HEADS

PhD PROGRAMME IN HEALTH DATA SCIENCE

Logistic Regression

Basics of Health Intelligent Data Analysis

PhD Programme in Health Data Science

Cláudia Camila Dias

Pedro Pereira Rodrigues

U.PORTO
FMUP FACULDADE DE MEDICINA UNIVERSIDADE DO PORTO

MEDCIDS
DEPARTAMENTO DE MEDICINA DA COMUNIDADE, INFORMAÇÃO E DECISÃO EM SAÚDE

CINTESIS
Health. Research.

Linear regression

- When we want to study the relationship of one continuous variable with another, it is common to use the linear regression model.
- This model assumes that the mean of the variable to model depends on a linear combination of the other variables.

- In the simplest case, involving only one independent variable (covariate), the model graphically corresponds to approximate the relationship of the two variables by a straight line.

Linear regression

- For example, we can approximate the relationship between mean systolic pressure and body mass index.

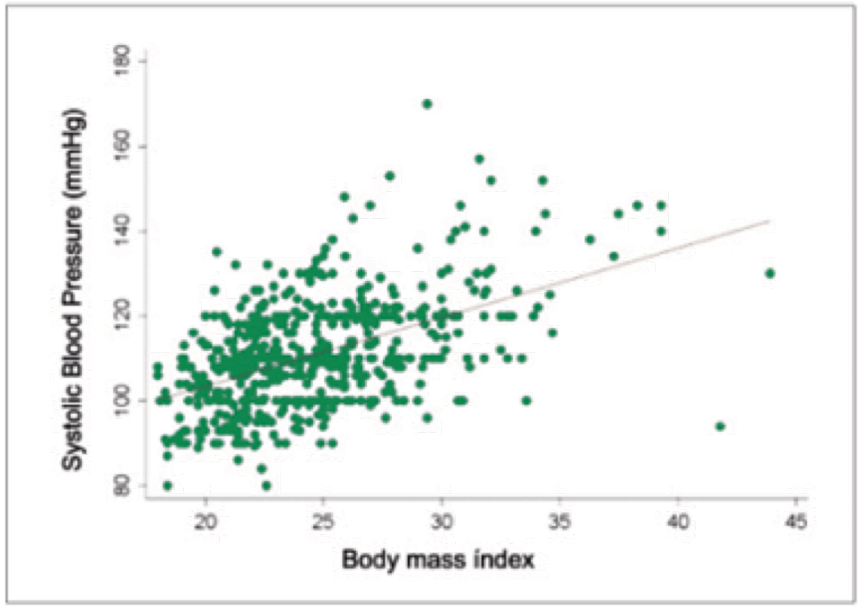


Figure. 1 - Simple linear regression of systolic blood pressure (SBP) in relation to body mass index (BMI) in a sample of 536 adolescents (r = 0.436; p=0.000; Beta coef. = 1.198).

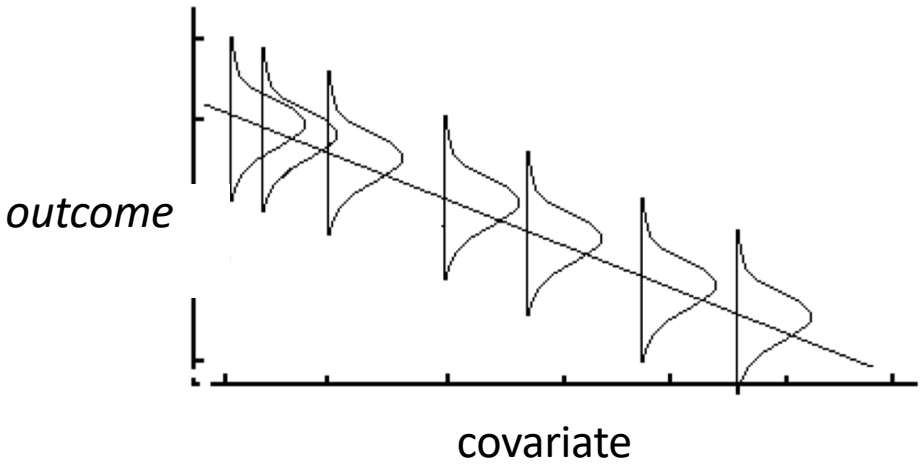
Simple linear regression

$$\mu_{y|x} = \alpha + \beta x$$

Linear regression

Linear model assumptions:

- The association of covariates and the outcome is linear
- Instances are independent
- The outcome (conditioned on the covariates) follows a normal distribution (i.e. the error follows a normal distribution)
- Homochedasticity – variance is homogeneous



Motivation for the logistic model

- **Objective**: 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.

Motivation for the logistic model

Example:

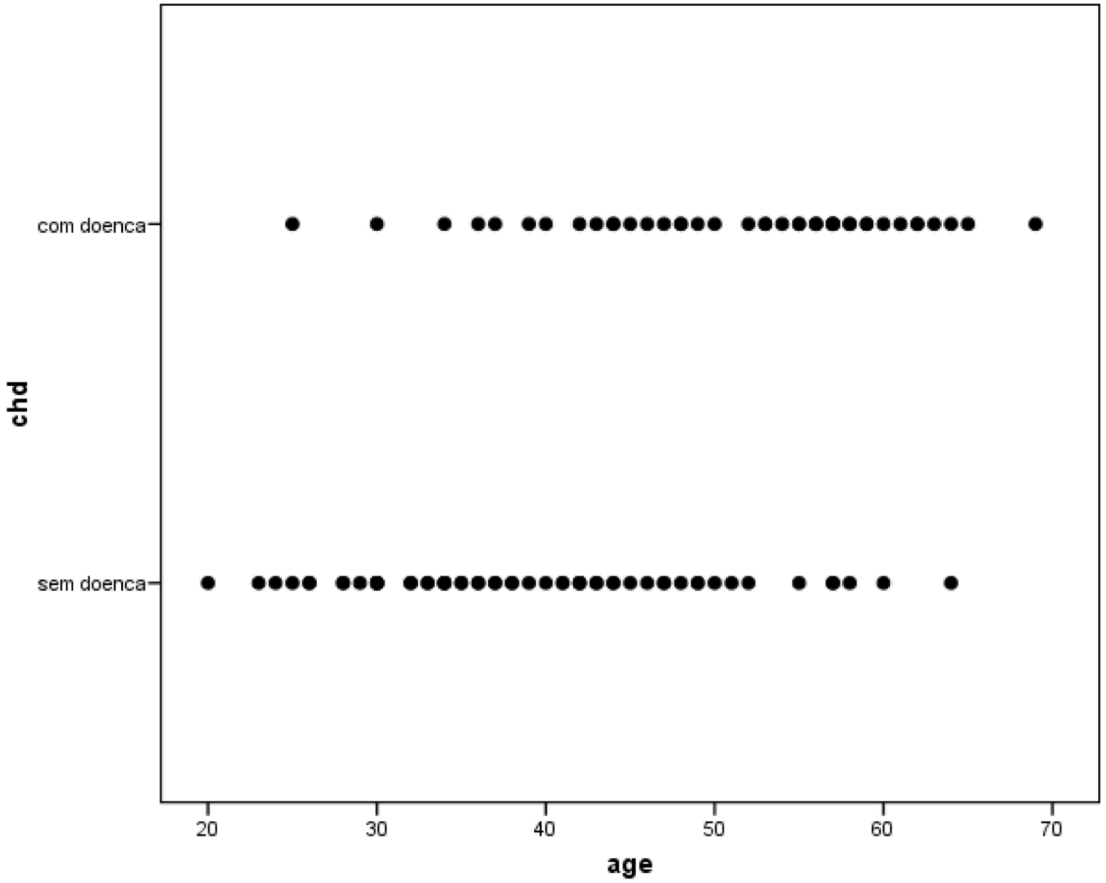
- Age and coronary heart disease (chd)
- 100 individuals

CHD = { 0 : without chd
1 : with chd

ID	AGE	CHD
1	20	0
2	23	0
3	24	0
4	25	0
5	25	1
6	26	0
…	…	…
98	64	1
99	65	1
100	69	1

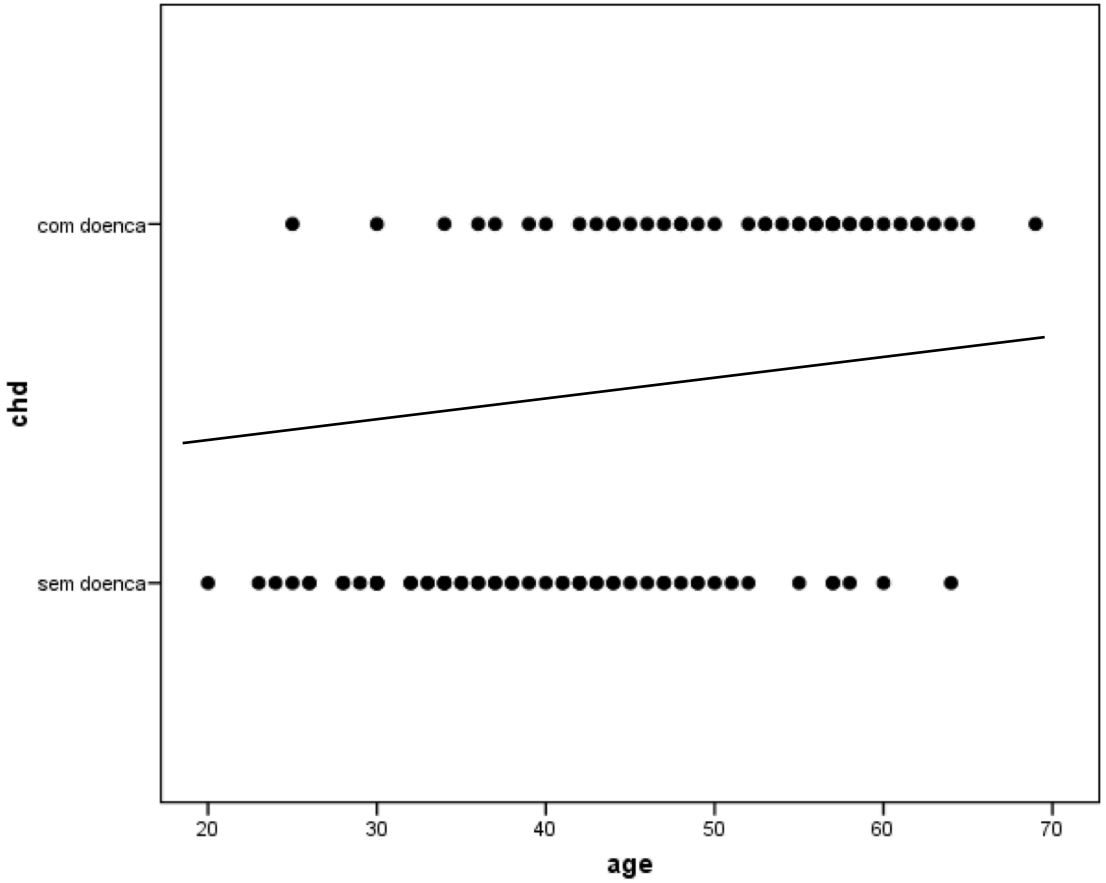
Motivation for the logistic model

- **Relation between CHD and AGE**



Motivation for the logistic model

Using a linear model…?



