Linear models for regression and classification

Federica Eduati

Eindhoven University of Technology
Department of Biomedical Engineering

2024

Learning goals

- **Fundamentals:** Understand the principles of linear regression (continuous outcomes) and logistic regression (binary outcomes).
- **Model Interpretation:** Formulate and interpret the equations of linear and logistic regression, focusing on the significance of coefficients.
- **Estimation:** Learn how to estimate parameters using least squares for linear regression and maximum likelihood for logistic regression, including the concept of gradient descent.
- **Model Evaluation:** Identify performance metrics for both models.

Material

- Chapter 3 of "*An introduction to statistical learning with applications in python*, G. James, D. Witten, T. Hastie, R. Tibshirani, J. Taylor"

Overview

Topics covered in this lecture

- Regression vs classification
- Linear regression
  - Simple linear regression
  - Multiple linear regression
- Logistic regression

Join at menti.com | use code 6203 8689

TU/e

Instructions

Go to

www.menti.com

Enter the code

6203 8689

Or use QR code

Introduction to linear models

Linear models: a linear combination of the *inputs* (also known as *predictors*, *features* or *independent variables*) is used to predict one or more outputs (also known as *response* or *dependent variables*).

The prediction task is defined as:

- **Regression**: when we predict *quantitative outputs*.
- **Classification**: when we predict *qualitative outputs* (also referred to as *categorical or discrete variables*).

Which ones are regression problems?

1. Predicting blood pressure based on age and weight.
2. Estimating kidney function as glomerular filtration rate (GFR) based on creatinine levels and age.
3. Predicting if a patient has diabetes based on fasting blood sugar and BMI.
4. Predicting tumour growth rate based on time and initial tumour size.
5. Diagnosing pneumonia from chest X-ray images.
6. Identifying malignant tumours from gene expression profiles.

For which ones can we use a linear model?

1. Predicting blood pressure based on age and weight.
2. Estimating kidney function as glomerular filtration rate (GFR) based on creatinine levels and age.
3. Predicting if a patient has diabetes based on fasting blood sugar and BMI.
4. Predicting tumour growth rate based on time and initial tumour size.
5. Diagnosing pneumonia from chest X-ray images.
6. Identifying malignant tumours from gene expression profiles.

Simple linear regression model

We want to predict a quantitative response Y based on one variable X.

$$Y \approx \beta_0 + \beta_1 X \qquad (\approx \text{ means approximation})$$

The model *coefficients* or *parameters* are unknown constants:

- β_0 is the *intercept*
- β_1 is the *slope*

Simple linear regression model

If we have a training data consisting in n observation pairs

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

we can estimate our model coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$ and use them to make predictions $\hat{y}$ for the Y on the bases of $X = x$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

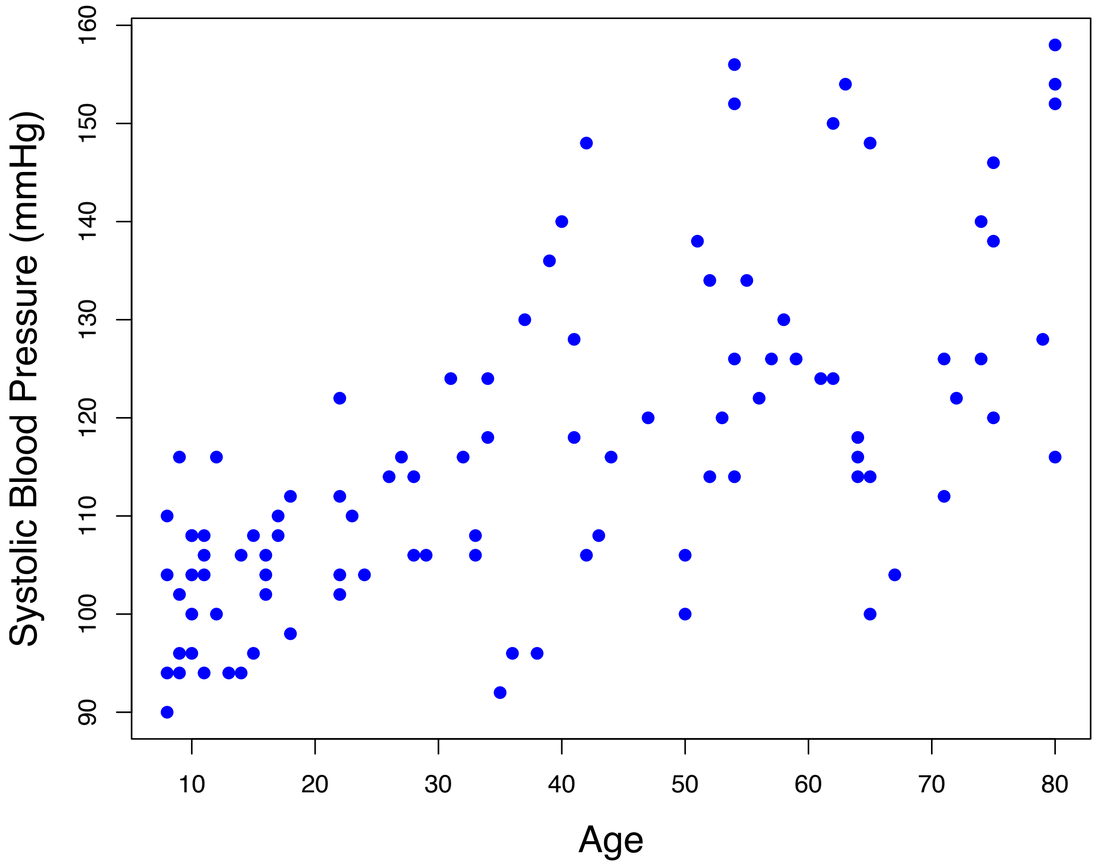
NOTE: ˆ is used to denote predicted value for an unknown parameter or to denote the predicted value of the response

Example: model formulation

We have measurement of Systolic Blood Pressure (SBP) and Age for 100 patients (n=100)

We can regress SBP (our Y) onto Age (our X) fitting the model

$$SBP \approx \beta_0 + \beta_1 \times Age$$



Data from the National Health and Nutrition Examination Survey (NHANES)

Which are the independent and the dependent variables?

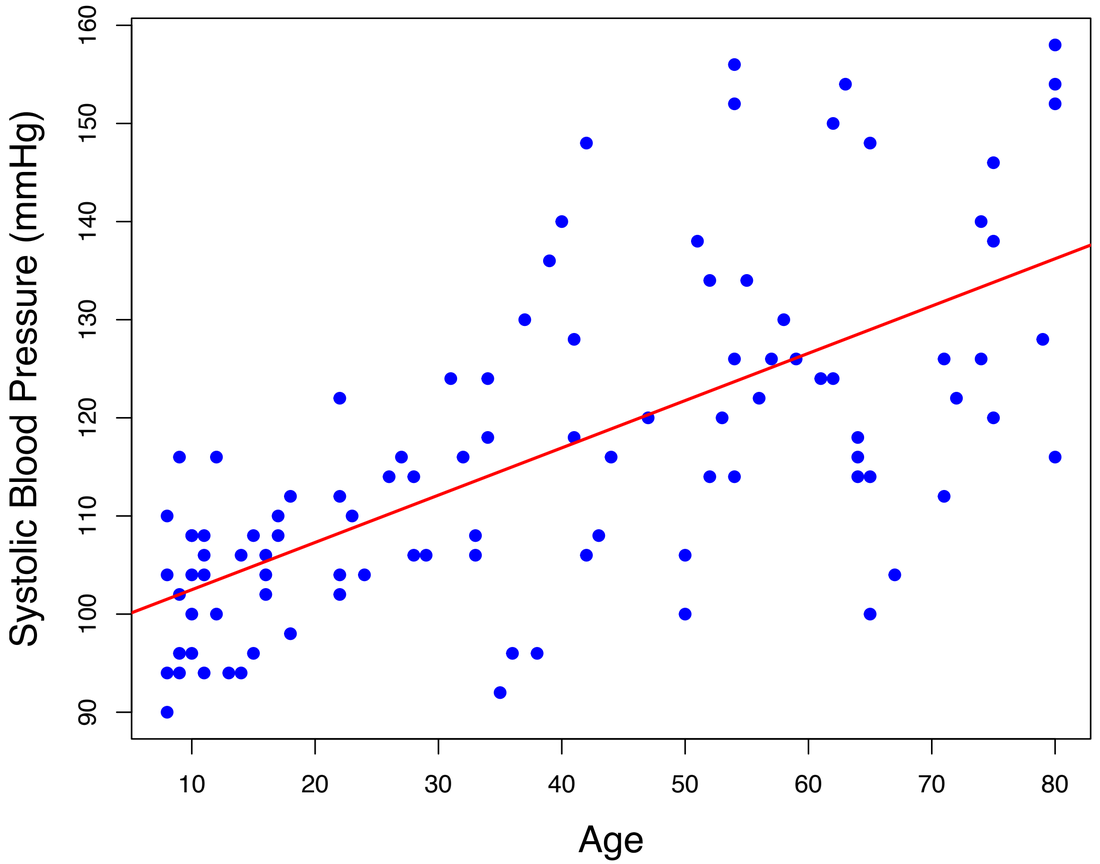
$$SBP \approx \beta_0 + \beta_1 \times Age$$

