# MSc Data Science for Business

# Introduction to Machine Learning

# Map 534

Julie Josse

# Logistic Regression

J. Josse

**Previous lecture: supervised learning**

- Framework
- SVM
- Linear discriminant analysis

**Today: Logistic regression**

# Outline

# Plan

# CHD example

Sample of males in a heart-disease high-risk region (Western Cape, South Africa).

```
- sbp      systolic blood pressure
- tobacco cumulative tobacco (kg)
- ldl     low density lipoprotein, bad cholesterol
- famhist family history of heart disease
- typea     type-A behavior
- obesity
- alcohol current alcohol consumption
- age     age
- chd     response, coronary heart disease
```

Model CDH as a bunch of coin flips with a success probability that depends on covariates

# CHD example

```
read.table("https://web.stanford.edu/~hastie/ElemStatLearn/datasets/SAheart.dat
```

```
  sbp tobacco  ldl adiposity famhist typea obesity alcohol age chd
1 160   12.00 5.73     23.11 Present    49   25.30   97.20  52 Yes
2 144    0.01 4.41     28.61  Absent    55   28.87    2.06  63 Yes
3 118    0.08 3.48     32.28 Present    52   29.14    3.81  46  No
4 170    7.50 6.41     38.03 Present    51   31.99   24.26  58 Yes
5 134   13.60 3.50     27.78 Present    60   25.99   57.34  49 Yes
6 132    6.20 6.47     36.21 Present    62   30.77   14.14  45  No
...
```

$n = 462$, 160 cases $(\text{cdh} = 1)$ and 302 control

# CHD example

## Questions

- **Analysis** : Understand which factors in this dataset are linked to the desease (chd)
  - Importance of the effect
  - Positive or negative effect
  - Significativity

  $\rightarrow$ Prevention
- **Prediction** : Predict, for a new patient, the risk to declare the desease.
  - Quality of the prediction
  - Parcimonious model

  $\rightarrow$ Better monitoring of the at-risk patient

# Scoring

## Credit Default, Credit Score, Bank Risk, Market Risk Management


- Data: Client profile, Client credit history...
- Input: Client profile
- Output: Credit risk

# Scoring exemple

We have at hands a dataset of 5380 customers

```
Observations: 5,380
Variables: 19
## $ Id_Customer         <int> 7440, 573, 9194, 3016, 6524, 3858, 2189, 9...
## $ Y                   <int> 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, ...
## $ Customer_Type       <chr> "Non Existing Client", "Existing Client", ...
## $ BirthDate           <chr> "07/08/1977", "13/06/1974", "07/11/1973", ...
## $ Customer_Open_Date  <chr> "13/02/2012", "04/02/2009", "03/04/2012", ...
## $ P_Client            <chr> "NP_Client", "P_Client", "NP_Client", "NP_...
## $ Educational_Level   <chr> "University", "University", "University", ...
## $ Marital_Status      <chr> "Married", "Married", "Married", "Married"...
## $ Number_Of_Dependant <int> 3, 0, 2, 3, 2, 0, 0, 0, 0, 4, 0, 0, 0, 0, ...
## $ Years_At_Residence  <int> 1, 12, 10, 3, 1, 28, 10, 15, 0, 35, 10, 10...
## $ Net_Annual_Income   <dbl> 36.000, 18.000, 36.000, 36.000, 36.000, 60...
## $ Years_At_Business   <int> 1, 2, 1, 1, 1, 2, 1, 1, 3, 2, 3, 2, 4, 1, ...
## $ Prod_Sub_Category   <chr> "C", "C", "C", "C", "C", "C", "C", "C", "P...
## $ Prod_Decision_Date  <chr> "14/02/2012", "30/06/2011", "04/04/2012", ...
## $ Source              <chr> "Sales", "Sales", "Sales", "Sales", "Sales...
## $ Type_Of_Residence   <chr> "Owned", "Parents", "Owned", "New rent", "...
## $ Nb_Of_Products      <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 1, 1, ...
## $ Prod_Closed_Date    <chr> NA, NA, NA, "31/12/2012", NA, NA, NA, "16/...
## $ Prod_Category       <chr> "B", "G", "B", "L", "D", "C", "B", "B", "E...
```

# Scoring exemple

## Questions

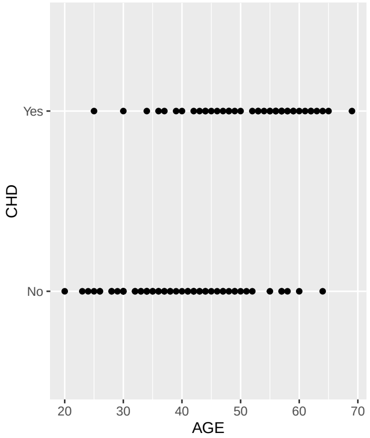
- **Analysis** : Understand which factors in this dataset are linked to the default
  - Importance of the effect
  - Positive or negative effect
  - Significativity

  $\rightarrow$ Customer analysis
- **Prediction** : Predict, for a new client, the risk to default.
  - Quality of the prediction
  - Parcimonious model

  $\rightarrow$ Decision aid.

- Similar questions in most setting: analysis / prediction...