Age and coronary heart disease

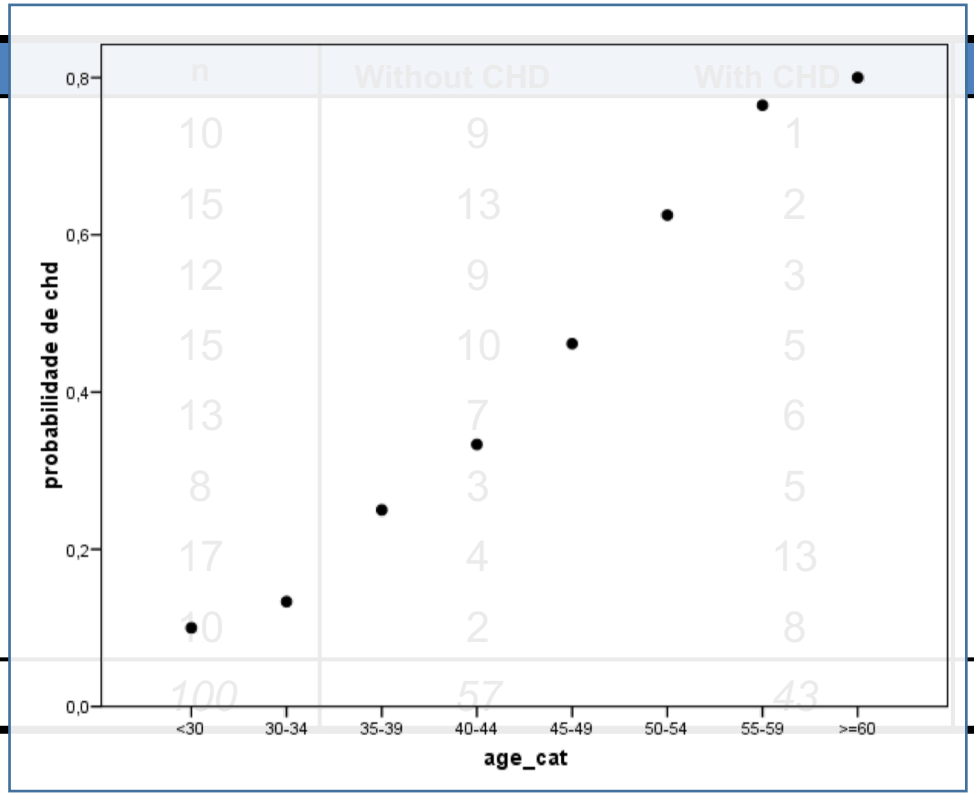
To better understand the relation between CHD and age we should consider age groups

Age Group	n	Without CHD	With CHD	% outcome
20-29	10	9	1	0.10
30-34	15	13	2	0.13
35-39	12	9	3	0.25
40-44	15	10	5	0.33
45-49	13	7	6	0.46
50-54	8	3	5	0.63
55-59	17	4	13	0.76
60-69	10	2	8	0.80
	100	*57*	*43*	

Age and coronary heart disease

To better understand the relation between CHD and age we should consider age groups

Age Group	n	Without CHD	With CHD	% outcome
20-29	10	9	1	0.10
30-34	15	13	2	0.13
35-39	12	9	3	0.25
40-44	15	10	5	0.33
45-49	13	7	6	0.46
50-54	8	3	5	0.63
55-59	17	4	13	0.76
60-69	10	2	8	0.80
	100	57	43	



Age and coronary heart disease

In case of a binary outcome $\mathbf{y_i}$ we'll model

$$\mathbf{P(y_i \mid x_i)}$$

In our example:

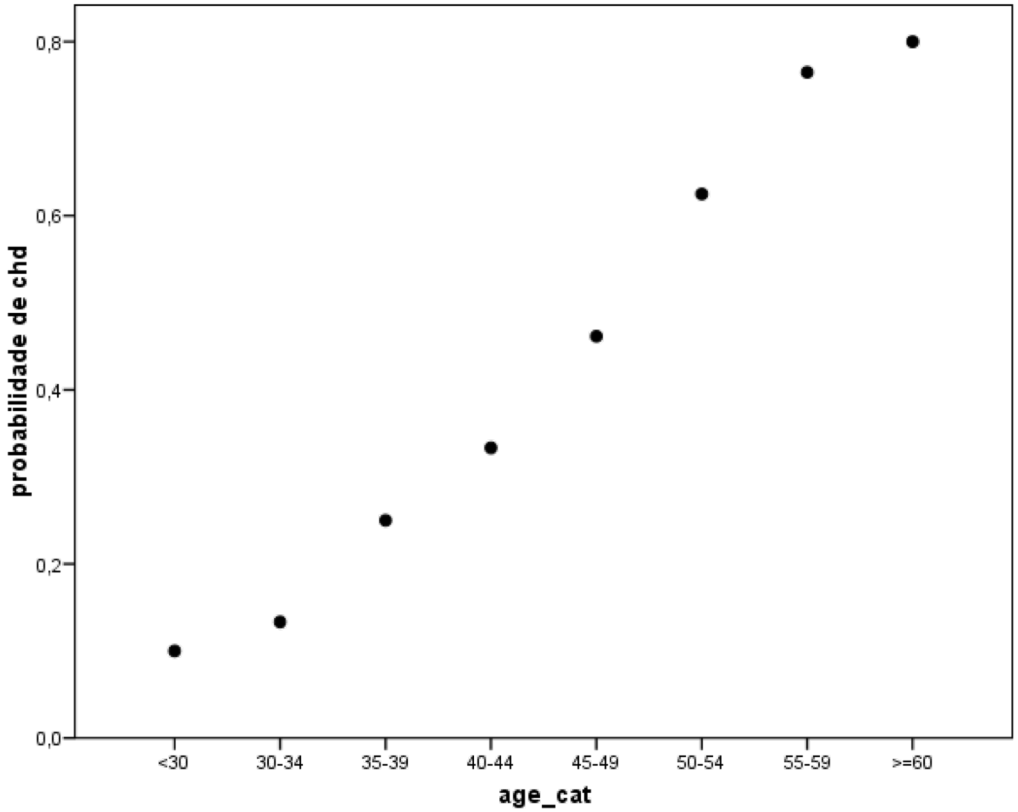
Prob (coronary heart disease | age)

