



# Logistic regression

## Basic Statistics for health researchers

Alessandra Meddis

Section of Biostatistics, University of Copenhagen



March 12, 2025

## Regression

Regression can be used when we want to adjust for confounding, or to account for risk factors to gain power in the analysis.

The type of outcome determines which kind of model is relevant:

### Continuous outcome

- ▶ Linear regression.
  - ▶ To model **means**.
  - ▶ Association parameters : **differences** between **mean** values

### Binary outcome

- ▶ Logistic regression.
  - ▶ To model **probabilities**.
  - ▶ Association parameters: **odds ratio** (OR) or equivalently **differences** between **log(odds)**.

## Outline

### Overview

ILO: to outline what the (univariate) logistic model is about

### One binary covariate

ILO: to interpret the model fit when using only one binary covariate

### One categorical (non binary) covariate

ILO: to interpret the model fit when using only categorical binary covariate  
ILO: to use the model to perform a powerful multiple testing adjustment

### One continuous covariate

ILO: to interpret and check the model, when using only one continuous covariate

### Multiple regression: two binary covariates

ILO: to interpret the fit of a multiple regression (i.e. an adjusted model)

### Multiple regression: one continuous and one binary covariate

ILO: to interpret the fit of a multiple regression (i.e. an adjusted model)

### Multiple regression: interaction

ILO: to interpret interactions and explain their meaning to others

2 / 68



## Binary outcome regression: why not linear?

If the outcome variable is binary:

$$Y_i = \begin{cases} 1 & \text{if } i \text{ is diseased} \\ 0 & \text{if } i \text{ is not diseased} \end{cases}$$

then **linear regression**

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

is **not good** for **many** reasons.



## Binary outcome regression: why not linear?

If the outcome variable is binary:

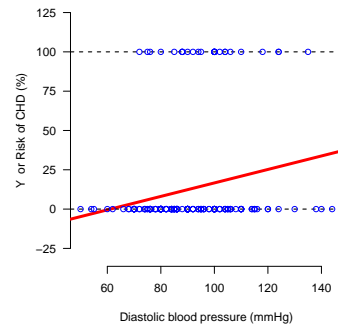
$$Y_i = \begin{cases} 1 & \text{if } i \text{ is diseased} \\ 0 & \text{if } i \text{ is not diseased} \end{cases}$$

then **linear regression**

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

is **not good** for **many** reasons.

One reason is that the regression line can go below 0 and above 1.



4/68



5/68

## Logistic regression

$$Y_i = \begin{cases} 1 & \text{if } i \text{ is diseased} \\ 0 & \text{if } i \text{ is not diseased} \end{cases}$$

We model the probability of the event  $Y_i = 1$  with predictor variable  $X_i$

$$P(Y_i = 1|X_i) = p_i$$

Instead of using a linear regression for  $p_i$ , which is bounded between 0 and 1, we apply **the logit function**:

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = a + bx_i$$

► It's a good idea as  $\log\left(\frac{p_i}{1-p_i}\right)$  can be both **negative** and **positive**.



$$\log\left(\frac{p_i}{1-p_i}\right) = a + bx_i$$

$$\rightarrow \left(\frac{p_i}{1-p_i}\right) = \exp(a + bx_i) = \exp(a) \cdot \exp(bx_i)$$

If we consider a categorical  $X$ , with two levels:

$$X = \begin{cases} 1 & \text{if in group 1} \\ 0 & \text{if in group 2} \end{cases}$$

The odds ratio are defined as:

$$OR = \frac{\left(\frac{p_1}{1-p_1}\right)}{\left(\frac{p_2}{1-p_2}\right)} = \frac{\exp(a + b \cdot 1)}{\exp(a + b \cdot 0)} = \frac{\exp(a) \cdot \exp(b)}{\exp(a) \cdot 1} = \exp(b)$$

thus,  $\exp(b)$  is interpreted as **odds ratio**, or equivalently the **difference between log(odds)**

$$b = \log(OR) = \log\left(\frac{\frac{p_1}{1-p_1}}{\frac{p_2}{1-p_2}}\right) = \log\left(\frac{p_1}{1-p_1}\right) - \log\left(\frac{p_2}{1-p_2}\right)$$