1. SBP is the dependent variable and Age the independent variable.
2. Age is the dependent variable and SBP the independent variable.
3. The are both dependent variables
4. The are both independent variables

Example: model fitting

We want to obtain the coefficient estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ so that the linear model fits the data well (i.e. model predictions are close to the data points)

$$y_i \approx \beta_0 + \beta_1 x_i \text{ for } i = 1, \ldots, n$$



How do we define 'closeness'?

Residual sum of squares (RSS)

If we define:

$\hat{y}_i = \hat{\beta}_0 - \hat{\beta}_1 x_i$ the prediction for Y based on the ith value of X

$e_i = y_i - \hat{y}_i$ the ith residual (i.e. difference between predicted and observed value)

The residual sum of squares (RSS) is:

$$RSS = e_1^2 + e_2^2 + \ldots + e_n^2 =$$
$$= (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \ldots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2$$

Least square estimates for simple linear regression

We want to choose the $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimise the RSS.

Closed form solution:

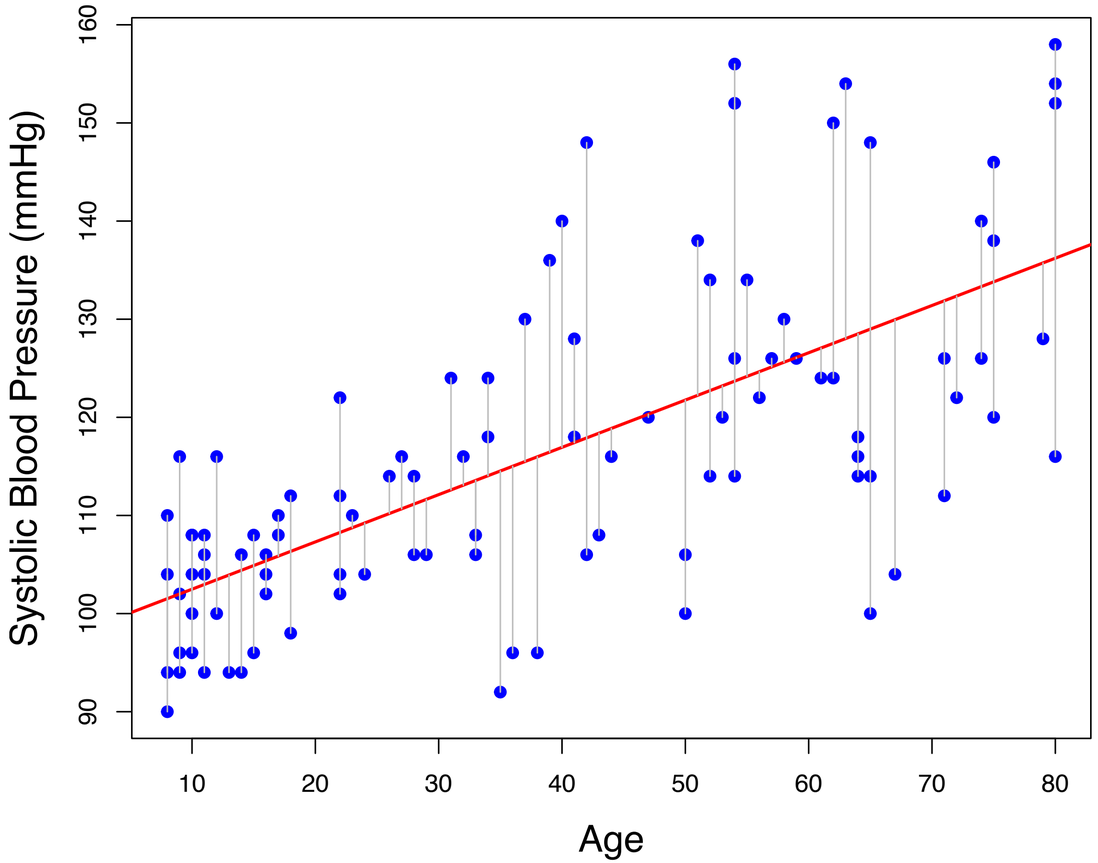
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

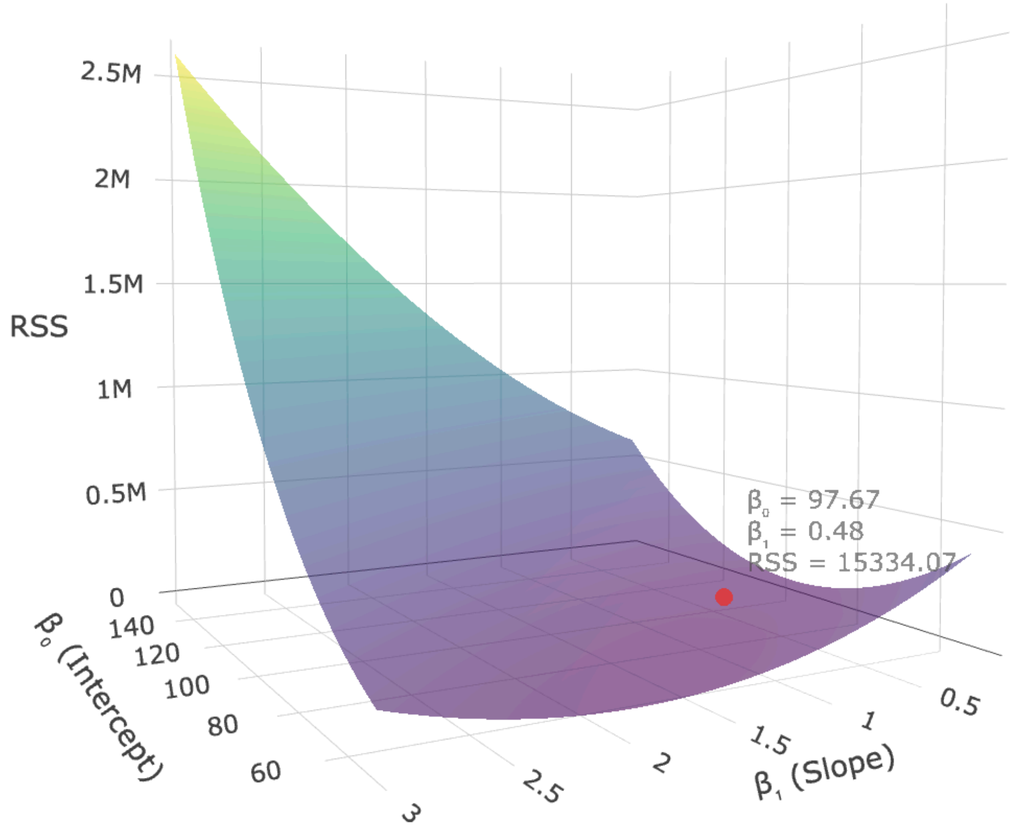
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Where $\bar{y} \equiv \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\bar{x} \equiv \frac{1}{n}\sum_{i=1}^{n} x_i$ are the sample means.