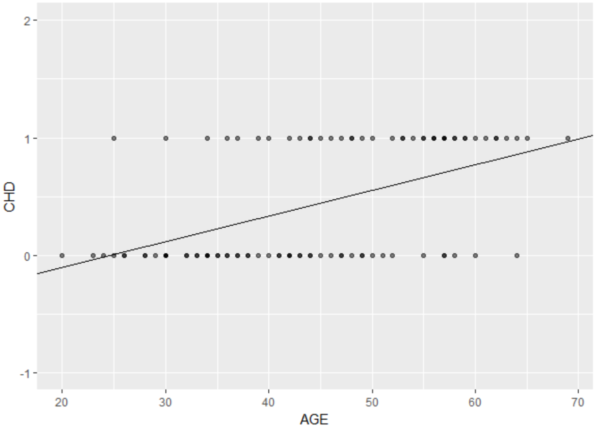
# CHD exemple




**Goal:** Predict **CHD** from the **AGE**.

# First attempt

Is it possible to use linear regression?




How can we interpret it?

# Binned Data

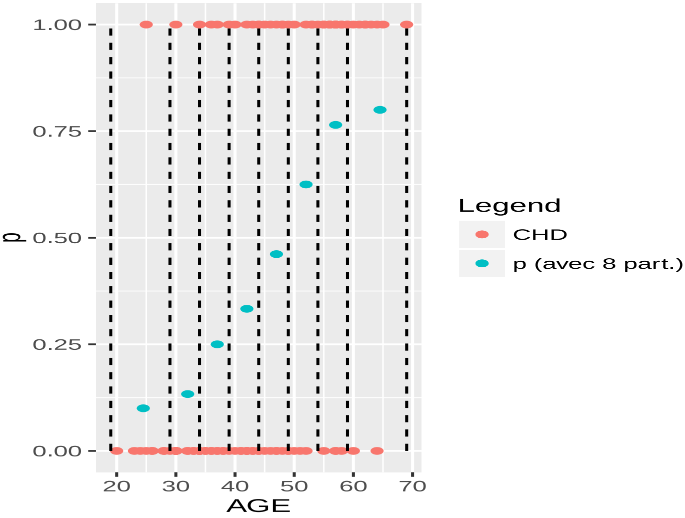
## Binned Data

- Gather all the observation having a similar **AGE** and estimate the proportion in this class.
- Easy estimation by the empirical proportion.

| $\mathrm{Age}_k$ | $c_k$ | $n_k$ | $n_k$[CHD=0] | $n_k$[CHD=1] | $\pi_k$ |
|---|---|---|---|---|---|
| [20,29] | 24.5 | 10 | 9 | 1 | 0.10 |
| [30,34] | 32 | 15 | 13 | 2 | 0.13 |
| [35,39] | 37 | 12 | 9 | 3 | 0.25 |
| [40,44] | 42 | 15 | 10 | 5 | 0.33 |
| [45,49] | 47 | 13 | 7 | 6 | 0.46 |
| [50,54] | 52 | 8 | 3 | 5 | 0.63 |
| [55,59] | 57 | 17 | 4 | 13 | 0.76 |
| [60,69] | 64.5 | 10 | 2 | 8 | 0.80 |

**Estimated proportions**

# Binned Data




**Estimated proportions**

# Bernoulli distribution

- Data: outcome $Y \in \{0,1\}$, covariates $\mathbf{X} \in \mathbb{R}^d$
- Goal: predict the probability that $Y$ is 0 or 1 given $\mathbf{X} \in \mathbb{R}^d$
- Bernoulli $\mathcal{B}(p)$: law on $\{0,1\}$ such that

$$Y \sim \mathcal{B}(p) \Leftrightarrow \begin{cases} \mathbb{P}\left(Y=1\right)=p \\ \mathbb{P}\left(Y=0\right)=1-p \end{cases}$$

## Conditional Bernoulli Law

$$\begin{aligned} Y_i|\mathbf{X}=\mathbf{x}_i &\sim \mathcal{B}\left(p_i\right) \\ Y_i|\mathbf{X}=\mathbf{x}_i &\sim \mathcal{B}\left(\mathbb{P}\left(Y_i=1|\mathbf{X}=\mathbf{x}_i\right)\right) \end{aligned}$$

$\Rightarrow$ Model the probability to have a disease.

$\Rightarrow$ It's a bunch of coin flips. However, the success probability will differ from one person to the other depending on their covariates

# Plan

# Logistic model

## Probabilistic model

- $Y_i | \mathbf{X} = \mathbf{x}_i \sim \mathcal{B}(p_i)$
- $E(Y_i | \mathbf{X} = \mathbf{x}_i) = p_i$
- We assume linearity in $\mathbf{X}$: $\eta_i = \sum_j \beta_j x_{ij}$

$\Rightarrow$ We need to apply some transformation to the $\eta_i$ that can vary between $-\infty$ and $+\infty$ so that it varies between 0 and 1.