6/68

## Case: Framingham study<sup>1</sup>

Data,  $n=1,363$ :

	AGE	FRW	SBP	DBP	CHOL	CIG	sex	disease
1	45	93	100	62	220	0	Female	0
2	48	93	108	70	340	0	Male	0
3	45	91	160	100	171	0	Female	0
4	50	110	110	70	224	0	Male	0
5	48	85	110	70	229	25	Male	0
6	55	101	134	84	224	0	Male	0



**Outcome:** coronary heart disease (CHD) during follow-up (1=yes/no=0).

<sup>1</sup>Mahmood et al. "The Framingham Heart Study and the epidemiology of cardiovascular disease: a historical perspective." The Lancet 383.9921 (2014): 999-1008.

Andersson, Charlotte, et al. "70-year legacy of the Framingham Heart Study." Nature Reviews Cardiology 16.11 (2019): 687-698.



## Categorical explanatory variable ( $K$ groups, $k = 1, \dots, K$ )

- ▶ **sex:** Male/Female
- ▶ **AGE:** age (years) at baseline (45-62)
- ▶ FRW: "Framingham relative weight" (pct.) at baseline (52-222; 11 persons have missing values)
- ▶ SBP: systolic blood pressure at baseline (mmHg) (90-300)
- ▶ **DBP:** diastolic blood pressure at baseline (mmHg) 50-160)
- ▶ CHOL: cholesterol at baseline (mg/100ml) (96-430)
- ▶ **CIG:** cigarettes per day at baseline (0-60; 1 person has missing value)
- ▶ **disease:** 1 if coronary heart disease (CHD) during follow-up, 0 otherwise

### Linear regression, continuous outcome $Y$

$$\text{mean}(Y|\text{group } k) - \text{mean}(Y|\text{reference group})$$

*E.g., the average blood pressure was higher in males compared to females.*

### Logistic regression, binary outcome

$$\text{OR} = \frac{\text{odds}(\text{group } k)}{\text{odds}(\text{reference group})}$$

*E.g., the risk (or the odds<sup>2</sup>) of coronary heart disease was higher in males compared to females.*

8 / 68



9 / 68

<sup>2</sup>remember:  $\text{odds}(p) = p/(1-p)$  and "higher odds" is equivalent to "higher risk".



## Quantitative (continuous) predictor variables

### Linear regression, continuous outcome $Y$

Differences in mean values per unit of  $X$ :

$$\text{mean}(Y|x+1) - \text{mean}(Y|x)$$

*E.g., the average systolic blood pressure increased with one unit of age.*

## Quantitative (continuous) predictor variables

### Linear regression, continuous outcome $Y$

Differences in mean values per unit of  $X$ :

$$\text{mean}(Y|x+1) - \text{mean}(Y|x)$$

*E.g., the average systolic blood pressure increased with one unit of age.*

### Logistic regression, binary outcome

Ratio of odds per unit of  $X$

$$\text{Odds ratio} = \frac{\text{odds}(x+1)}{\text{odds}(x)}$$

Differences in  $\log(\text{odds})$  per unit of  $X$

$$\log(\text{OR}) = \log\{\text{odds}(x+1)\} - \log\{\text{odds}(x)\}$$

*E.g., the risk (odds) of coronary heart disease increased with age.*

10 / 68



## Linearity in regression models

For a continuous variable  $X$  (e.g. age), linearity means that the effect of a unit change of  $X$  on the outcome does not depend on the value of  $X$ .

