HEADS
PhD PROGRAMME IN HEALTH DATA SCIENCE

Logistic regression

STATS – Modelação Estatística

PhD Programme in Health Data Science

Cristina Costa Santos & Andreia Teixeira

Logistic Regression

Aim:

To find the best and simplest model that describes the relationship between an outcome (*dependent variable*) and a set of covariates (*independent variables*).

The main difference between the logistic model and the linear model is that the outcome is a binary (dichotomous) variable.

Logistic Regression

Aim:

To find the best and simplest model that describes the relationship between an outcome (*dependent variable*) and a set of covariates (*independent variables*).

The main difference between the logistic model and the linear model is that the outcome is a binary (dichotomous) variable.

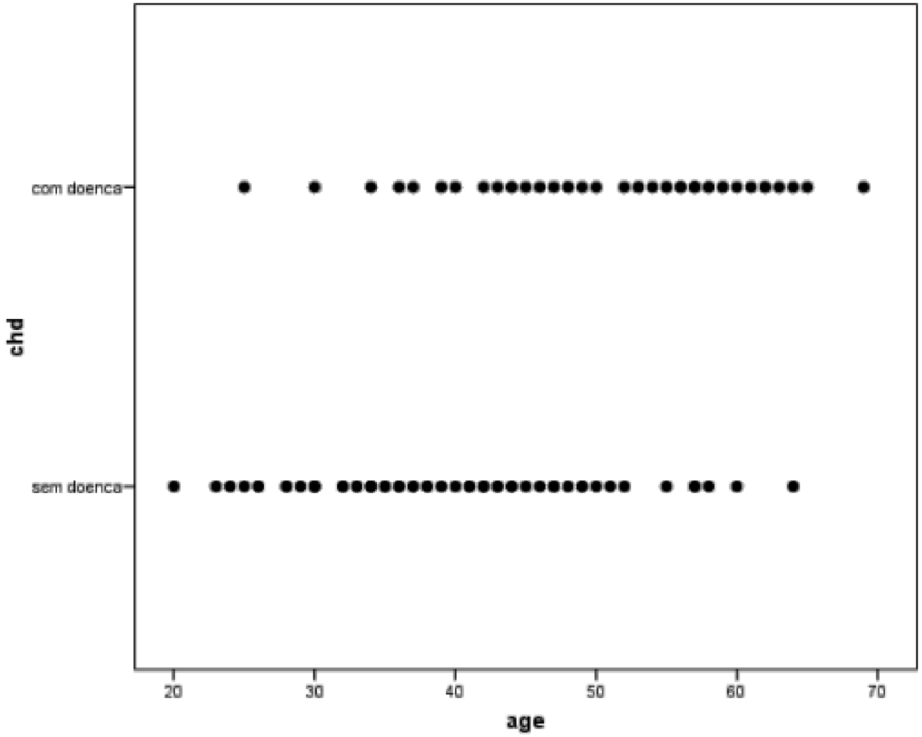
Logistic regression allows estimating the probability associated with the occurrence of a given event given a set of covariates.