## Logistic regression, S shape

$E(Y_i | \mathbf{X} = \mathbf{x}_i) = p_i = \frac{e^{\eta_i}}{1+e^{\eta_i}}$

$p_\beta(\mathbf{x}_i) = \frac{e^{\mathbf{x}_i^t \beta}}{1+e^{\mathbf{x}_i^t \beta}}$

# Logistic model

## The expit function

$$E(Y_i|\mathbf{X} = \mathbf{x}_i) = p_i = p_\beta(\mathbf{x}_i) = \frac{e^{\eta_i}}{1+e^{\eta_i}} = \frac{e^{\mathbf{x}_i^t\beta}}{1+e^{\mathbf{x}_i^t\beta}}$$

## Logistic Regression: odds: $(0;+\infty)$ log odds $(-\infty;+\infty)$ $p_i$ $(0;1)$

$$\begin{aligned}
\log\left(\frac{p_i}{1-p_i}\right) &= \eta_i. \\
g : t \mapsto \log\left(\frac{t}{1-t}\right), \text{logit function } g(p_i) &= \eta_i, \\
\log\left(odds\right) &= \eta_i. \\
\log\left(\frac{\mathbb{P}\left(Y_i = 1|\mathbf{X} = \mathbf{x}_i\right)}{\mathbb{P}\left(Y_i = 0|\mathbf{X} = \mathbf{x}_i\right)}\right) &= \sum_j \beta_j x_{ij}.
\end{aligned}$$

# Logistic model

## The expit function

$$E(Y_i|\mathbf{X}=\mathbf{x}_i) = p_i = p_\beta(\mathbf{x}_i) = \frac{e^{\eta_i}}{1+e^{\eta_i}} = \frac{e^{\mathbf{x}_i^t\beta}}{1+e^{\mathbf{x}_i^t\beta}}$$

## Logistic Regression: odds: $(0;+\infty)$ log odds $(-\infty;+\infty)$ $p_i$ $(0;1)$

$$\begin{aligned}
\log\left(\frac{p_i}{1-p_i}\right) &= \eta_i.\\
g: t \mapsto \log\left(\frac{t}{1-t}\right), \text{logit function } g(p_i) &= \eta_i,\\
\log\left(odds\right) &= \eta_i.\\
\log\left(\frac{\mathbb{P}\left(Y_i=1|\mathbf{X}=\mathbf{x}_i\right)}{\mathbb{P}\left(Y_i=0|\mathbf{X}=\mathbf{x}_i\right)}\right) &= \sum_j \beta_j x_{ij}.
\end{aligned}$$

# Logistic model

## All in all: Logistic Regression

$Y \in \{0, 1\}$

$$P(Y_i = 1|\mathbf{X} = \mathbf{x}_i) = p_i = \frac{e^{\mathbf{x}_i^t \beta}}{1 + e^{\mathbf{x}_i^t \beta}} = \frac{1}{1 + e^{-\mathbf{x}_i^t \beta}}$$

$$P(Y_i = 0|\mathbf{X} = \mathbf{x}_i) = 1 - p_i = \frac{1}{1 + e^{\mathbf{x}_i^t \beta}}$$

## All in all: Logistic Regression

$Y \in \{-1, 1\}$

$$P(Y_i = y_i|\mathbf{X} = \mathbf{x}_i) = \frac{1}{1 + e^{-y_i \mathbf{x}_i^t \beta}}$$

# Logistic model

## All in all: Logistic Regression

$Y \in \{0, 1\}$

$$
\begin{aligned}
P(Y_i = 1 | \mathbf{X} = \mathbf{x}_i) &= p_i = \frac{e^{\mathbf{x}_i^t \beta}}{1 + e^{\mathbf{x}_i^t \beta}} = \frac{1}{1 + e^{-\mathbf{x}_i^t \beta}} \\
P(Y_i = 0 | \mathbf{X} = \mathbf{x}_i) &= 1 - p_i = \frac{1}{1 + e^{\mathbf{x}_i^t \beta}}
\end{aligned}
$$

## All in all: Logistic Regression

$Y \in \{-1, 1\}$

$$
P(Y_i = y_i | \mathbf{X} = \mathbf{x}_i) = \frac{1}{1 + e^{-y_i \mathbf{x}_i^t \beta}}
$$

How to estimate $\beta$?

# Plan

# Maximum Likelihood Estimate

Simulations

# Maximum Likelihood Estimate

## Likelihood

The likelihood of the model is defined as:

$$L_n(y_1, \dots, y_n, \beta) = \prod_{i=1}^{n} \mathbb{P}\left(Y = y_i | \mathbf{X} = \mathbf{x}_i\right)$$

which is simply denoted by $L_n(\beta)$.