Example: least square estimates

Least squared fit found minimising the residual sum of squares





Gradient descent for parameter estimation

Gradient descent updates the parameters β_0 and β_1 by taking steps proportional to the negative gradient of the RSS with respect to these parameters

$$\beta_0^{(t+1)} = \hat{\beta}_0^{(t)} - \rho \frac{\partial RSS}{\partial \beta_0}$$

$$\beta_1^{(t+1)} = \hat{\beta}_1^{(t)} - \rho \frac{\partial RSS}{\partial \beta_1}$$

Where:

ρ is the learning rate controlling the size of the steps

t is the iteration index

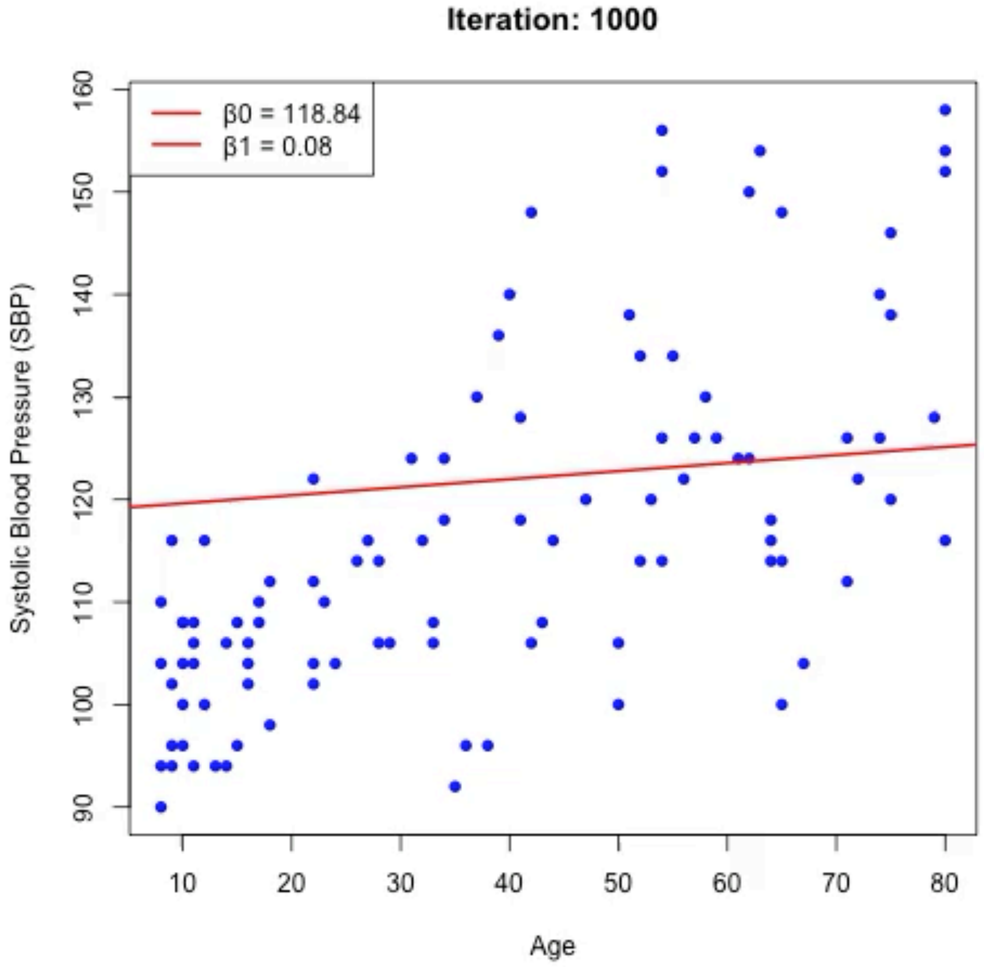
For linear regression the gradient of RSS with respect to β_0 and β_1 are:

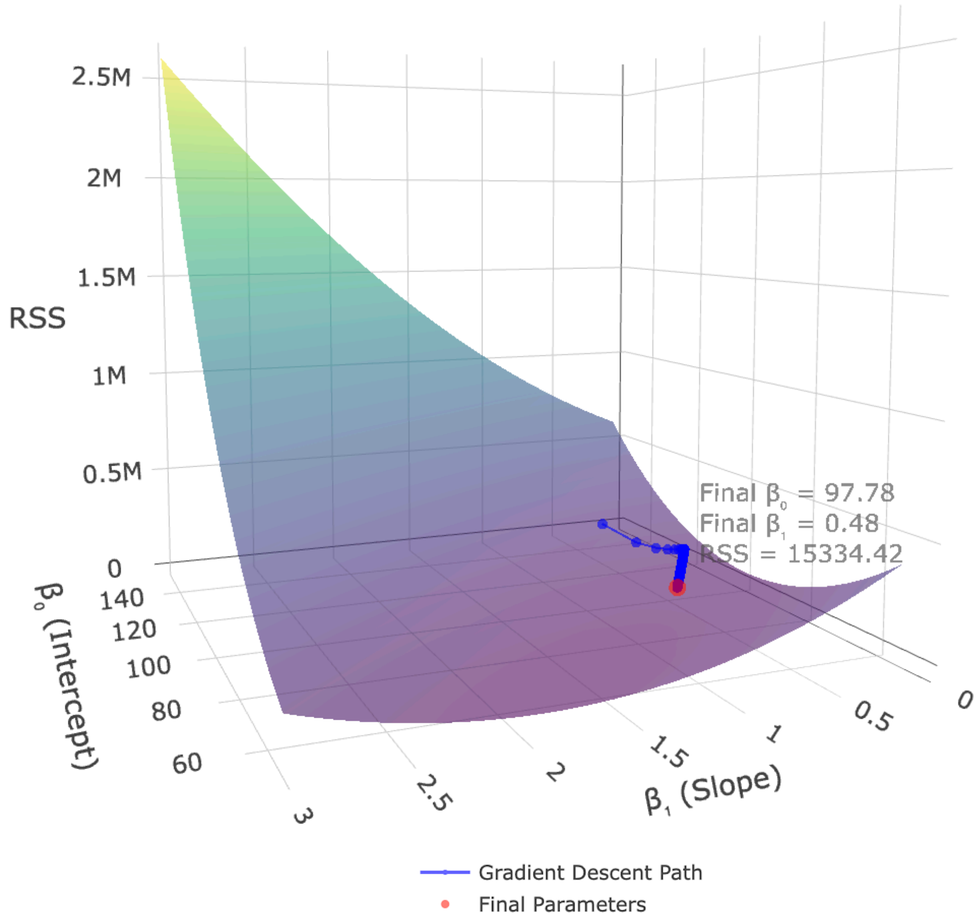
$$\frac{\partial RSS}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))$$

$$\frac{\partial RSS}{\partial \beta_1} = -2 \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i)) x_i$$


Steep slope,
take large steps
Less steep slope,
take smaller steps
Goal

Example: gradient descent





With higher values of ρ will the algorithm converge

$$\beta_0^{(t+1)} = \hat{\beta}_0^{(t)} - \rho \frac{\partial RSS}{\partial \beta_0}$$

$$\beta_1^{(t+1)} = \hat{\beta}_1^{(t)} - \rho \frac{\partial RSS}{\partial \beta_1}$$

Where:

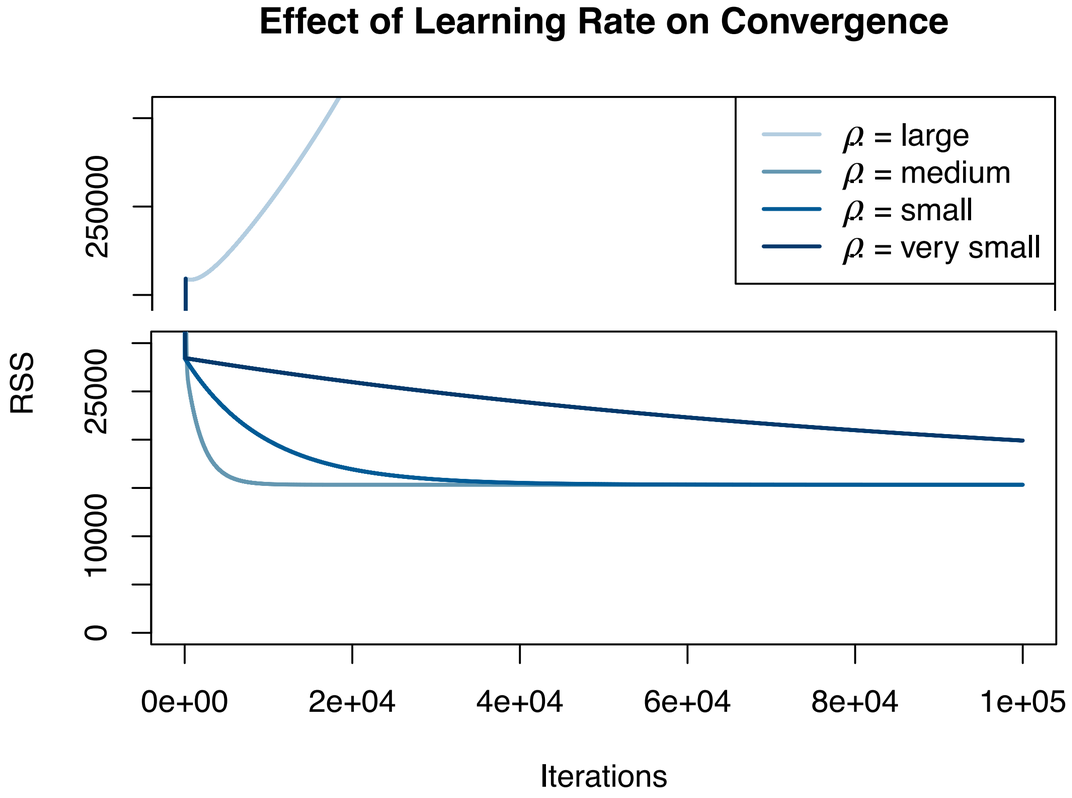
ρ is the learning rate controlling the size of the steps

t is the iteration index

1. Faster
2. Slower
3. Does not affect the speed of convergence

Effect of the learning rate on convergence

With gradient descent there are two hyperparameters that need to be defined: the learning rate ρ and the total number of iterations



Large	The algorithm overshoots the minimum, leading to oscillations or even divergence (increasing error).
Medium/ small	The algorithm quickly converges to the minimum without overshooting.
Very small	The algorithm takes many iterations to reach the minimum because each step is tiny

Select characteristics of least squares

- Exact solution
- May get stuck in local minima
- Not scalable to complex data/models
- Memory efficient
- Requires hyper parameters tuning
- Flexible use for different models
- Fast

Summary least squares vs gradient descent

Least squares

- Exact solution
- Fast
- Not scalable to complex data/ models

Gradient descent

- Flexible use for different models
- Memory efficient
- Requires hyper parameters tuning
- May get stuck in local minima

Multiple linear regression

When multiple predictor variables are available they can be combined using a multiple linear regression model

$$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

where

p is the number of predictors

X_j us the jth predictor

β_j is the jth coefficient (i.e. association between predictor j and the response)

Estimating multiple linear regression coefficients

Also for the multiple linear regression, given a training dataset we can estimate the values for the coefficients and we can use them to make predictions

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_p x_p$$

If we have a training data consisting in n observation pairs

$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, where each $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$

using the least squares approach we can choose $\beta_0, \beta_1, \ldots, \beta_p$ to minimise the RSS

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \ldots \hat{\beta}_p x_{ip})^2$$

Example: model formulation

Apart from Age there can be other factors affecting Systolic Blood Pressure (SBP), I can use multiple linear models to test the effect of adding more predictors

$$SBP \approx \beta_0 + \beta_1 \times Age + \beta_2 \times Gender + \beta_3 \times BMI + \beta_4 \times Cholesterol + \beta_4 \times Teeth$$

From the NHANES study we have data where SBP, age, gender, BMI, total cholesterol and number of teeth were measured for 2000 individuals and we can use these data to estimate model parameters.

What are the values of n and p in this example?

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_p x_p$$

Training data consisting in n observation pairs

$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, where each $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$

In our example we have 1000 individuals with measures of SBP, age, gender, BMI, total cholesterol and number of teeth.

n = , p =

Quantitative and qualitative predictors

Predictors can be

- quantitative (e.g. Age, BMI, Cholesterol, Teeth)

  They should be standardised (subtract mean and divide by standard deviation) to be able to compare the corresponding coefficient. Standardisation does not affect model performances.

- qualitative (e.g. Gender).

  How they are encoded does not affect model performances just their interpretation.

Use of linear regression models

Linear regression models can be used for different purposes:

- Understand the effect of the different input variables
  - Which variables are more informative in predicting the output?
  - What is the association between the input and output variables?

- Make predictions
  - Can I predict the output variable given the input variables and and which model should I use?

Example: understanding estimated coefficients

We run linear regression on the example data:

- The quantitative variables (Age, BMI, Cholesterol, Teeth) were standardised
- The qualitative variable Gender was encoded as 0 (Female), 1 (Male)

And obtained the following estimates for the coefficients

	Estimate	**P-value**	
(Intercept)	119.9923	< 2e-16	***
Age	10.0512	< 2e-16	***
Gender	2.3140	0.000674	***
BMI	2.9384	3.11e-16	***
TC	1.4717	3.56e-05	***
Teeth	-0.5508	0.148024	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

NOTE: The estimated coefficients are affected by error as they depend on the used training set. It is possible to assess the accuracy of the coefficients (as standard error or as p-value).

Is there a positive or negative association between BMI and SBP?

We run linear regression on the example data:

- The quantitative variables (Age, BMI, Cholesterol, Teeth) were standardised
- The qualitative variable Gender was encoded as 0 (Female), 1 (Male)

	Estimate	**P-value**	
(Intercept)	119.9923	< 2e-16	***
Age	10.0512	< 2e-16	***
Gender	2.3140	0.000674	***
BMI	2.9384	3.11e-16	***
TC	1.4717	3.56e-05	***
Teeth	-0.5508	0.148024	

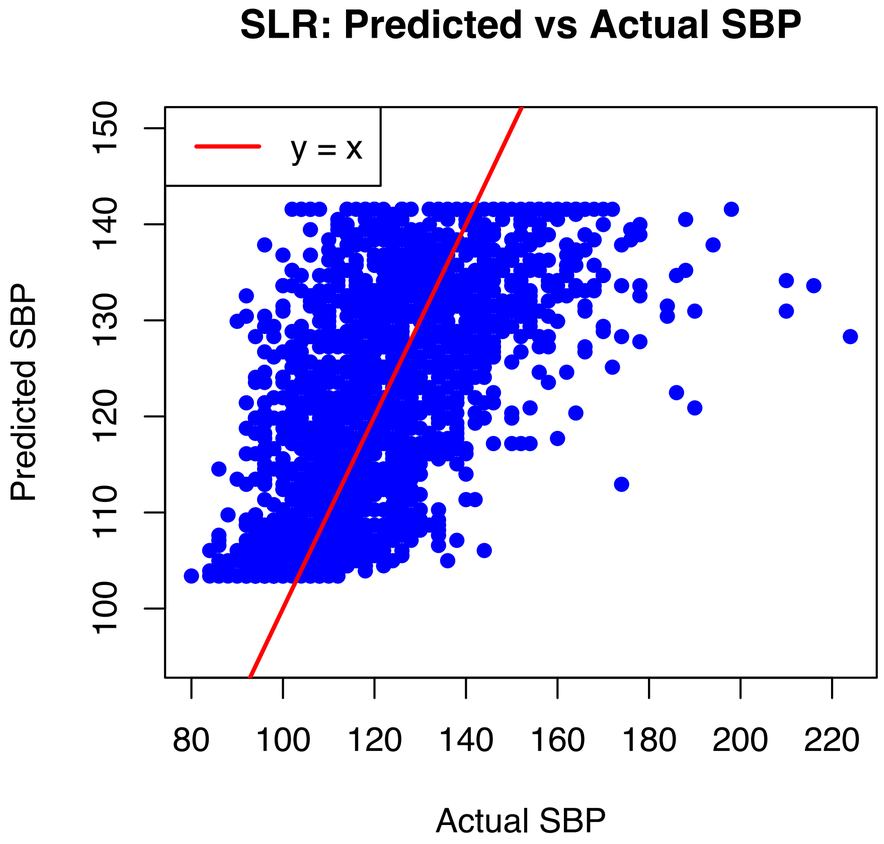
Is SBP higher in men or women?