### ► Linear regression, continuous outcome $Y$

$$\begin{aligned}\text{mean}(Y|45+1) - \text{mean}(Y|45) &= \text{mean}(Y|46+1) - \text{mean}(Y|46) \\ &= \dots = \text{mean}(Y|61+1) - \text{mean}(Y|61)\end{aligned}$$

### ► Logistic regression, binary outcome

$$\frac{\text{odds}(45+1)}{\text{odds}(45)} = \frac{\text{odds}(46+1)}{\text{odds}(46)} = \dots = \frac{\text{odds}(61+1)}{\text{odds}(61)}$$

Linearity is a model assumption which should be checked!<sup>3</sup>

<sup>3</sup>Categorizing a continuous covariate can be useful when linearity does not hold.

## Software parametrization

By default, **software** report  $\log(\text{Odds ratio}) = \text{difference in } \log(\text{odds})$ .

$$\begin{aligned}\log(\text{OR}) &= \log \left\{ \frac{\text{odds}(\text{group } k)}{\text{odds}(\text{reference group})} \right\} \\ &= \log \left\{ \text{odds}(\text{group } k) \right\} - \log \left\{ \text{odds}(\text{reference group}) \right\}\end{aligned}$$

But it does not matter for the **interpretation**.

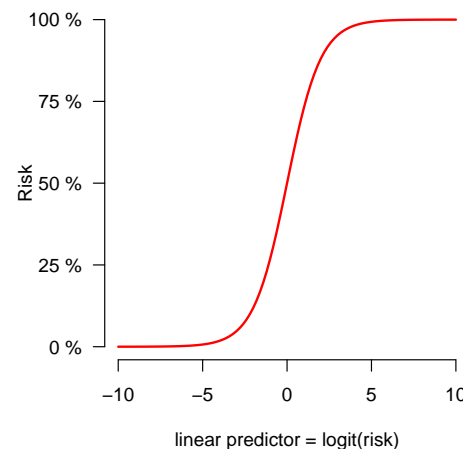
- $OR > 1 \Leftrightarrow \log(OR) > 0 \Leftrightarrow RR > 1$  (higher risk)
- $OR = 1 \Leftrightarrow \log(OR) = 0 \Leftrightarrow RR = 1$  (same risk)
- $OR < 1 \Leftrightarrow \log(OR) < 0 \Leftrightarrow RR < 1$  (lower risk)

## Appendix: further details

Equivalently, the (univariate) logistic model is:

$$p_i = \frac{\exp(a + bx_i)}{1 + \exp(a + bx_i)}$$

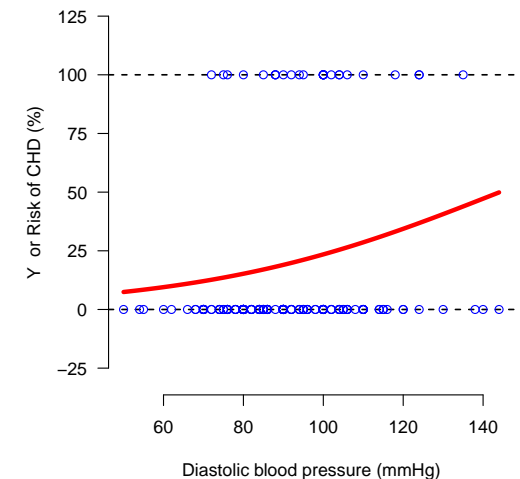
### ► $a + bx_i$ : linear predictor



Example of model fit, with  $x$  being the diastolic blood pressure (mmHg):

$$p_i = \frac{\exp(-3.86 + 0.027x_i)}{1 + \exp(-3.86 + 0.027x_i)}$$

Here the **linear predictor** ranges from  
 $-3.86 + 0.027 \cdot 50 = -3.52$  to  
 $-3.86 + 0.027 \cdot 144 = -0.13$   
 because the pressure ranges from 50 to 144.