Let us write the likelihood as a function of $\beta$:

$$L_n(\beta) = \prod_{i=1}^{n} \mathbb{P}\left(Y = y_i | \mathbf{X} = \mathbf{x}_i\right) = \prod_{i=1}^{n} p_\beta(\mathbf{x}_i)^{y_i} \left(1 - p_\beta(\mathbf{x}_i)\right)^{1-y_i}.$$

$$L_n(\beta) = \prod_{i=1}^{n} \mathbb{P}\left(Y = y_i | \mathbf{X} = \mathbf{x}_i\right) = \prod_{i=1}^{n} g^{-1}(\mathbf{x}_i^t \beta)^{y_i} (1 - g^{-1}(\mathbf{x}_i^t \beta))^{1-y_i}.$$

# Maximum Likelihood Estimate

$$
\begin{aligned}
L_n(\beta) &= \prod_{i=1}^{n} \left( \frac{e^{\mathbf{x}_i^t \beta}}{1 + e^{\mathbf{x}_i^t \beta}} \right)^{y_i} \left( \frac{1}{1 + e^{\mathbf{x}_i^t \beta}} \right)^{1-y_i} \\
&= \prod_{i=1}^{n} \left( \frac{e^{\mathbf{x}_i^t \beta y_i}}{1 + e^{\mathbf{x}_i^t \beta}} \right)
\end{aligned}
$$

## log-likelihood

$$
\log(L_n(\beta)) = \sum_{i=1}^{n} \left( y_i \mathbf{x}_i^t \beta - \log(1 + e^{\mathbf{x}_i^t \beta}) \right)
$$

# Score equation

$$\nabla\left(\log(L_n(\beta))\right) = \left(\frac{\partial}{\partial\beta_0}(\beta), \cdots, \frac{\partial}{\partial\beta_d}(\beta)\right)$$

$$\begin{aligned}\frac{\partial\left(\log(L_n(\beta))\right)}{\partial\beta_j} &= \sum_{i=1}^{n}\left(y_i x_{ij} - \frac{x_{ij}e^{\mathbf{x}_i^t\beta}}{(1+e^{\mathbf{x}_i^t\beta})}\right)\\ &= \sum_{i=1}^{n} x_{ij}\left(y_i - p_\beta(\mathbf{x}_i)\right)\end{aligned}$$

$$S(\beta) = \mathbf{X}'(Y - P_\beta) = 0$$

# Maximum Likelihood Estimate

## Unfortunately...

- For linear regression, we have an explicit expression for the maximizer $\widehat{\beta}$ of the likelihood:

$$\widehat{\beta} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t Y.$$

- For logistic regression, there is no explicit expression for the maximizer $\widehat{\beta}$ of the likelihood.

## Fortunately...

Generally, likelihood has a unique maximum, and there exist plenty of numeric algorithms to find this maximum

# Maximum Likelihood Estimate

## Hypothesis

- H1: rank of $\mathbf{X} = d$
- H2: No separability: there exists $\beta$ $\forall i$, $Y_i = 1$ when $\mathbf{x}_i^t \beta \geq 0$ and $\forall i$, $Y_i = 0$ when $\mathbf{x}_i^t \beta \leq 0$

1. Under H1 and H2, the log-likelihood $\beta \to L_n(\beta)$ is strictly concave so that $\hat{\beta}$ exists and is unique.

1. Maximize $L_n(\beta)$ or minimizing $\mathcal{D} = -2L_n(\beta)$

# IRWLS algorithm

Iterative Reweighted Least Squares - Newton-Raphson.

## Univariate $\beta$

- Init: $\beta^0$
- Let us note $\beta^1 = \beta^0 + h$ a candidate solution of $S(\beta) = 0$, $S(\beta^0 + h) = 0$
- First order Taylor: $S(\beta^0 + h) \approx S(\beta^0) + hS'(\beta^0)$ $h = \frac{-S(\beta^0)}{S'(\beta^0)}$

$$\beta^1 = \beta^0 - \frac{S(\beta^0)}{S'(\beta^0)}$$

# IRWLS algorithm

## Vector $\beta \in \mathbb{R}^d$

$$S(\beta) = \nabla(\log L_n(\beta))$$

$$H = \nabla^2(\log L_n(\beta))_{(k,l)} = \frac{\partial^2 \log L_n}{\partial_{\beta_k}\partial_{\beta_l}}$$

## Algorithm

- Init: $\beta^0$
- $\beta^1 = \beta^0 - \{\nabla^2(\log L_n(\beta))\}^{-1}\nabla(\log L_n(\beta))$
- $\beta^{k+1} = \beta^k + A^k\nabla(\log L_n(\beta^k))$
- $k \leftarrow k+1$
- Stop when $\beta^{k+1} \approx \beta^k$

# IRWLS algorithm

## Exercice

Show that

$$H = \mathbf{X}^t \mathbf{W}_\beta \mathbf{X}$$

with $\mathbf{W}_{\beta_{n\times n}} = -\operatorname{diag}(p_\beta(\mathbf{x}_i)(1 - p_\beta(\mathbf{x}_i))_{i=1,\cdots,n})$

## Weighted regression

$$\begin{aligned}
\beta^{k+1} &= \beta^k + (\mathbf{X}^t \mathbf{W}_{\beta^\mathbf{k}} \mathbf{X})^{-1} \mathbf{X}'(\mathbf{Y} - \mathbf{P}_{\beta^\mathbf{k}}) \\
&= (\mathbf{X}^t \mathbf{W}_{\beta^\mathbf{k}} \mathbf{X})^{-1} \mathbf{X}^\mathbf{t} \mathbf{W}_{\beta^\mathbf{k}} \left( \mathbf{X}\beta^\mathbf{k} + \mathbf{W}_{\beta^\mathbf{k}}^{-1} (\mathbf{Y} - \mathbf{P}_{\beta^\mathbf{k}}) \right) \\
&= (\mathbf{X}^t \mathbf{W}_{\beta^\mathbf{k}} \mathbf{X})^{-1} \mathbf{X}^\mathbf{t} \mathbf{W}_{\beta^\mathbf{k}} \mathbf{Z}^\mathbf{k}
\end{aligned}$$