It is important to mention that:

- If y is coded as 0 and 1, $P(y_i|x_i) = \mu_{y|x}$
- $0 <= P(y_i|x_i) <= 1$

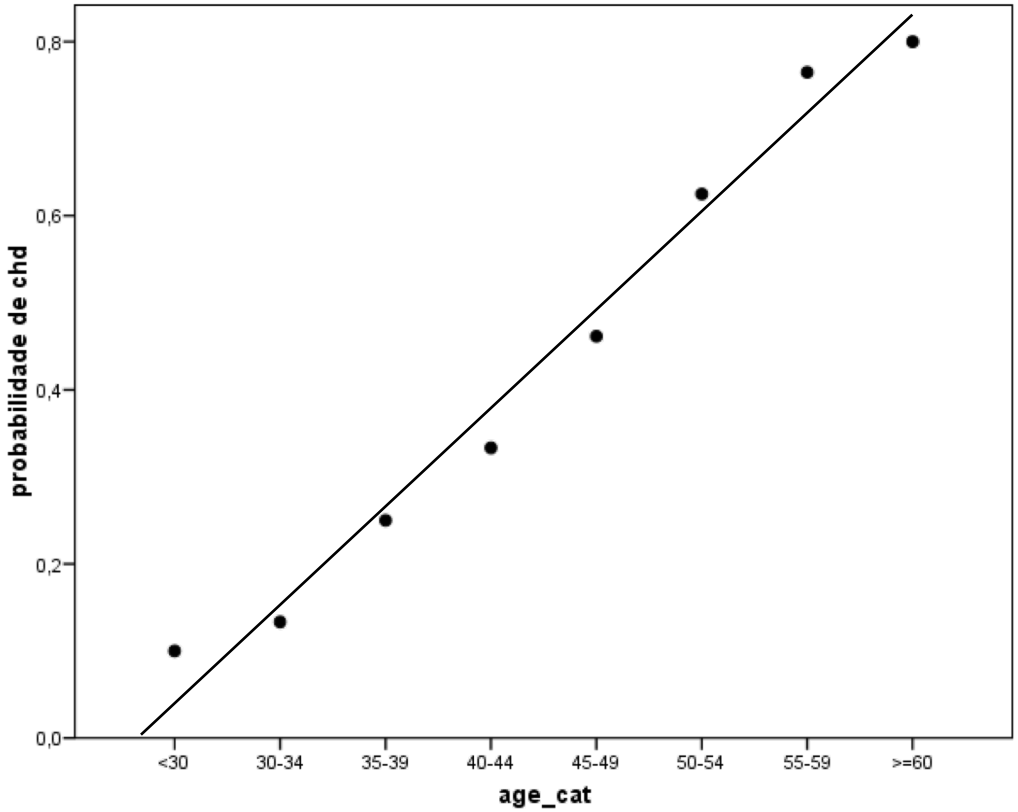
Age and coronary heart disease

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



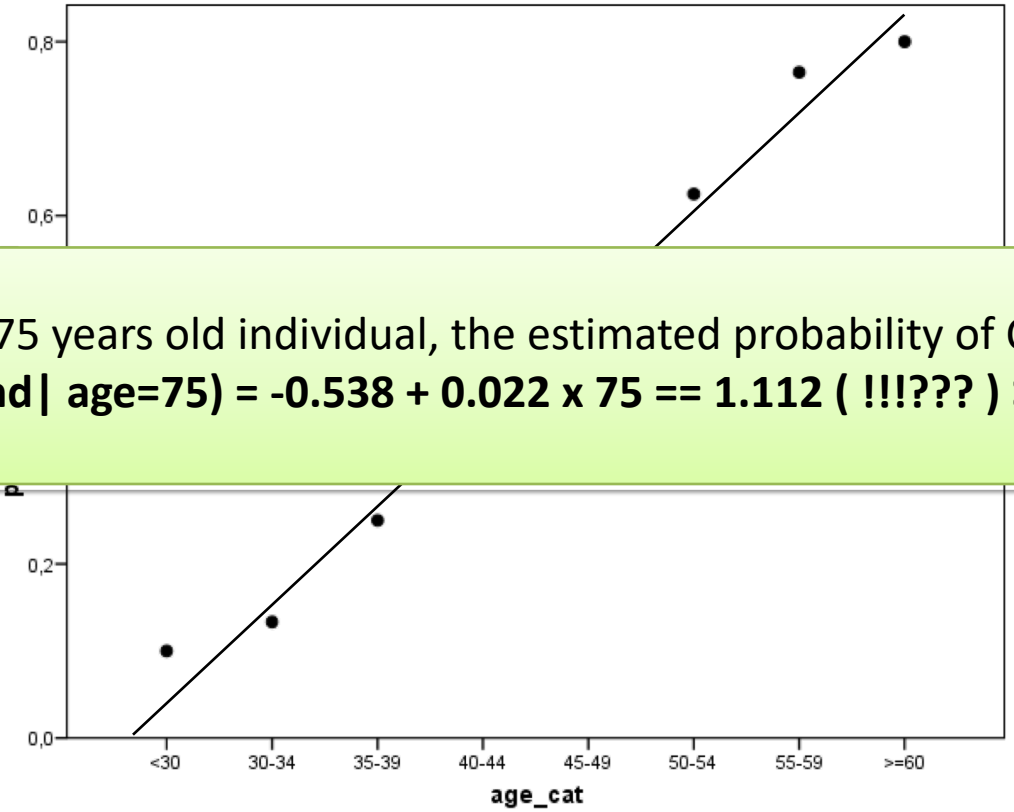
Age and coronary heart disease

Using linear regression ***P (chd | age) = -0.538 + 0.022 x age***



Age and coronary heart disease

Using linear regression ***P (chd | age) = -0.538 + 0.022 x age***

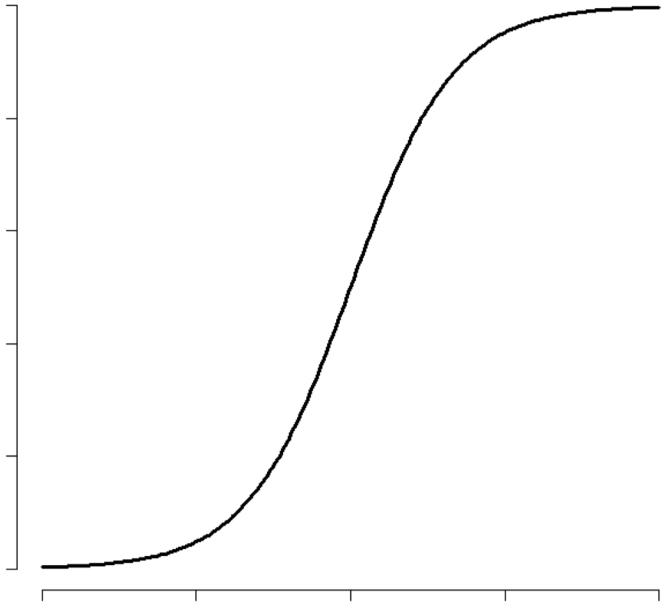


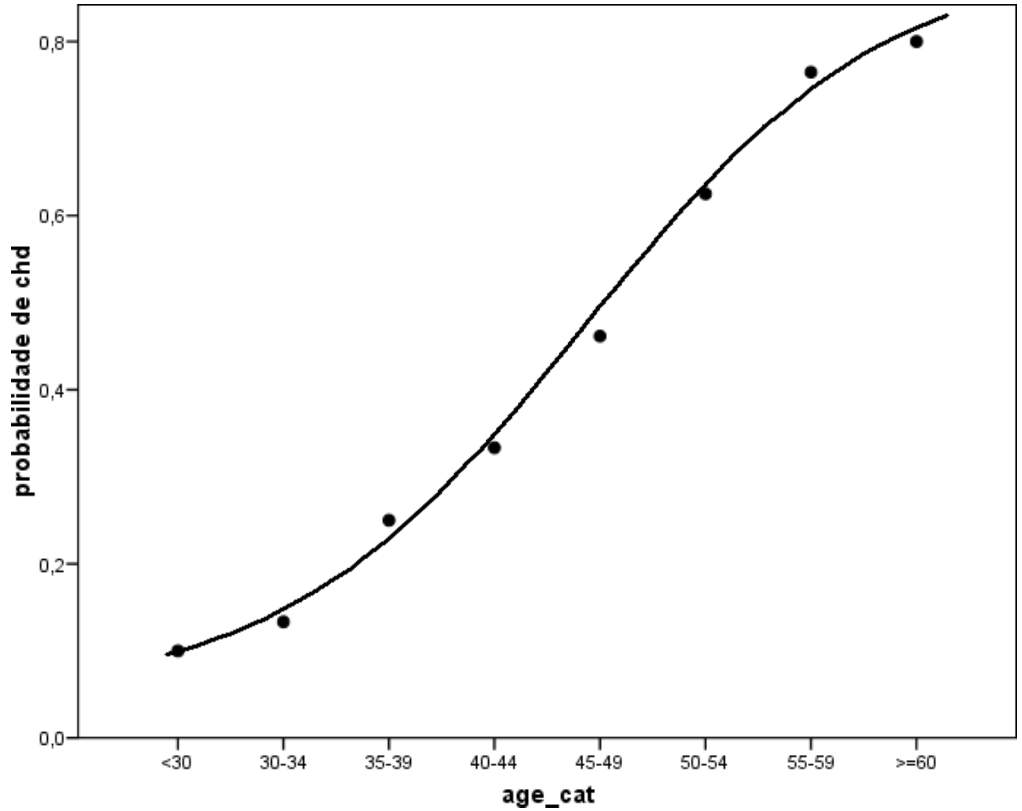
For a 75 years old individual, the estimated probability of CHD is:

P (chd| age=75) = -0.538 + 0.022 x 75 == 1.112 (!!!???) > 1!!!

The logit function

Can we find a better model than the linear?





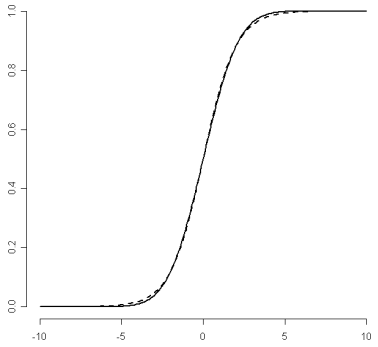
The logit function

There are several “s-shaped” functions which are also limited between 0 and 1

The most common is:

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

The logit function comes after that to help get us back linear



logit (P(e)) = log (odds(e))

The logit function

The logit function stretches the “s-shape” into a straight line.