We run linear regression on the example data:

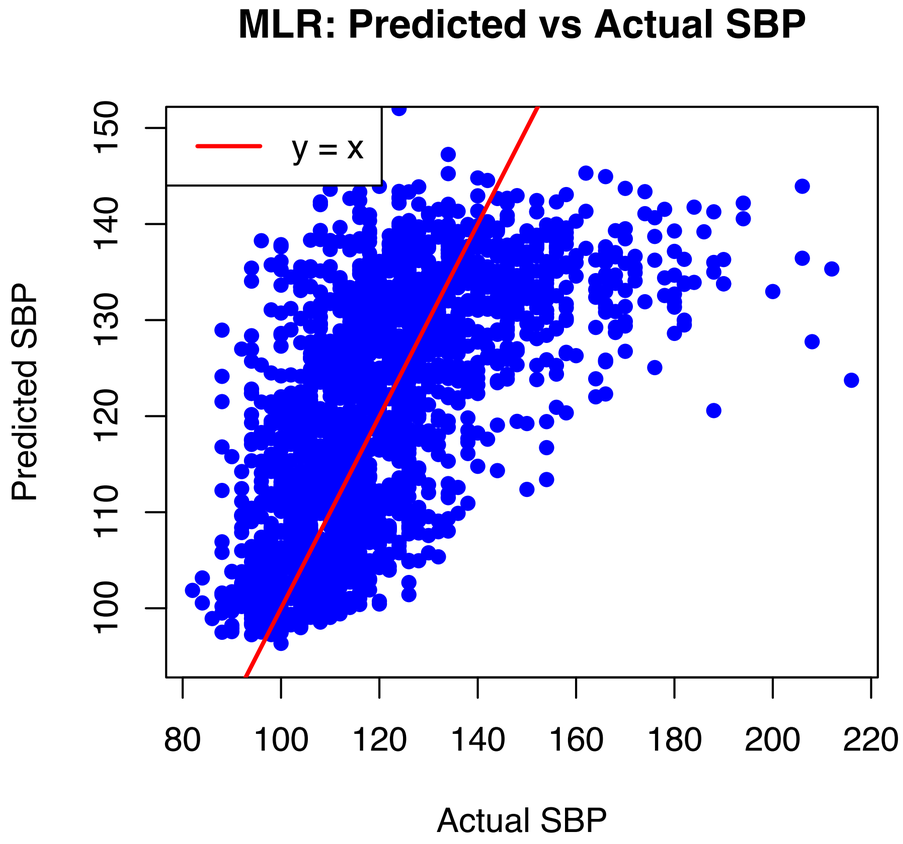
- The quantitative variables (Age, BMI, Cholesterol, Teeth) were standardised
- The qualitative variable Gender was encoded as 0 (Female), 1 (Male)

	Estimate	**P-value**	
(Intercept)	119.9923	< 2e-16	***
Age	10.0512	< 2e-16	***
Gender	2.3140	0.000674	***
BMI	2.9384	3.11e-16	***
TC	1.4717	3.56e-05	***
Teeth	-0.5508	0.148024	

Example: model predictions



RSS = 445863.3
MSE = 222.93
Pearson correlation = 0.62
AIC = 16495.48



RSS = 416501.2
MSE = 208.25
Pearson correlation = 0.65
AIC = 16367.24

Extension of the linear models

The model remains linear in the parameters even if it is not additive with respect to the input variables.

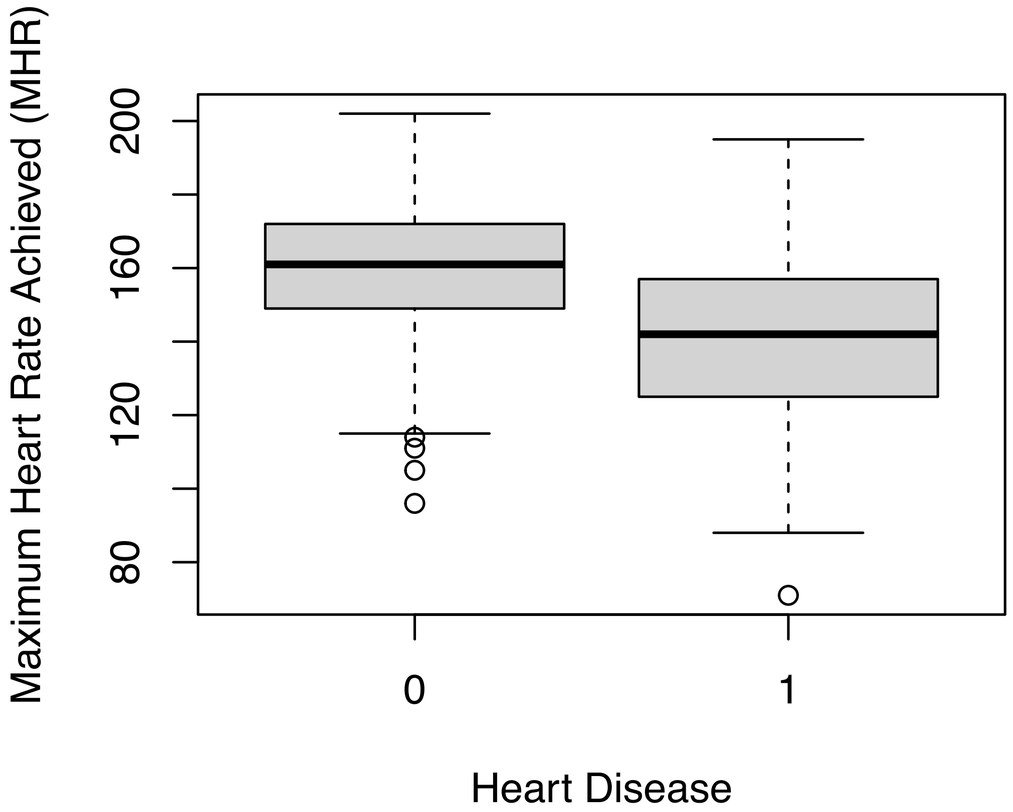
The model is still linear if the input variables X_j:

- Are polynomial, e.g. $X_2 = X_1^2$
- Derive from interactions between variables, e.g. $X_3 = X_1 \cdot X_2$

Linear models for qualitative outputs

Often in biomedical problems we want to predict a categorical variable.

Example: predict heart disease based on the maximum heart rate achieved during a stress test



Data from the Cleveland Heart Disease Dataset (downloaded from UCI Machine Learning Repository)

Classification problems: the Bayes classifier

We want to assign each observation to the most likely class, given its predictor values.

This means to assign an observation with predictor vector x_0 to a class j for which:

$$Pr(Y = j \,|\, X = x_0)$$

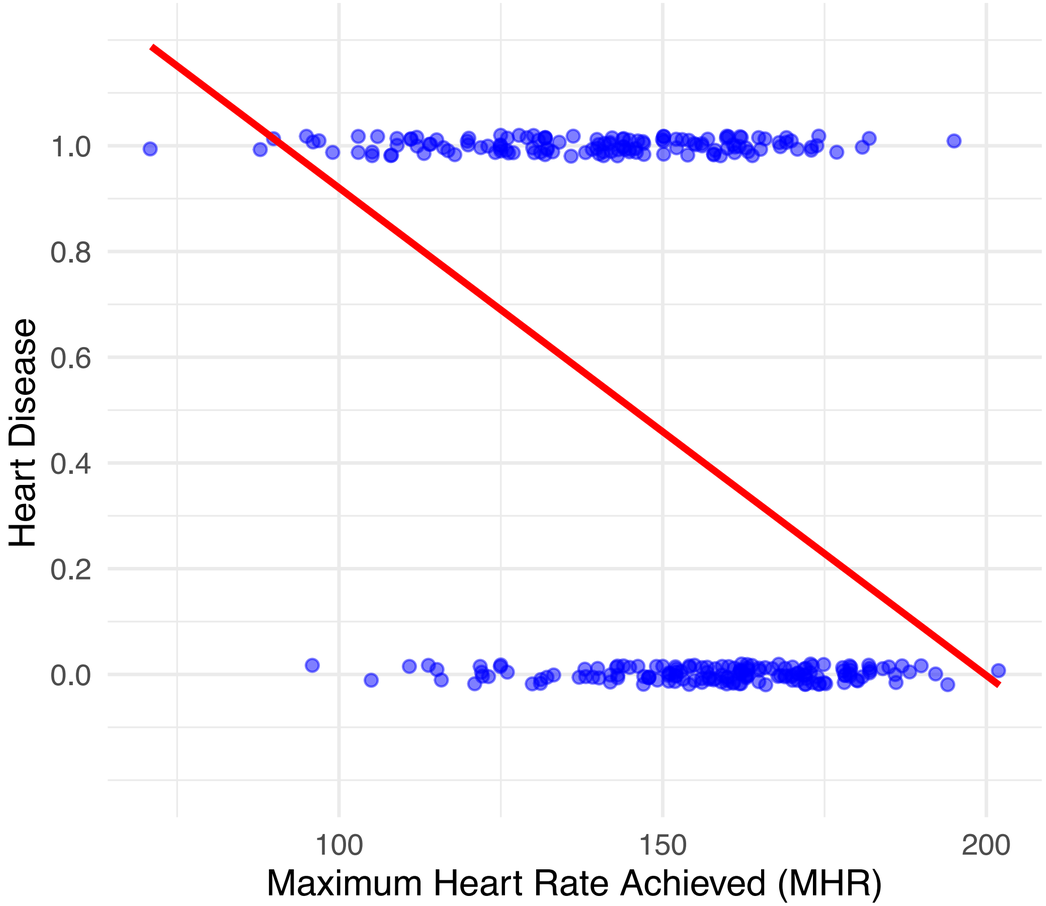
This is the conditional probability that $Y = j$, given the observed predictor vector x_0.