# Plan

# Logistic Regression CDH on Age

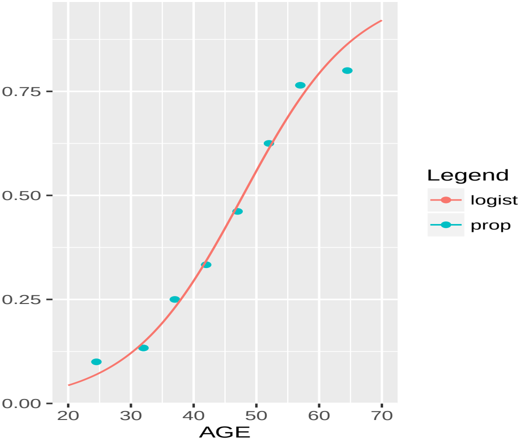
```
> CHD.logit = glm(CHD~AGE, family=binomial(link="logit"))
> summary(CHD.logit)

Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.30945    1.13365  -4.683 2.82e-06 ***
AGE          0.11092    0.02406   4.610 4.02e-06 ***
---

Number of Fisher Scoring iterations: 4
```

$$\log\left(\frac{\mathbb{P}\left(\mathrm{CHD}=1|\mathbf{X}\right)}{\mathbb{P}\left(\mathrm{CHD}=0|\mathbf{X}\right)}\right)=\underbrace{-5.31}_{\hat{\beta}_0}+\underbrace{0.11}_{\hat{\beta}_1}\times\mathrm{AGE}.$$

# Logistic Regression CDH on Age



# Logistic coefficient interpretation: odds-ratio

$$\log\left(\frac{\mathbb{P}\left(\mathrm{CHD}=1|\mathbf{X}\right)}{\mathbb{P}\left(\mathrm{CHD}=0|\mathbf{X}\right)}\right)=\underbrace{-5.31}_{\hat{\beta}_0}+\underbrace{0.11}_{\hat{\beta}_1}\times\mathrm{AGE}.$$

$\beta_j$: increases in the log-odds when $x^{(j)}$ increases by one unit, **the other variables begin fixed** (Be careful when interpreting, simple regression different from multiple regression)

# Logistic coefficient interpretation: odds-ratio

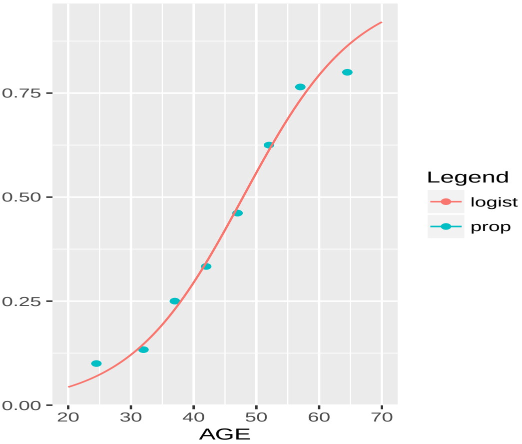
$$\log\left(\frac{\mathbb{P}\left(\mathrm{CHD}=1|\mathbf{X}\right)}{\mathbb{P}\left(\mathrm{CHD}=0|\mathbf{X}\right)}\right)=\underbrace{-5.31}_{\hat{\beta}_0}+\underbrace{0.11}_{\hat{\beta}_1}\times\mathrm{AGE}.$$

$\beta_j$: increases in the log-odds when $x^{(j)}$ increases by one unit, **the other variables begin fixed** (Be careful when interpreting, simple regression different from multiple regression)

## Interpretation

- A coefficient $\beta_1 = 0$ corresponds to no influence of $x^{(j)}$ on $Y$.
- A coefficient $\beta_1 < 0$ corresponds to a negative influence of $x^{(j)}$ on $Y$.
- A coefficient $\beta_1 > 0$ corresponds to a positive influence of $x^{(j)}$ on $Y$.

# Logistic Regression CDH on Age




$\beta_1$ does not correspond to the change in the proba associated with a one-unit increase in $X$. The amount that changes due to a one-unit change in $X$ will depend on the current value of $X$.

# Logistic coefficient interpretation: odds-ratio

$$Odds(\mathbf{x}_i) = \frac{p_i}{1 - p_i}$$

$$OddsRatio(\mathbf{x}_i, \tilde{\mathbf{x}}_i) = \frac{Odds(\mathbf{x}_i)}{Odds(\tilde{\mathbf{x}}_i)}$$

$\mathbf{x}_i$ and $\tilde{\mathbf{x}}_i$ differ only on the variable $j$ of one unit. (ex: To check)

$$OddsRatio(\mathbf{x}_i, \tilde{\mathbf{x}}_i) = e^{\beta_j}$$

The OR for $\mathrm{AGE}$ is $\exp(\beta_1) = \exp(0.11) = 1.117$

$$\log(\mathrm{OR}) = \beta_j$$

## Odds-ratio

It measures the evolution of the ratio of the probability of the event $Y = 1$ by the probability of the event $Y = 0$ when the variable $j$ goes from $x^{(j)}$ to $x^{(j)} + 1$, the other variables being constant. When $x^{(j)}$ increases by one unit, it multiplies the odds by $e^{\beta_j}$.