Coronary Heart Disease

- Age
- Coronary heart disease (CHD): 0=No | 1=Yes

ID	Age	CHD
1	20	0
2	23	0
3	24	0
4	25	0
5	25	1
6	26	0
…	…	…

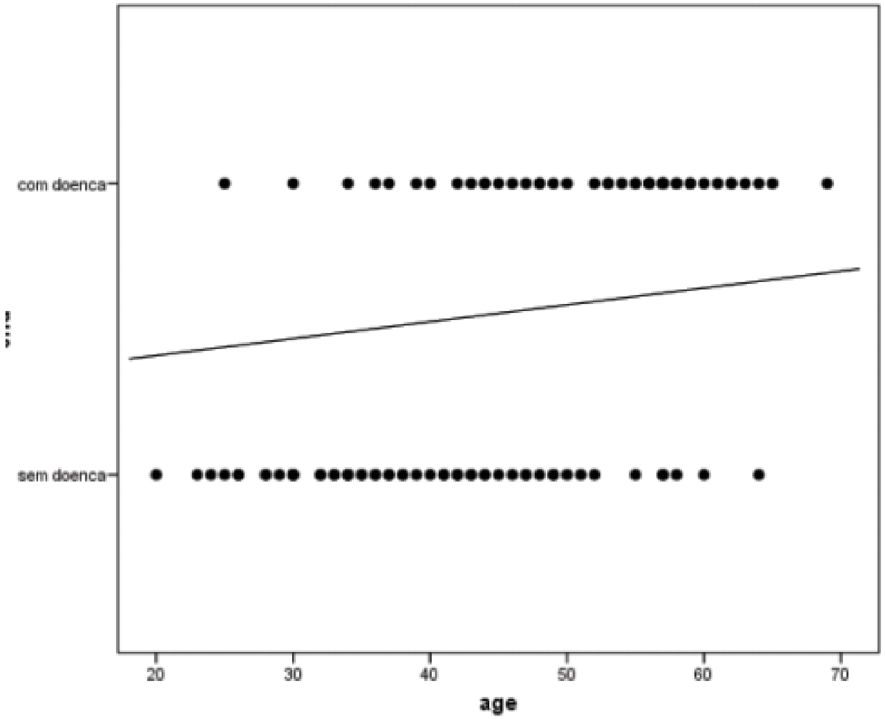
Coronary Heart Disease



we can see that people with coronary disease are older

Coronary Heart Disease

Can we use the linear model?



Coronary Heart Disease

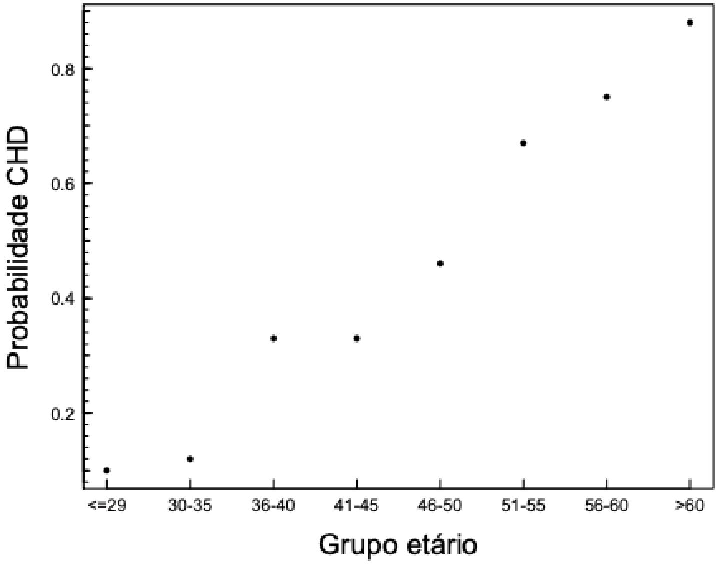
To better understand the relationship between CHD and Age, we can consider age groups.

Age group	n	No CHD	CHD	% CHD
<=29	10	9	1	10
30-35	17	15	2	12
36-40	12	8	4	33
41-45	15	10	5	33
46-50	13	7	6	46
51-55	9	3	6	67
56-60	16	4	12	75
>60	8	1	7	88
	100	**57**	**43**	

We calculated the probability of CHD for each age group.