That is, if $P(y_i|x_i)$ is “s-shaped”,

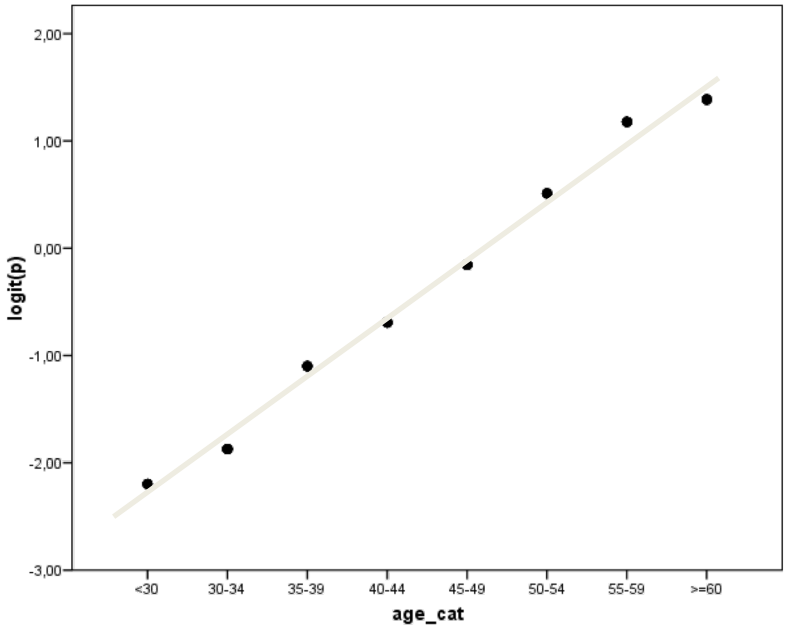
$logit\ (P(y_i|x_i))$ is approximately a straight line.

So, we can use the linear form to model $logit\ (P(y_i|x_i))$

Logistic regression model

Age group	n	With	Without	P(o)=p	Logit (p)
20-29	10	9	1	0.10	**-2.20**
30-34	15	13	2	0.13	**-1.87**
35-39	12	9	3	0.25	**-1.10**
40-44	15	10	5	0.33	**-0.69**
45-49	13	7	6	0.46	**-0.15**
50-54	8	3	5	0.63	**0.51**
55-59	17	4	13	0.76	**1.18**
60-69	10	2	8	0.80	**1.39**
	100	*57*	*43*		

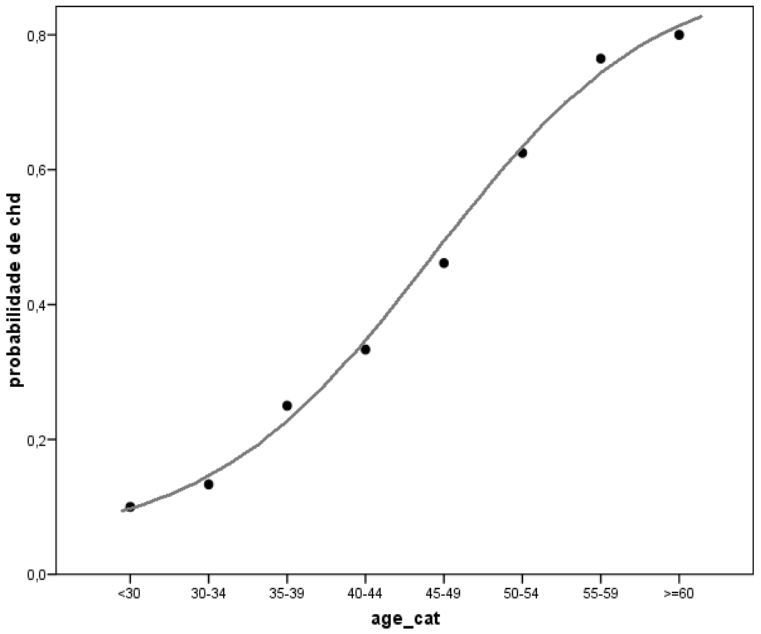
$$Logit\ [\ P\ (chd|age)\ ] = \beta_0 + \beta_1\ age$$



Logistic regression model

Age group	n	With	Without	P(o)=p	Logit (p)
20-29	10	9	1	0.10	**-2.20**
30-34	15	13	2	0.13	**-1.87**
35-39	12	9	3	0.25	**-1.10**
40-44	15	10	5	0.33	**-0.69**
45-49	13	7	6	0.46	**-0.15**
50-54	8	3	5	0.63	**0.51**
55-59	17	4	13	0.76	**1.18**
60-69	10	2	8	0.80	**1.39**
	100	*57*	*43*		

$$P\,(chd|age) = \frac{\exp(\beta_0 + \beta_1 \text{age})}{1 + \exp(\beta_0 + \beta_1 \text{age})}$$



Estimation

In the example, the logistic regression gives

Variables in the Equation

	B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a] age	,111	,024	21,254	1	,000	1,117
Constant	-5,309	1,134	21,935	1	,000	,005

a. Variable(s) entered on step 1: age.

$\hat{\beta}_1$

$\hat{\beta}_0$

$$P(chd|age) = \frac{\exp(-5.309 + 0.111 \times age)}{1 - \exp(-5.309 + 0.111 \times age)}$$

Estimation

We can calculate the probability of coronary disease estimated by the model for a given age.

For example, for a 50-year-old, the estimated probability is:

$$P(chd \mid age = 50) = \frac{\exp(-5.309 + 0.111 \times 50)}{1 - \exp(-5.309 + 0.111 \times 50)} = 0.56$$

Interpretation

Consider the case of *X* being a dichotomous variable.