# Logistic coefficient interpretation: odds-ratio

It comes from the horse bets....
Be careful!

$$\mathrm{OR} = \left( \frac{\frac{\mathbb{P}\left(CDH=1|x^{(j)}=1\right)}{\mathbb{P}\left(CDH=0|x^{(j)}=1\right)}}{\frac{\mathbb{P}\left(CDH=1|x^{(j)}=0\right)}{\mathbb{P}\left(CDH=0|x^{(j)}=0\right)}} \right) = \exp(\beta_j)$$

so

$$\mathrm{OR} \approx \left( \frac{\mathbb{P}\left(CDH=1|x^{(j)}=1\right)}{\mathbb{P}\left(CDH=1|x^{(j)}=0\right)} \right)$$

only when the probabilities are very small...

# Plan

# Asymptotic Behavior of the MLE

**Theorem:**

a $\hat{\beta} \rightarrow_{a.s} \beta$ when $n \rightarrow \infty$

b $\sqrt{(}n)(\hat{\beta} - \beta) \rightarrow \mathcal{N}(0, I(\beta)^{-1})$ with the Fisher information $I(\beta)_{k,l} = -E\left(\frac{\partial^2 \log L_n}{\partial_{\beta_k} \partial_{\beta_l} L_n(\beta)}\right)$

**Theorem:**

$(\hat{\beta} - \beta)' nI(\beta)(\hat{\beta} - \beta) \rightarrow \chi^2_d$

**Law of large number + Slutsky Lemma + Convergence in Law**

- $\hat{I}(\beta) = \frac{1}{n}\mathbf{X}^t\mathbf{W}_\beta\mathbf{X}$ converge a.s. to $I(\beta)$
- $(\hat{\beta} - \beta)'(\mathbf{X}^t\mathbf{W}_{\hat{\beta}}\mathbf{X})(\hat{\beta} - \beta) \rightarrow \chi^2_{\mathbf{d}}$

# Asymptotic Confidence interval. Wald

## Asymptotic distributions

$$\frac{(\hat{\beta}_j - \beta_j)^2}{\hat{\sigma}_j^2} \to \chi_1^2$$

$$\frac{(\hat{\beta}_j - \beta_j)}{\hat{\sigma}_j} \to \mathcal{N}(0,1)$$

## Confidence interval

$$IC_{\mathbb{1}-\alpha}(\beta_j) = \left[\hat{\beta}_j - u_{1-\alpha/2}\hat{\sigma}_j ; \hat{\beta}_j - u_{1-\alpha/2}\hat{\sigma}_j\right]$$

# Test

## Significativity test of the coefficients

To test

$$H_0 : \beta_j = 0 \qquad \text{against} \qquad H_1 : \beta_j \neq 0$$

one uses

$$Z_j = \frac{\hat{\beta}_j}{\hat{\sigma}_j} \sim \mathcal{N}(0, 1)$$

## Significativity test of all the coefficients

To test

$$H_0 : \beta_j = 0 \quad \forall j \qquad \text{against} \quad H_1 : \text{it exists at least one } j \quad \beta_j \neq 0$$

one uses Wald test

$$(\hat{\beta} - 0)(\mathbf{X}^t \mathbf{W}_{\hat{\beta}} \mathbf{X}^{\mathbf{t}})(\hat{\beta} - \mathbf{0}) \sim \chi^2_{\mathbf{d}}$$

# Logistic Regression CDH on Age

```
> CHD.logit = glm(CHD~AGE, family=binomial(link="logit"))
> summary(CHD.logit)

Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.30945    1.13365  -4.683 2.82e-06 ***
AGE          0.11092    0.02406   4.610 4.02e-06 ***
---

Number of Fisher Scoring iterations: 4
```

Test: $\beta_1 = 0$: test if the proba of CDH is independent of Age:
$\hat{p}_i = \frac{e^{\hat{\beta}_0}}{1+e^{\hat{\beta}_0}}$

# Plan

# Prediction Error (CHD example)

## Prediction (threshold $= 0.5$)

Then, for a given $\mathbf{x}$,

- If $\widehat{\mathbb{P}}(Y = \mathrm{Yes}|\mathbf{X} = \mathbf{x}) > 0.5$, we predict $\widehat{y} = \mathrm{Yes}$;
- If $\widehat{\mathbb{P}}(Y = \mathrm{Yes}|\mathbf{X} = \mathbf{x}) \leq 0.5$, we predict $\widehat{y} = \mathrm{No}$.