For a binary classifier where $Y = \{0,1\}$, the Bayes classifier predicts:

- class 1 if $Pr(Y = 1 \,|\, X = x_0) > 0.5$
- class 0 if $Pr(Y = 1 \,|\, X = x_0) \leq 0.5$

Example: Can we use a linear regression model?

$$\textit{Heart Disease} \approx \beta_0 + \beta_1 \times \textit{MHR}$$

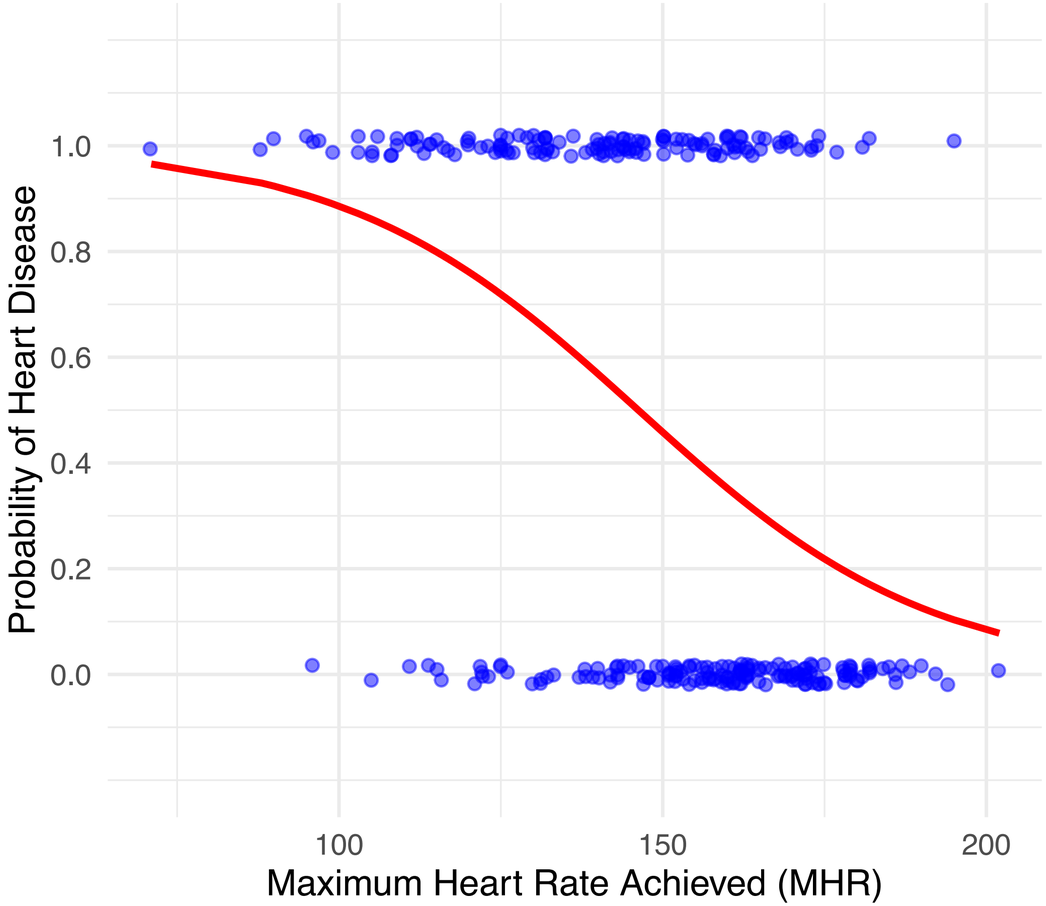


Problems:

- It can produce predictions outside the [0,1] range that cannot be interpreted as probabilities.
- Performs poorly at the extremes (close to true probability 0 or 1) as it can't capture binary nature of data.
- Cannot be applied to multi class problems.

Example: Modelling probabilities

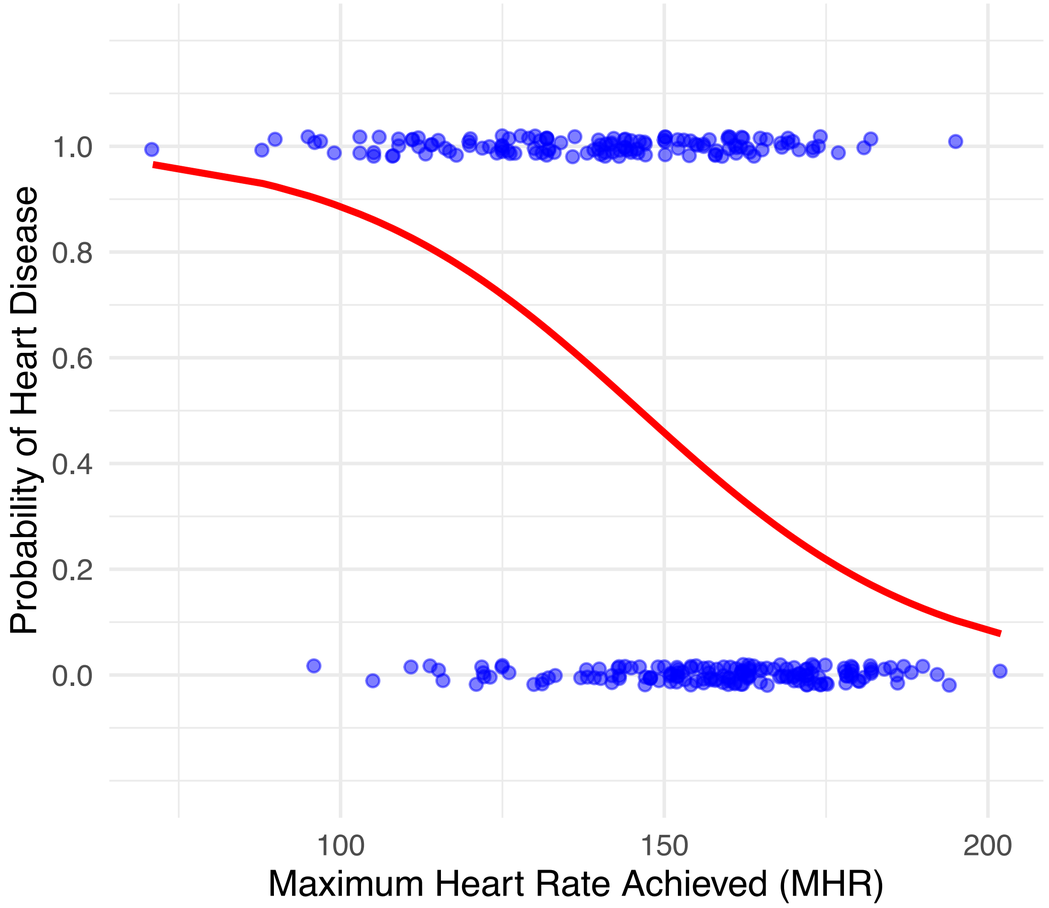
$$Pr(Heart\,Disease = Yes \mid MHR) \quad \text{or} \quad p(MHR)$$



- Probabilities range between 0 and 1
- Thresholds for binary classification can be set based on the application

Which threshold would you select in the following scenario?