Assuming the model

$$P(y=1|x)=\frac{\exp(\beta_0+\beta_1 x)}{1+\exp(\beta_0+\beta_1 x)}$$

the odds of y=1 given x=1 is:

$$\frac{P(y=1|x=1)}{P(y=0|x=1)}=\frac{P(y=1|x=1)}{1-P(y=1|x=1)}=\frac{\frac{\exp(\beta_0+\beta_1)}{1+\exp(\beta_0+\beta_1)}}{\frac{1}{1+\exp(\beta_0+\beta_1)}}=\exp(\beta_0+\beta_1)$$

Interpretation

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

$$\frac{P(y=1|x=0)}{1-P(y=1|x=0)} = \frac{\frac{\exp(\beta_0)}{1+\exp(\beta_0)}}{\frac{1}{1+\exp(\beta_0)}} = \exp(\beta_0)$$

Thus, the odds ratio is given by:

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

$$OR = e^{\beta_1}$$

Interpretation

The coefficients of logistic regression can thus be interpreted as (log) odds ratio.

That is, the odds of y=1 increases $\exp(\beta_1)$ times when *x* increases one unit (also true for continous *x*).

Interpretation

In the example

Variables in the Equation

	B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a] age	,111	,024	21,254	1	,000	1,117
Constant	-5,309	1,134	21,935	1	,000	,005

For an increase of one year, there is an increase of 1.117 times the odds of coronary heart disease.

Multiple logistic regression

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

$$P\left(y_i = 1 \mid x_{1i}, x_{2i}, \ldots, x_{pi}\right) = \frac{\exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi})}{1 + \exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi})}$$

β_j coefficients are interpreted as odds ratio adjusted for the remaining variables.

Model building and evaluation

We have been studying the estimation, interpretation and inference (CI and hypothesis testing).

Often the object 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 into the model
- The evaluation of the adequacy of the model in general and in terms of individually covariates

Selecting covariates

- Include covariates that are clinically relevant and / or derived from scientific knowledge of the area

- Use statistical methods to decide on the inclusion of covariates

Selecting covariates

I - Univariate Analysis

- Test each of the covariates individually
    - E.g. use contingency tables for categorical variables and t-tests for continuous variables, or use logistic regressions for each covariate
- The advantage of contingency tables is the ease of identifying possible empty categories. This situation may lead to instability in the algorithms used in logistic regression.
- If there are empty categories, or with few individuals, we should consider joining categories

Selecting covariates

II - Select the subset of variables to use in the multivariate model

- A commonly used rule is to select covariates with $p < 0.25$ in the univariate analysis as candidates for the multivariate model.

- Also included in this subgroup are covariates that have a scientific justification or are of direct interest to research (e.g. treatment) even if $p > 0.25$

Selecting covariates

III - Running the multivariate model with the subset of covariates

- Examine the significance (statistical and scientific) of covariates in the model
- Compare the estimation of multivariate model parameters with the estimation of univariate models
- Variables that do not contribute to the model (and there is no other justification for maintaining them in the model) should be removed one by one and the model should be reevaluated at each step.

Selecting covariates

IV. Look more closely

- Having obtained a preliminary version of the multivariate model with the set of covariates that we think are the most important, we must now look more closely at each of the variables.
- Categorical variables must have been analyzed in univariate models.
- Regarding continuous variables we should check the linearity in the logit scale.

Selecting covariates

V – Study interactions

- In the last covariate selection step we studied potential interactions between covariates.
- Due to the number of possible interactions, we usually restrict ourselves to testing interactions that make sense in the context of the problem.

Automatic selection (stepwise)

Stepwise methods are automatic covariate selection procedures for integrating the model.

Some authors criticize this procedure because it is based only on statistical criteria and there is no more active intervention by the researcher.

However these methods may be useful for a first approach to model building or when the goal is to get a model with better predictability

Automatic selection (stepwise)

There are **two** stepwise algorithms:

- **Backward stepwise**
- **Forward stepwise**

Although there are slight variations to the backward and forward stepwise algorithms the fundamental idea is as follows.

Automatic selection (stepwise)

Backward stepwise

- The process starts with all variables in the model and the definition of the output criterion (e.g. max p=0.1).
- Using the Wald test or likelihood ratio test, we removed from the model the variable with the highest p value, provided that p exceeds the defined output criterion.
- We run the model again but now without this variable.
- The process is repeated until there are no more covariates that meet the exit criteria.

Automatic selection (stepwise)

Forward stepwise

- The process begins by making separate logistic regressions for each covariate and defining an input and output criterion (e.g. the input criterion is 0.05 and the output criterion is 0.1).
- The most significant variable (based on the Wald test or likelihood ratio test) that satisfies the input criterion is chosen to enter the model.
- We ran regressions separately for the remaining covariates but now with the covariate selected in the previous step.
- It is possible that by adding one more covariate, the previous covariate will no longer matter. If the p value for the preceding covariate has been changed to a value greater than the output criterion, then the previous covariate is removed.
- The previous steps are repeated until there are no more covariates that meet the input criteria.

Model evaluation

Quality of adjustment (sometimes referred to as calibration)

- Compare the observed outcome with the predicted model

Ability to discriminate

- Given the model covariates what is the ability to classify an individual

Calibration

Hosmer-Lemeshow Test (chi-squared based)

General idea:

- Compare what is predicted by the model and what is observed

Problem:

- Observations are binary but the model predicts probabilities

We calculate for each individual the probability predicted by the model.

Then, we consider the deciles (percentiles 10, 20, ..., 90) of the probabilities predicted by the model.