# Outline

## Overview

ILO: to outline what the (univariate) logistic model is about

## One binary covariate

ILO: to interpret the model fit when using only one binary covariate

## One categorical (non binary) covariate

ILO: to interpret the model fit when using only categorical binary covariate

ILO: to use the model to perform a powerful multiple testing adjustment

## One continuous covariate

ILO: to interpret and check the model, when using only one continuous covariate

## Multiple regression: two binary covariates

ILO: to interpret the fit of a multiple regression (i.e. an adjusted model)

## Multiple regression: one continuous and one binary covariate

ILO: to interpret the fit of a multiple regression (i.e. an adjusted model)

## Multiple regression: interaction

ILO: to interpret interactions and explain their meaning to others

15 / 68

## Research question:<sup>4</sup>

Do men and women have the same risk of coronary heart disease?

<sup>4</sup>A bit made up, just for pedagogical purpose, to illustrate the concepts.

## A binary explanatory variable

$$Y_i = \begin{cases} 1 & \text{subject } i \text{ develops coronary heart diseased (CHD)} \\ 0 & \text{subject } i \text{ does not develop CHD} \end{cases}$$

$$X_i = \begin{cases} 1 & \text{subject } i \text{ is a female} \\ 0 & \text{if subject } i \text{ a male (REF group)} \end{cases}$$

Univariate logistic regression for  $p_i = P(Y_i = 1 | X_i = x_i)$ :

$$\log\left(\frac{p_i}{1-p_i}\right) = a + bx_i = \begin{cases} a & \text{males} \\ a+b & \text{females} \end{cases}$$

17 / 68

## A binary explanatory variable

$$Y_i = \begin{cases} 1 & \text{subject } i \text{ develops coronary heart diseased (CHD)} \\ 0 & \text{subject } i \text{ does not develop CHD} \end{cases}$$

$$X_i = \begin{cases} 1 & \text{subject } i \text{ is a female} \\ 0 & \text{if subject } i \text{ a male (REF group)} \end{cases}$$

Univariate logistic regression for  $p_i = P(Y_i = 1 | X_i = x_i)$ :

$$\log\left(\frac{p_i}{1-p_i}\right) = a + bx_i = \begin{cases} a & \text{males} \\ a+b & \text{females} \end{cases}$$

That means,

$$\begin{aligned} \log(OR_{\varnothing \text{ vs } \sigma}) &= \log\left(\frac{\text{odds for } \varnothing}{\text{odds for } \sigma}\right) = \\ &= \log(\text{odds for } \varnothing) - \log(\text{odds for } \sigma) = (a+b) - a = b \end{aligned}$$

$$\text{and } \log(OR_{\sigma \text{ vs } \varnothing}) = -(\log(\text{odds for } \varnothing) - \log(\text{odds for } \sigma)) = -b$$

17 / 68

## Logistic regression in R

```
fit1 <- glm(disease~sex, data=framingham, family=binomial)
```

- ▶ `disease ~ sex`: tells R that `disease` is the outcome and `sex` the predictor variable.
- ▶ `data=framingham`: tells R *where* to find the variables `disease` and `sex`.
- ▶ `glm`: means “generalized linear model”.
- ▶ `family=binomial`: tells R that the outcome is *binary* and the *logit* link function should be used.

19 / 68



## R code: only sex variable

R code:

```
fit1 <- glm(disease~sex, data=framingham, family=binomial)
summary(fit1)
```

Output (partial):

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-1.07183	0.09047	-11.847	< 2e-16 ***
sexFemale	-0.70702	0.13937	-5.073	3.92e-07 ***

Note: Pay attention to the default reference group ! Here it is “male” as in the previous slide...

