.22	0.29	0.18

DATAtab

And now comes the most **important table**.
The **table** with the **model coefficients**.

The most important parameters are
the **coefficient B**, the **p-value** and the **odds ratio**.

Model

Copy Word Copy Excel

	Coefficient B	Standard error	z	p	Odds Ratio	95% conf. interval
Age	0.04	0.03	1.68	.092	1.04	0.99 - 1.1
Male	0.87	0.8	1.08	.28	2.39	0.49 - 11.55
Smoker	1.34	0.79	1.7	.089	3.81	0.82 - 17.76
Constant	-2.73	1.26	2.16	.03		

Coefficients B

In the first column we can read the calculated **coefficients** from our model.

Model

Copy Word Copy Excel

	Coefficient B	Standard error	z	p	Odds Ratio	95% conf. interval
Age	0.04	0.03	1.68	.092	1.04	0.99 - 1.1
Male	0.87	0.8	1.08	.28	2.39	0.49 - 11.55
Smoker	1.34	0.79	1.7	.089	3.81	0.82 - 17.76
Constant	-2.73	1.26	2.16	.03		

We can insert these into the **regression equation**.

$$\frac{1}{1+e^{-\boxed{(b_1\cdot x_1+\ldots+b_k\cdot x_k+a)}}}$$

DATAtab

If we insert the **coefficients**, we get the following **regression equation**:

$$\frac{1}{1+e^{-(0.04\cdot Age+0.87\cdot Gender+1.34\cdot Smoker-2.73)}}$$

Model

Copy Word Copy Excel

	Coefficient B	Standard error	z	p	Odds Ratio	95% conf. interval
Age	0.04	0.03	1.68	.092	1.04	0.99 - 1.1
Male	0.87	0.8	1.09	.28	2.39	0.49 - 11.55
Smoker	1.34	0.79	1.7	.089	3.81	0.82 - 17.76
Constant	-2.73	1.26	2.16	.03		

With this we can now calculate the **probability** that a **person is diseased**.

$$P(is\ diseased)=\frac{1}{1+e^{-(b_1\cdot Age+b_2\cdot Male+b_3\cdot Smoker+a)}}$$

Example:

We want to know how likely a person who is **55 years old**, **female**, and **smoker** is to be diseased.

$$P(is\ diseased)=\frac{1}{1+e^{-(b_1\cdot Age+b_2\cdot Male+b_3\cdot Smoker+a)}}$$

We insert:

55 for the age

0, because the person is female

and **1**, as the person is a smoker.

This gives us 0.69 or 69%.

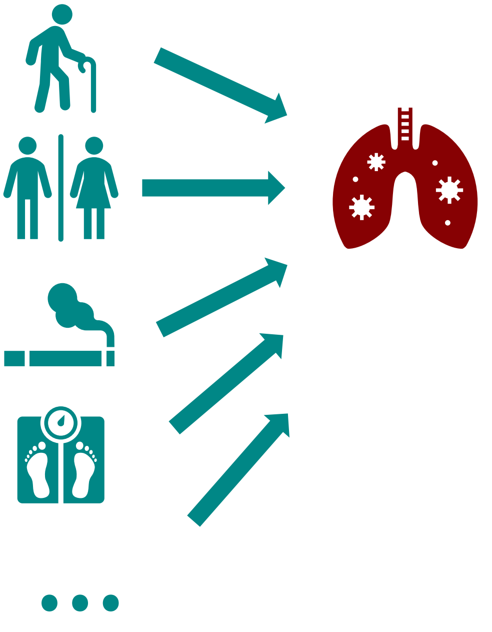
$$P(is\ diseased) = \frac{1}{1+e^{-(0.04\cdot55+0.87\cdot0+1.34\cdot1-2.73)}} = 0.69$$

Thus, it is **69% likely** that a 55-year-old female smoker is diseased.

Based on this **prediction**, it could now be decided whether to do another extensive investigation.

The example is **purely fictitious**.

In reality, there would certainly be many **other** and **different independent variables**.



But now back to the table!

In this **column** we can read whether the **coefficient** is **significantly different** from **zero**.

Model

Copy Word Copy Excel

	Coefficient B	Standard error	z	p	Odds Ratio	95% conf. interval
Age	0.04	0.03	1.68	.092	1.04	0.99 - 1.1
Male	0.87	0.8	1.08	.28	2.39	0.49 - 11.55
Smoker	1.34	0.79	1.7	.089	3.81	0.82 - 17.76
Constant	-2.73	1.26	2.16	.03		

The following **null hypothesis** is tested:

The coefficient is zero in the population.

So, if the value is smaller than **0.05**, the respective **coefficient** has a **significant influence**.

In our example, we see that **none** of the **coefficients** have a **significant** impact, as all **p-values are greater than 0.05.**

Odds ratio

In this column we can then read the **odds ratio**.

Model

Copy Word Copy Excel

	Coefficient B	Standard error	z	p	Odds Ratio	95% conf. interval
Age	0.04	0.03	1.68	.092	1.04	0.99 - 1.1
Male	0.87	0.8	1.08	.28	2.39	0.49 - 11.55
Smoker	1.34	0.79	1.7	.089	3.81	0.82 - 17.76
Constant	-2.73	1.26	2.16	.03		

For example, the **odds ratio** of 1.04 means that a one **unit increase** in the variable age **increases the probability** that a person is sick by **1.04 times**.

If you liked this Playbook

feel free to **share it**!

Of course we are also happy if you visit us on datatab.net.