**Confusion Matrix : Cross table of the prediction vs the truth**

```
##       pred
## CHD    No Yes
##   No   45  12
##   Yes  14  29
```

**Prediction error** : $(14 + 12)/100 = 0.26$

# Classical Metrics

| | | Predicted Class | |
|---|---|---|---|
| | | Yes | No |
| Actual Class | Yes | TP | FN |
| | No | FP | TN |

## Score

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

# Classical Metrics

## Scores

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

$$\text{True Positive Rate} = \text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{False Positive Rate} = 1 - \text{Specificity} = \frac{\text{FP}}{\text{FP} + \text{TN}}$$

$$\text{Precision} = \frac{\text{TP}}{\#(\text{predicted P})} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

- Many other metrics...

# ROC Curve

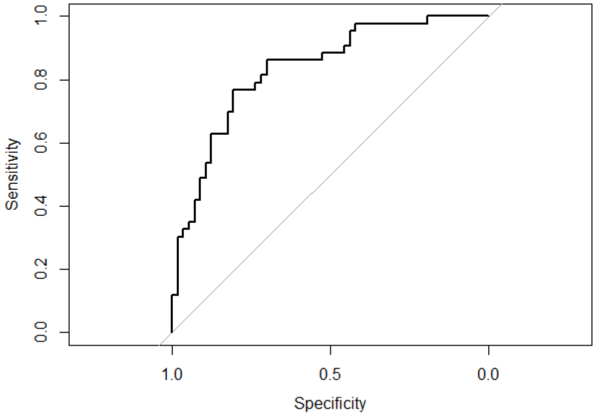
## **ROC Curve** (Receiver Operating Characteristic)

- For binary classification
- True positive Rate TPR $= \frac{\text{TP}}{\text{TP}+\text{FN}}$
- False positive Rate FPR $= \frac{\text{FP}}{\text{FP}+\text{TN}}$
- $x$-axis: FPR, $y$-axis: TPR
- Each point $A_t$ of the curve has coordinates $(\text{FPR}_t, \text{TPR}_t)$, where $\text{FPR}_t$ and $\text{TPR}_t$ are FPR and TPR of the confusion matrix obtained by the classification rule

$$\hat{y}_i = \mathbf{1}\hat{p}_i \geq t$$

- AUC score is the Area Under the ROC Curve

# ROC Curve




**ROC Curve**

Area under the curve: 0.8331

# How to draw benefits from the estimated model?

**Prediction for a new individual:**

- A new individual $\mathbf{x}_{\text{new}}$ appears and we want to predict if he has the disease ($y_{\text{new}} = 1$) or not ($y_{\text{new}} = 0$).
- We have estimated the logistic coefficients $\widehat{\beta}$ so that, for all $\mathbf{x}$,

$$\mathbb{P}\left(Y = 1 | \mathbf{X} = \mathbf{x}\right) \approx \frac{e^{\widehat{\beta}^T \mathbf{x}}}{1 + e^{\widehat{\beta}^T \mathbf{x}}}. \tag{1}$$

- Any ideas to predict $y_{\text{new}}$?

# Plan

# Generalized linear model

Reference: Generalized Linear Models. P. McCullagh, John A. Nelder, Chapman and Hall/CRC.

## GLM

- Distribution: $E[Y_i] = \mu_i$
- Linear predictor $\eta_i = \sum_j x_{ij}\beta_j$
- Link function $g(\mu) = \eta$

## GLM

- Gaussian: linear regression
- Bernouilli: logistic regression
- Poisson: poisson regression (counts)

# Much more...

## Logistic regression

Interpret the coefficients for categorical predictors
Interpret interaction
Inspect residuals - Model checking
With more than two categories: multinomial regression multinom function nnet package
clm function of ordinal package
polr mass package
Model selection - Likelihood Ratio
In high dimension, regularization
R package: glmnet

# Model Selection

1. Given $Y$ a variable to explain by $d$ variables $X^{(1)}, \dots, X^{(d)}$ , how to select (systematically) the most interesting subset of variables to do the prediction?

**Variable selection**

Find automatically a sub-group of variables to explain $Y$.

2. More generaly, given $k$ models $\mathcal{M}_1, \dots, \mathcal{M}_k$, which one to use?

**Model selection**

Criterion to compare the performance of different models.

# Embedded model testing I

- Assume we have two competing models $\mathcal{M}_S$ (with $S$ parameters) and $\mathcal{M}_L$ (with $L$ parameters) such that $\mathcal{M}_S \subset \mathcal{M}_L$ .
- Can we test if $\mathcal{M}_S$ is sufficient?

## Example (S=2 and L=4)

- Models:
  - $\mathcal{M}_S$ : $\mathrm{logit} p_\beta^{(S)}(\mathbf{x}) = \beta_0 + \beta_1 X^{(1)} + \beta_2 X^{(2)}$
  - $\mathcal{M}_L$ : $\mathrm{logit} p_\beta^{(L)}(\mathbf{x}) = \beta_0 + \beta_1 X^{(1)} + \beta_2 X^{(2)} + \beta_3 X^{(3)} + \beta_4 X^{(4)}$
- Test

$$H_0 : \beta_3 = \beta_4 = 0 \quad \text{against} \quad H_1 : \beta_3 \neq 0 \quad \mathrm{or} \quad \beta_4 \neq 0$$