In each decile we can calculate the number of individuals with event observed and compare it with the expected number.

Discrimination

Sensitivity & Specificity

- To classify an individual based on the probability predicted by the model we need a rule:
  - For example, classify as
    - 0 if $P(y = 1|x) \leq 0.5$
    - 1 if $P(y = 1|x) > 0.5$
- Now we can compare the classification with what is observed

Discrimination

We can consider other “cutoffs” as classification criteria instead of 0.5.

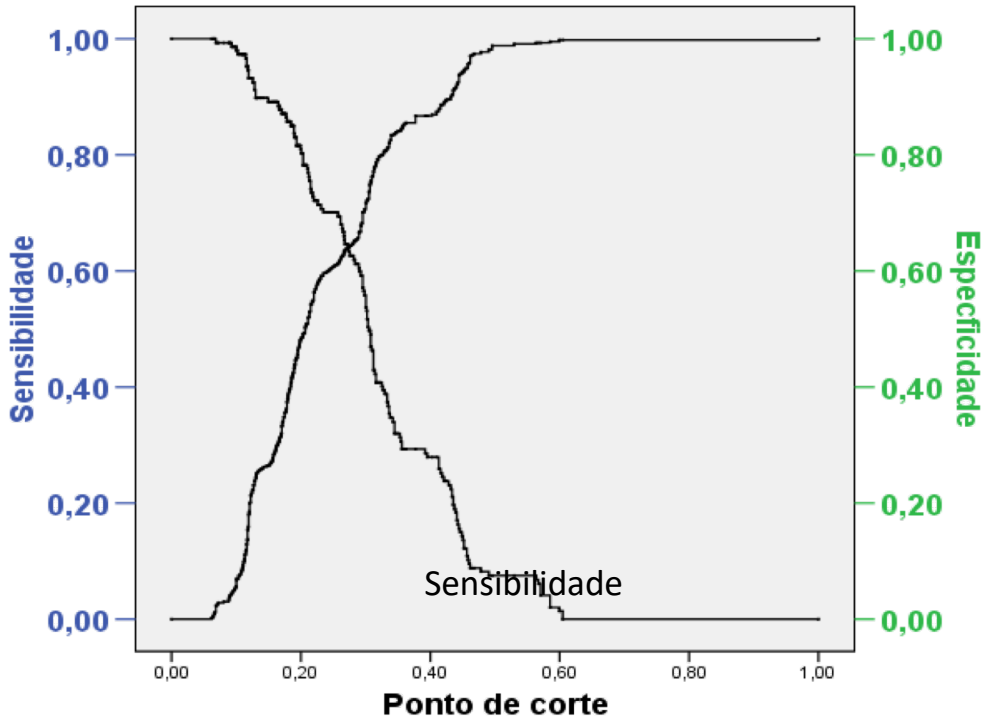
Any criterion used is not sufficient to give a correct idea of the discriminatory capacity of the model.

As an extreme case, imagine a situation where all probabilities predicted by the model are less than 0.45; using the “cutoff” 0.5, no individual would be classified as positive outcome.

However, the model predicts that some individuals will have this outcome.

Discrimination

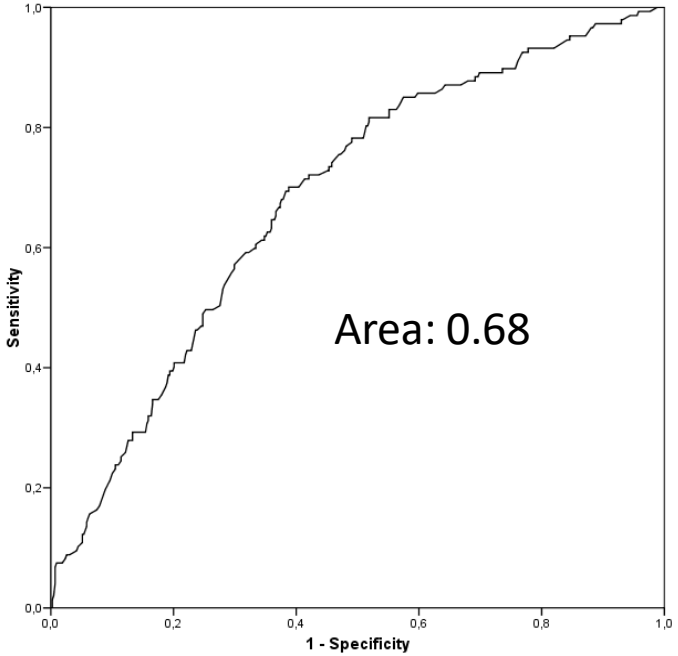
One option for choosing a “cutoff point” is to consider all possible “cutoffs” and calculate their sensitivity and specificity for each situation.



Discrimination – ROC Curve

The graph representing the sensitivity and specificity pairs is called the receiver operating characteristic (ROC) curve. The area under the ROC curve is a measure of the model's ability to discriminate.

= 0.5	indiscriminative
.7 - .8	acceptable
.8 - .9	excellent
> .9	hummmm...



HEADS

PhD PROGRAMME IN HEALTH DATA SCIENCE

Cheers!

Title

Logistic Regression

Acknowledgments

Armando Teixeira Pinto, University of Sydney, Australia

Basics of Health Intelligent Data Analysis

PhD Programme in Health Data Science

Cláudia Camila Dias

Pedro Pereira Rodrigues

U.PORTO FMUP FACULDADE DE MEDICINA UNIVERSIDADE DO PORTO

MEDCIDS DEPARTAMENTO DE MEDICINA DA COMUNIDADE, INFORMAÇÃO E DECISÃO EM SAÚDE

CINTESIS Health. Research.