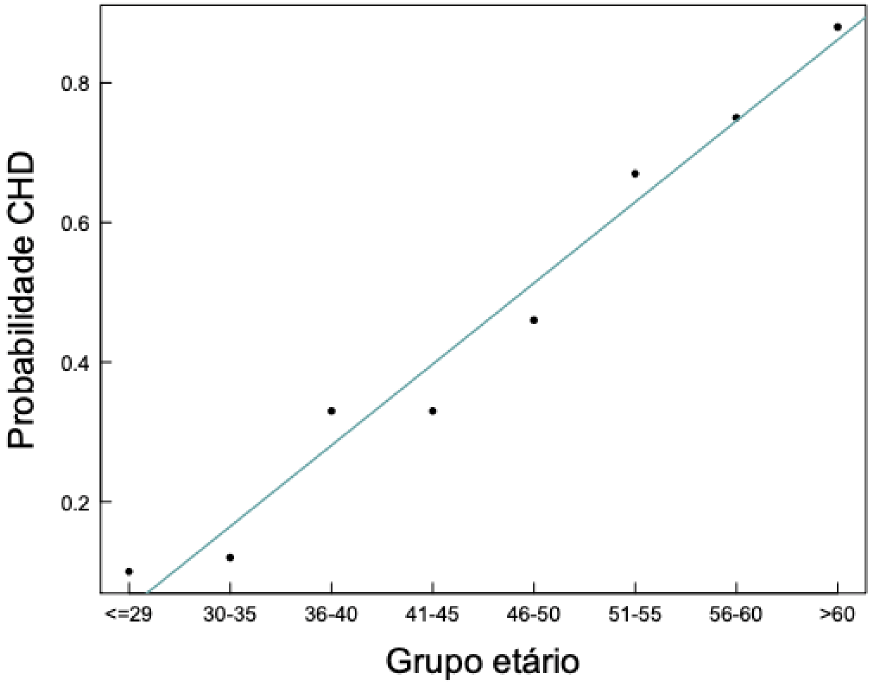
Coronary Heart Disease

Age group	n	No CHD	CHD	% CHD
<=29	10	9	1	10
30-35	17	15	2	12
36-40	12	8	4	33
41-45	15	10	5	33
46-50	13	7	6	46
51-55	9	3	6	67
56-60	16	4	12	75
>60	8	1	7	88
	100	**57**	**43**	



Coronary Heart Disease

Can we use a linear model for Prob(CHD|age)?



Can we use a linear model for Prob(CHD|age)?

If we use a **LINEAR model**:

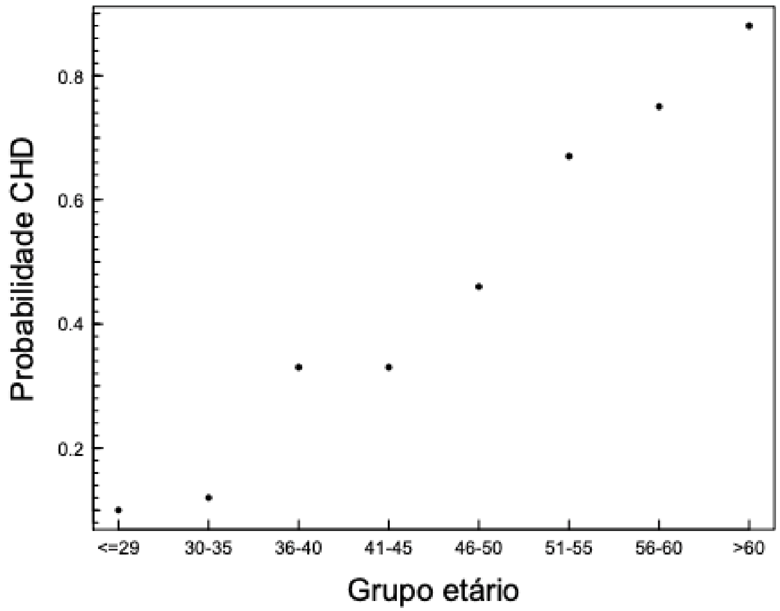
Prob(CHD|age) = - 0.538 + 0.022 x age

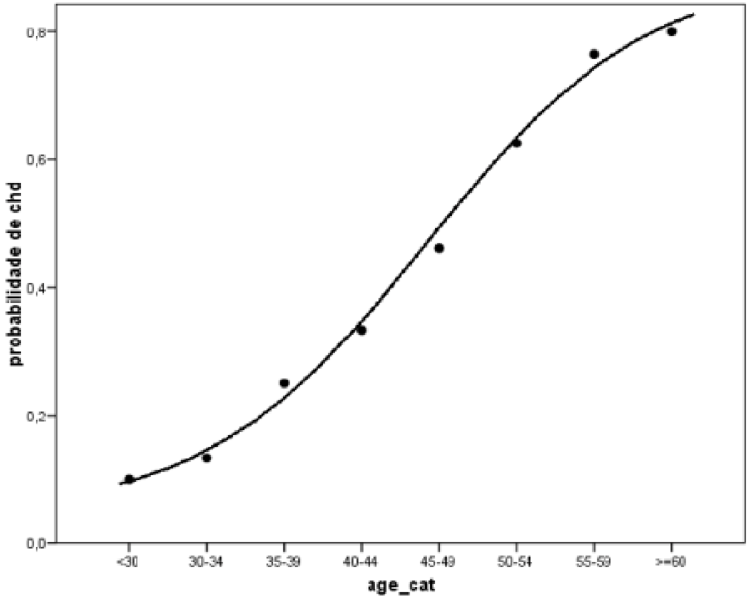
So, for a 75-year-old individual, the estimated probability of coronary heart disease is:

Prob(CHD|age) = - 0.538 + 0.022 x 75 = **1.112 ?!?!?!**

Coronary Heart Disease

Is there anything better than a straight line?





logit

There are several functions with the "S form" and limited between 0 and 1. The most used to model dichotomous variables are:

Logit: $P(|y_i = 1|x_i) = \frac{\exp(\beta_0+\beta_1 x_i)}{1+\exp(\beta_0+\beta_1 x_i)}$

Probit: $P(y_i = 1|x_i) = \Phi^{-1}(\beta_0 + \beta_1 x_i)$, onde

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} \partial t$$

logit

$$P(y_i = 1|x_i) = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}$$

$$\boxed{\log\left(\frac{P(y_i = 1|x_i)}{1 - P(y_i = 1|x_i)}\right)} = \beta_0 + \beta_1 x_i$$

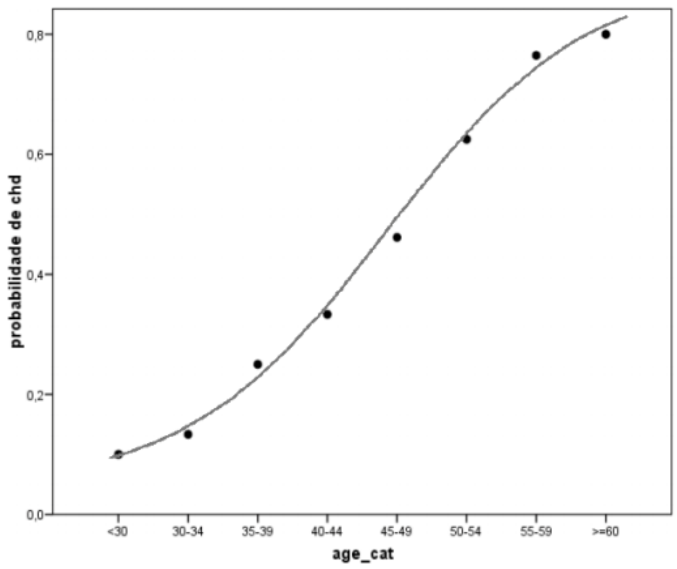
$$\Downarrow$$

$$\boldsymbol{logit(P(y_i = 1|x_i))}$$

The logit function ‘stretches’ the “S-shape” into a straight line.

logit

Age group	n	No CHD	CHD	% CHD	Logit(p)
<=29	10	9	1	10	-2.20
30-35	17	15	2	12	-1.99
36-40	12	8	4	33	-0.71
41-45	15	10	5	33	-0.71
46-50	13	7	6	46	-0.16
51-55	9	3	6	67	0.71
56-60	16	4	12	75	1.10
>60	8	1	7	88	1.99
	100	**57**	**43**		↑