# Embedded model testing II

## Deviance

- Log-likelihood $\mathcal{M}_S$: $\mathcal{L}_S = \log(L_S(\hat{\beta}, \mathcal{D}_n))$
- Log-likelihood of model $\mathcal{M}_L$: $\mathcal{L}_L = \log(L_L(\hat{\beta}, \mathcal{D}_n))$
- Likelihood Ratio: $L_S/L_L$. $Log(R) = log(S) - log(L)$. $-2Log(R) = 2log(L) - 2log(S)$ Deviance between the two models:

$$\mathcal{D}_{L-S} = 2(\mathcal{L}_L - \mathcal{L}_S) = 2(\log(L_L(\hat{\beta}, \mathcal{D}_n)) - \log(L_S(\hat{\beta}, \mathcal{D}_n)))$$

Asymptotically under $H_0$ : $(\mathcal{D}_{L-S}) \sim \chi^2(L-S)$

Under R : If `W` and `V` are two `glm` objects such that `W` is a submodel of `V`, the command anova(W,V,test="Chisq") performs this test.

# Embedded model testing III

## LR to test the significance of all the coefficients

Deviance between the complete model and the null model with the intercept:

$$\mathcal{D}_{d-0} = 2(\mathcal{L}_d - \mathcal{L}_0) = 2(\log(L_d(\hat{\beta}, \mathcal{D}_n)) - \log(L_0(\hat{\beta}, \mathcal{D}_n))) \sim \chi^2(d)$$

# AIC, BIC

- Let $\mathcal{M}$ be a generic logistic model and denote $p$ its number of parameters.
- Let $\hat{\beta}$ be the ML estimate in this model $\mathcal{M}$.

- The AIC and BIC consist in minimizing

$$-2 \times \log(L(\hat{\beta}, \mathcal{D}_n)) + \kappa(n) \times p$$

  over all models.
- Different choices for the factor $\kappa(n)$:
  - AIC : $\kappa(n) = 2$.
  - BIC : $\kappa(n) = \log n$.
- The BIC criterion leads to the selection of a model with a smaller dimension than AIC.

# Logistic model

## Other choices for $g$

- Classical choice for $g$ such that $g : \mathbb{R} \to [0, 1]$,

$$
\begin{aligned}
g^{-1}(t) &= \frac{e^t}{1 + e^t} && \text{logit} \\
g^{-1}(t) &= F_{\mathcal{N}}(t) && \text{probit} \\
g^{-1}(t) &= 1 - e^{-e^t} && \text{log-log}
\end{aligned}
$$

where $F_{\mathcal{N}}$ is the cumulative distribution function of a standard Gaussian $\mathcal{N}(0, 1)$.

# In high dimension

## Penalized Likelihood

- Minimization of

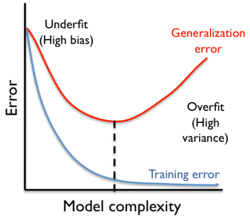
$$\operatorname*{argmin}_{\beta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \log(1 + e^{-y_i(\beta^t \mathbf{x}_i)}) + \mathrm{pen}(\beta)$$

  where $\mathrm{pen}(\beta)$ is a (sparsity promoting) penalty
- Variable selection if $\beta$ is sparse.

## Classical Penalties

- AIC: $\mathrm{pen}(\beta) = \lambda \|\beta\|_0$ (non convex / sparsity)
- Ridge: $\mathrm{pen}(\beta) = \lambda \|\beta\|_2^2$ (convex / no sparsity)
- Lasso: $\mathrm{pen}(\beta) = \lambda \|\beta\|_1$ (convex / sparsity)
- Elastic net: $\mathrm{pen}(\beta) = \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2$ (convex / sparsity)

# Regularization Parameter Issue



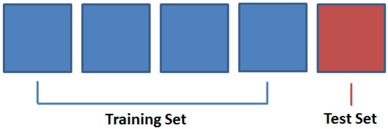

- Need to choose $\lambda$ from the data!

## Error behaviour

- Learning/training error (error made on the learning/training set) decays when the regularization parameter decreases.
- Quite different behavior when the error is computed on new observations (generalization error).

- Overfit for complex models: parameters learned are too specific to the learning set!
- General situation! (Think of polynomial fit...)
- Need another criterion than the training error!

# Cross Validation




- **Very simple idea:** use a second learning/verification set to compute a verification error.
- Sufficient to avoid over-fitting!

## Cross Validation

- Use $\frac{V-1}{V}n$ observations to train and $\frac{1}{V}n$ to verify!
- Validation for a learning set of size $(1-\frac{1}{V}) \times n$ instead of $n$!

- Most classical variations:
  - Leave One Out,
  - $V$-fold cross validation.
- Accuracy/Speed tradeoff: $V = 5$ or $V = 10$!

# Penalization and Cross-Validation

## Practical Selection Methodology

- Choose a penalty shape $\widetilde{\text{pen}}(\beta)$.
- Compute a CV error for a penalty $\lambda\widetilde{\text{pen}}(\beta)$ for all $\lambda \in \Lambda$.
- Determine $\widehat{\lambda}$ the $\lambda$ minimizing the CV error.
- Compute the final logistic regression with a penalty $\widehat{\lambda}\widetilde{\text{pen}}(\beta)$.