Imagine you are in clinical setting where you want to apply the model as a first screening approach to define which patients should have a follow up exam to confirm the presence of heart disease.



- 0.9
- 0.1

Logistic regression

In logistic regression the relationship between $p(X) = Pr(Y = 1 \mid X)$ and X is modelled using the *logistic function:*

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}$$

This can be rewritten as *odds* in the following way:

$$\frac{p(X)}{1 - p(X)} = e^{\beta_0+\beta_1 X}$$

And taking the logarithm we obtain the *log odds* or *logit*, which are linear in X

$$\log\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta_1 X$$

Estimating the coefficients with maximum likelihood

Given a training data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ we can estimate the model coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$ that maximise the *likelihood function*:

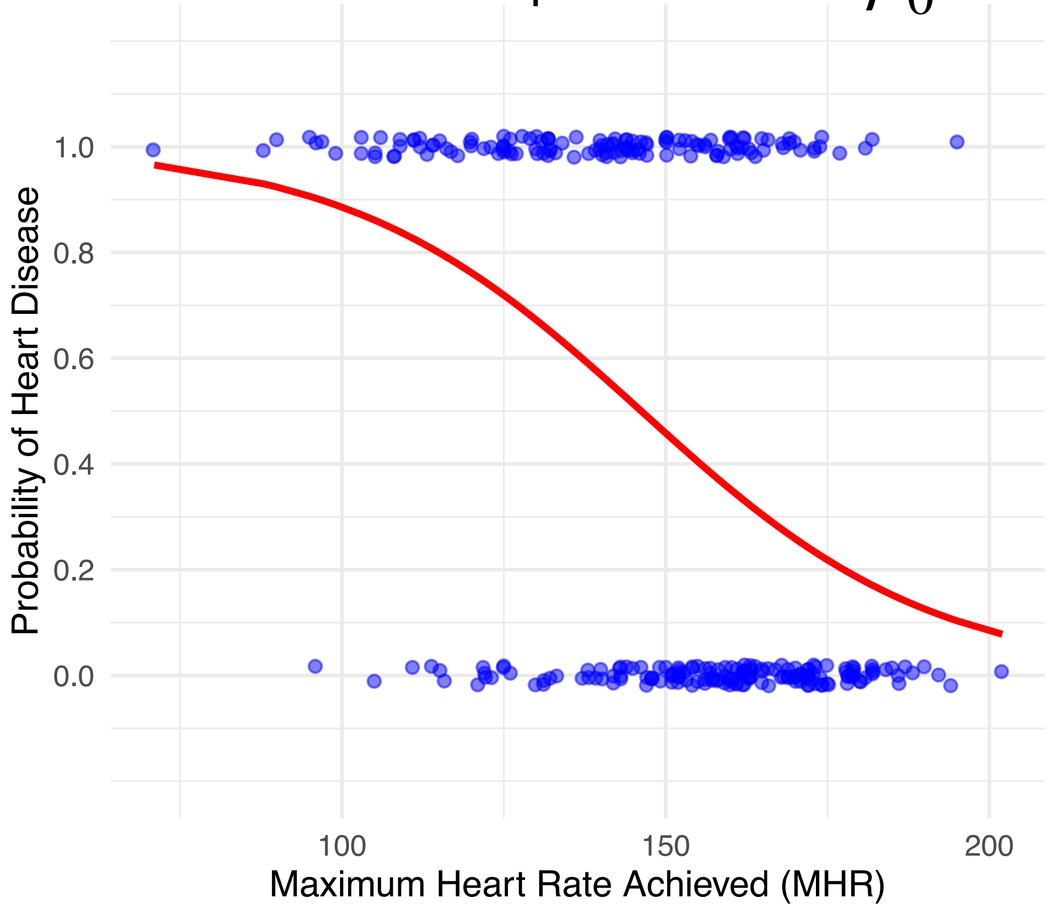
$$\ell(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0} (1 - p(x_i'))$$

i.e. it gives a probability close to 0 for observations in class 0 and close to 1 for observations in class 1.

NOTE: similar to what we have seen for linear regression, gradient descent is not the standard way to solve this optimisation problem, but is a flexible approach that can be used as you will see during the practical.

Example: Interpreting parameters and making predictions

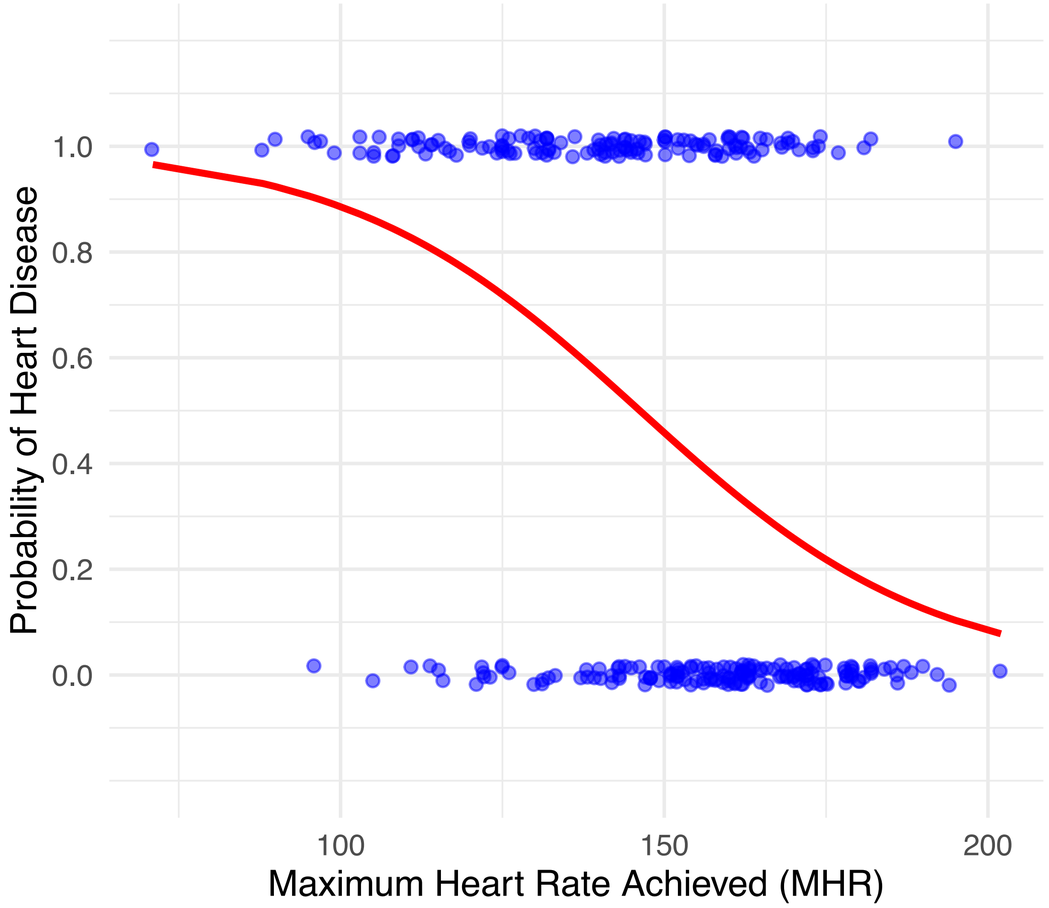
Using observations from 296 patients, we used maximum likelihood to estimated the model parameters $\hat{\beta}_0 = 6.48$ and $\hat{\beta}_1 = -0.04$



We can now make predictions using the formula:

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{e^{6.48-0.04X}}{1+e^{6.48-0.04X}}$$

What is the predicted probability of heart disease for MHR = 100



- 1
- 0.88
- 0.23

$$\hat{p}(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}} = \frac{e^{6.48-0.04X}}{1+e^{6.48-0.04X}}$$

Multiple logistic regression

Can be used to predict a categorical response using multiple predictors.
In case of a binary output:

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X + \ldots + \beta_p X_p$$

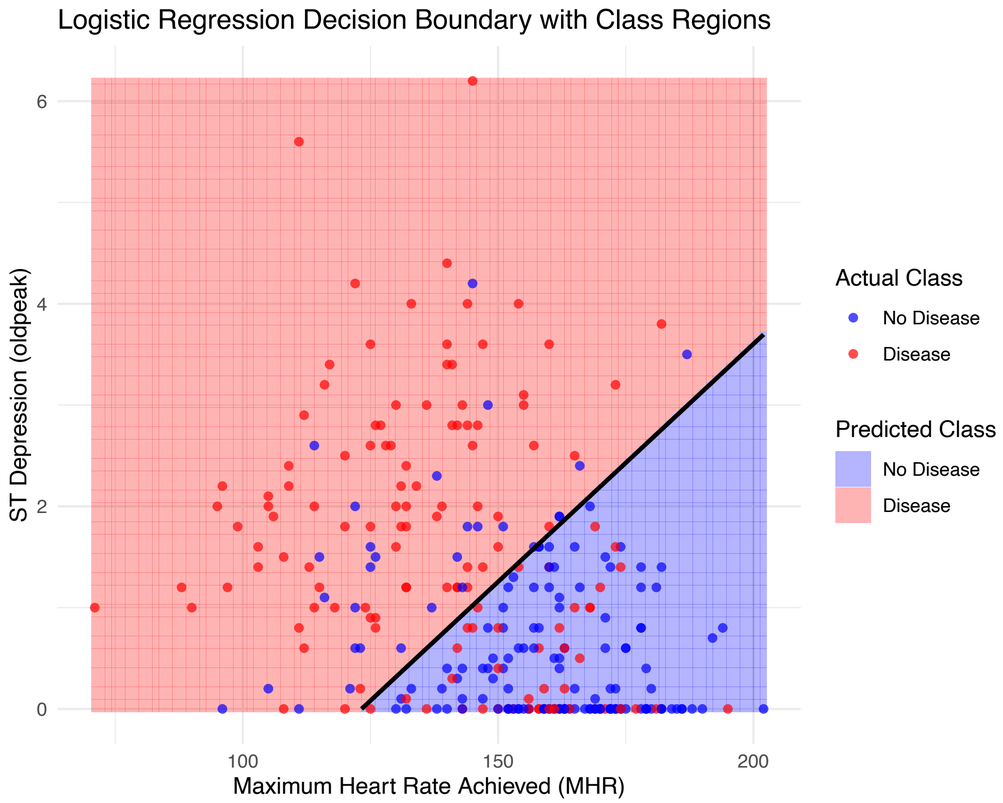
Where $X = (X_1, \ldots, X_p)$ are the p predictors. This equation can be rewritten as:

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\ldots+\beta_p X_p}}{1 + e^{\beta_0+\beta_1 X_1+\ldots+\beta_p X_p}}$$

Also for multiple logistic regression models maximum likelihood can be used to estimate the model parameters $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$

Example: Bivariate logistic regression

We now use both *MHR* and *oldpeak*, which is another indicator of heart health measuring ST depression induced by exercise relative to rest



	Estimate	P-value	
(Intercept)	4.195315	7.89e-05	***
MHR	-0.034037	5.89e-07	***
oldpeak	0.724979	2.91e-07	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Example: Evaluating model performances

After setting a threshold (e.g. at 0.5) we can compare model predictions using a confusion matrix.

Bivariate model

		Actual	
		0	**1**
Predicted	**0**	130	48
	1	29	89

Class 1 corresponds to diseased patients

What is the number of true positives (TP)

		Actual	
		0	**1**
Predicted	**0**	130	48
	1	29	89

- 130
- 48
- 29
- 89

What is the number of false negatives (FN)

		Actual	
		0	**1**
Predicted	**0**	130	48
	1	29	89

- 130
- 48
- 29
- 89

Example: comparing model performances

Several metrics can be useful and which one are more relevant depend on the problem

$$Accuracy = \frac{TP + TN}{P + N}$$

$$Sensitivity = \frac{TP}{P}$$

$$Specificity = \frac{TN}{N}$$

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

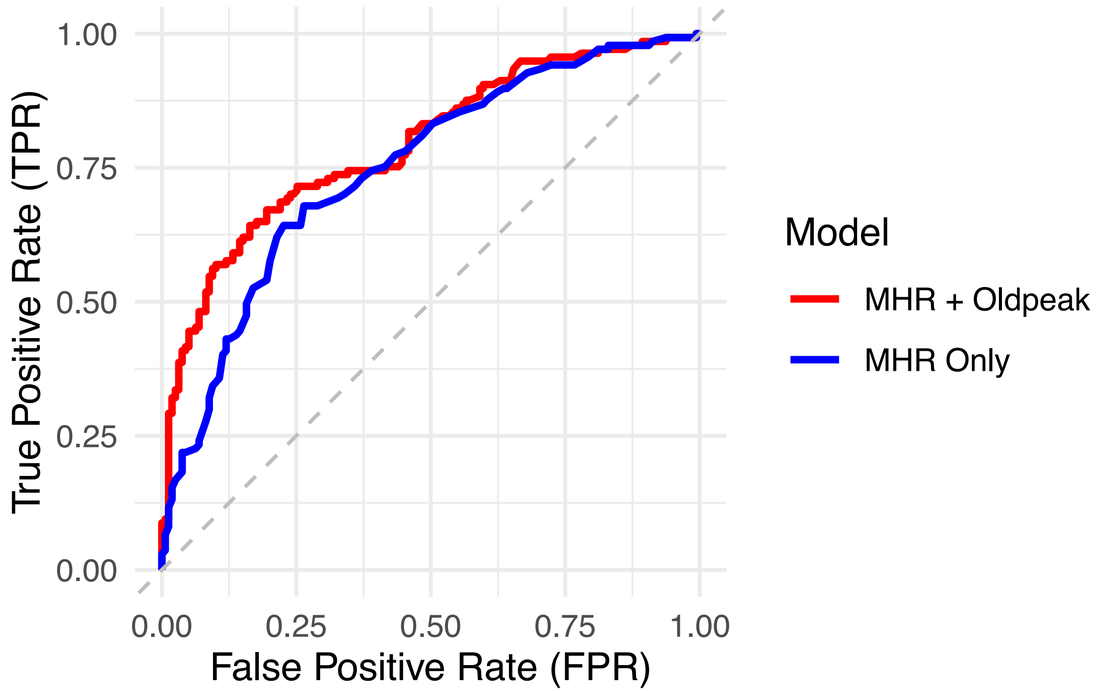
Metric	Bivariate	Univariate
Accuracy	0.74	0.71
Sensitivity	0.82	0.79
Specificity	0.65	0.62
Precision	0.73	0.71

Where $P = TP + FN$ and $N = FP + TN$

NOTE: this is not a full list of relevant evaluation metrics

Comparing classification performances

We can also use Receiver Operating Characteristic (ROC) curve to compare model performances across different thresholds.



$$TPR = \frac{TP}{TP + FN}$$ (Also called sensitivity or recall)

$$FPR = \frac{FP}{FP + TN}$$

MHR + Oldpeak: AUC = 0.79

MHR Only: AUC = 0.75

Diagonal line correspond to the performance of a random classifier (AUC = 0.5)

Multinomial logistic regression

Logistic regression can be extended to classify a response variable with more than two classes.

When we have K classes ($K > 2$) we consider an arbitrary K^{th} class as baseline.

$$Pr(Y = k|X = x) = \frac{e^{\beta_{k0}+\beta_{k1}x_1+\ldots+\beta_{kp}x_p}}{1 + \sum_{l=1}^{K-1} e^{\beta_{l0}+\beta_{l1}x_1+\ldots+\beta_{lp}x_p}} \qquad \text{for } k = 1,\ldots,K-1$$

$$Pr(Y = K|X = x) = \frac{1}{1 + \sum_{l=1}^{K-1} e^{\beta_{l0}+\beta_{l1}x_1+\ldots+\beta_{lp}x_p}}$$

Multinomial logistic regression

For $k = 1,\ldots,K-1$ this corresponds to

$$log\left(\frac{Pr(Y=k\,|\,X=x)}{Pr(Y=K\,|\,X=x)}\right) = \beta_{k0} + \beta_{k1}x_1 + \ldots + \beta_{kp}x_p$$

Depending of the choice of which class to use as baseline the estimated coefficients will change but the predictions and the log odds will not.

Questions?