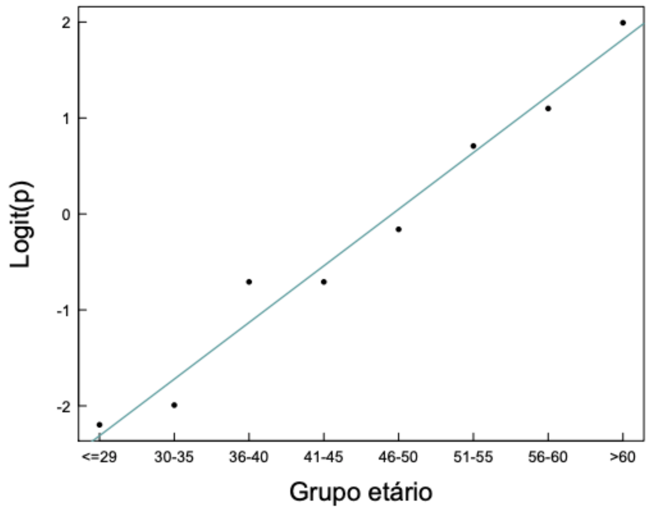


$$P(CHD = 1|idade) = \frac{\exp(\beta_0 + \beta_1 idade)}{1 + \exp(\beta_0 + \beta_1 idade)}$$

$$\log\left(\frac{P(CHD = 1|idade)}{1 - P(CHD = 1|idade)}\right)$$

logit

Age group	n	No CHD	CHD	% CHD	Logit(p)
<=29	10	9	1	10	-2.20
30-35	17	15	2	12	-1.99
36-40	12	8	4	33	-0.71
41-45	15	10	5	33	-0.71
46-50	13	7	6	46	-0.16
51-55	9	3	6	67	0.71
56-60	16	4	12	75	1.10
>60	8	1	7	88	1.99
	100	**57**	**43**		↑



$$\boldsymbol{logit\big(P(CHD = 1|idade)\big) = \beta_0 + \beta_1 idade}$$

Logistic Regression

In the example of the association of age with CHD, the result of the logistic regression is:

$$\mathrm{P}(\boldsymbol{CHD}|\boldsymbol{idade}) = \frac{\boldsymbol{e}^{-5.309+0.111*\boldsymbol{idade}}}{\mathbf{1} + \boldsymbol{e}^{-5.309+0.111*\boldsymbol{idade}}}$$

Logistic Regression

In the example of the association of age with CHD, the result of the logistic regression is:

$$\mathrm{P}(\boldsymbol{CHD}|\boldsymbol{idade}) = \frac{\boldsymbol{e^{-5.309+0.111*idade}}}{\mathbf{1}+\boldsymbol{e^{-5.309+0.111*idade}}}$$

We can calculate the probability of coronary disease estimated by the model, for a given age. For example, for a 75-year-old subjet, the estimated probability is:

$$\boxed{\mathrm{P}(\boldsymbol{CHD}|\boldsymbol{idade}) = \frac{\boldsymbol{e^{-5.309+0.111*75}}}{\mathbf{1}+\boldsymbol{e^{-5.309+0.111*75}}} = \mathbf{0.95}}$$

Coefficients

Let's consider the case that X is a dichotomous variable (takes values 0 and 1)

Assuming the model: $P(y_i = 1|x_i) = \frac{\exp(\beta_0+\beta_1 x_i)}{1+\exp(\beta_0+\beta_1 x_i)}$

We designate **odds** as the quotient between the probability of the outcome occurring and the probability of its non-occurrence.

Therefore, the *logit transform is nothing more than the logarithm of an odds*.

Coefficients

The odds of y=1 for x=1 is:

$$Odds(x=1) = \frac{P(y_i=1|x=1)}{P(y_i=0|x=1)} = \frac{P(y_i=1|x=1)}{1-P(y_i=1|x=1)} = \frac{\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}}{\frac{1}{1+e^{\beta_0+\beta_1}}} = e^{\beta_0+\beta_1}$$

The odds of y=1 for x=0 is:

$$Odds(x=0) = \frac{P(y_i=1|x=0)}{P(y_i=0|x=0)} = \frac{P(y_i=1|x=0)}{1-P(y_i=1|x=0)} = \frac{\frac{e^{\beta_0}}{1+e^{\beta_0}}}{\frac{1}{1+e^{\beta_0}}} = e^{\beta_0}$$

So, the odds ratio is:

$$\boldsymbol{OR = \frac{Odds(x=1)}{Odds(x=0)} = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{\beta_0+\beta_1-\beta_0} = e^{\beta_1}}$$

Coefficients

The logistic regression coefficients **(β)** can then be interpreted as (log) odds ratios.

That is, the odds of y=1 increase exp(β) times, when the variable x increases by one unit (this is also true for continuous x).

Coefficients

The logistic regression coefficients **(β)** can then be interpreted as (log) odds ratios.

That is, the odds of y=1 increase exp(β) times, when the variable x increases by one unit (this is also true for continuous x).

In our example, a one-year increase corresponds to an increase of $e^{0.111} = 1.117$ times the odds of coronary disease.

Attention: the increase of 10 years corresponds to an increase of $e^{0.111*10} = e^{1.11} = 3.03$ (not 1.117 x 10 = 11.17) in the odds of coronary disease.

Inference

Confidence intervals (CI) for the regression coefficients can be obtained using the asymptotic distribution of maximum likelihood estimators.

$$\boxed{IC_{95\%}(\beta_1) = [\widehat{\beta_1} - 1.96 * SE; \widehat{\beta_1} + 1.96 * SE]}$$

We can also obtain the CI for $\exp(\beta)$, that is, the CI for the odds ratio.

Inference

There are several hypothesis tests to test whether the regression coefficients are different from 0:

$$H_0: \beta_1 = 0$$

One such test is called the **Wald test** and is based on the asymptotic distribution of the estimator.

Inference

There are several hypothesis tests to test whether the regression coefficients are different from 0:

$$H_0: \beta_1 = 0$$

One such test is called the **Wald test** and is based on the asymptotic distribution of the estimator.

$$W^2 = (\frac{0.111}{0.024})^2 = 21.254$$

$$P(X_1^2 > 21.254) < 0.001$$

Inference

There are several hypothesis tests to test whether the regression coefficients are different from 0:

$$H_0: \beta_1 = 0$$

One such test is called the **Wald test** and is based on the asymptotic distribution of the estimator.

$$W^2 = \left(\frac{0.111}{0.024}\right)^2 = 21.254$$

$$P(\mathrm{X}_1^2 > 21.254) < 0.001$$

In this case, we reject the null hypothesis ($H_0: \beta_1 = 0$).
That is, **the effect of age on the likelihood of coronary disease is significant.**

Inference

Another way to test the contribution of a variable to the model is to use the likelihood ratio test.

The idea is to compare the likelihood of a model with the covariate (model A) and without the covariate (model B).

$$G = -2\,ln\left(\frac{verosimilhança\ modelo\ B}{verosimilhança\ modelo\ A}\right) \sim \chi^2_{dfA-dfB}$$

Variation

"Pseudo" r^2 – measure the amount of variation explained by the model.

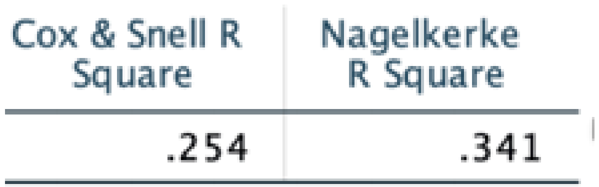
"Pseudo" r^2 Cox & Snell e Nagelkerke are based on the log-likelihood of the model.

r^2 Cox & Snell has a maximum theoretical value (for the "perfect" model) less than 1.

r^2 Nagelkerke is based on the previous one but adjusted so that the maximum theoretical value is 1.

Variation

In our exemple:



Cox & Snell R Square	Nagelkerke R Square
.254	.341

The model (age) explains 34% of the CHD variability.

Multiple logistic regression

The extension of the logistic regression model for multiple covariates is immediate:

$$\boxed{P\left(y_i = 1 \middle| x_{1i}, x_{2i}, \dots, x_{pi}\right) = \frac{\exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + +\beta_p x_{pi})}{1 + \exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + +\beta_p x_{pi})}}$$

Coefficients are interpreted as odds ratios adjusted for other covariates.

Multiple logistic regression

Often the objective of the study is to select the covariates that constitute the "best" model within the scientific context of the problem under analysis.

There are two key points in this process:

- The choice of variables to integrate in the model
- The assessment of the adequacy of the model in general and in terms of individual covariates