Note: Remember that the results are for  $\log(\text{OR})$

19 / 68



## Comparison with results from the 2x2 table

```
TabSex <- table(relevel(framingham$sex,ref="Female"),
               factor(framingham$disease,levels=c(1,0)))
table2x2(TabSex,stat=c("table","or"))
```

2x2 contingency table

	1	0	Sum
Female	104	616	720
Male	164	479	643
--	--	--	--
Sum	268	1095	1363

Odds ratio =  $\text{OR} = (p1/(1-p1))/(p2/(1-p2)) = 0.4931$   
 Standard error =  $\text{SE.OR} = \text{sqrt}((1/a+1/b+1/c+1/d)) = 0.1394$

And we can see the same results:

- ▶  $\widehat{\text{OR}} = \exp(-0.7070219) = 0.493$
- ▶ Standard error of  $\log(\text{OR}) = 0.1394$ .

For this simple case with only one binary predictor variable, logistic regression is equivalent to what we have seen last week.

20 / 68



## Confidence intervals for the odds ratio

```
library(Publish)
publish(fit1)
```

Variable	Units	OddsRatio	CI.95	p-value
Sex	Male	1.00	[1.00;1.00]	1
	Female	0.49	[0.38;0.65]	<0.0001

Note:  $0.49 = \exp(-0.71)$ .

21 / 68



## Confidence intervals for the odds ratio

```
library(Publish)
publish(fit1)
```

Variable	Units	OddsRatio	CI.95	p-value
Sex	Male	1.00	[1.00;1.00]	1
	Female	0.49	[0.38;0.65]	<0.0001

**Note:**  $0.49 = \exp(-0.71)$ .

**“Typical”/possible conclusion sentence:**

*Women have a significantly lower risk to develop coronary heart disease than men (odds ratio: 0.49, 95%-CI: [0.38; 0.65], p-value <0.0001).*

21 / 68



22 / 68

## Changing the reference level

```
framingham$sexF <- relevel(framingham$sex,ref="Female")
fit1a <- glm(disease~sexF, data=framingham, family=binomial)
publish(fit1a)
```

## Changing the reference level

```
framingham$sexF <- relevel(framingham$sex,ref="Female")
fit1a <- glm(disease~sexF, data=framingham, family=binomial)
publish(fit1a)
```

Variable	Units	OddsRatio	CI.95	p-value
sexF	Female	1.00	[1.00;1.00]	1
	Male	2.03	[1.54;2.66]	<0.0001

**Note:**  $2.03 = \exp(0.71)$ .

**“Typical”/possible conclusion sentence:**

*Men have a significantly higher risk to develop coronary heart disease than women (odds ratio: 2.03, 95%-CI: [1.5; 2.7], p-value <0.0001).*

22 / 68



23 / 68

## Outline

### Overview

ILO: to outline what the (univariate) logistic model is about

### One binary covariate

ILO: to interpret the model fit when using only one binary covariate

### One categorical (non binary) covariate

ILO: to interpret the model fit when using only categorical binary covariate

ILO: to use the model to perform a powerful multiple testing adjustment

### One continuous covariate

ILO: to interpret and check the model, when using only one continuous covariate

### Multiple regression: two binary covariates

ILO: to interpret the fit of a multiple regression (i.e. an adjusted model)

### Multiple regression: one continuous and one binary covariate

ILO: to interpret the fit of a multiple regression (i.e. an adjusted model)

### Multiple regression: interaction

ILO: to interpret interactions and explain their meaning to others



## Model with only one categorical explanatory variable

Assume that we want to **compare several groups**, e.g. four **age** groups.<sup>6</sup>

Research questions:<sup>5</sup>

Is age associated with the risk of coronary heart disease?

Are some age groups more at risk of coronary heart disease than others?

		Age				
		45-48	49-52	53-56	57-62	
Outcome	$Y = 1$	51	61	64	92	268
(CHD)	$Y = 0$	308	298	254	235	1095
		359	359	318	327	1363

We can either use:

- Fisher's exact test or Pearson  $\chi^2$  for the global null hypothesis  
 **$H_0$ : "the risk is the same for all age groups"** (see Lecture 7).
- or **logistic regression** to make all-pairwise comparisons (via OR) and use the **min-P approach to efficiently account for multiple testing**.<sup>7</sup>

<sup>5</sup>A bit made up, just for pedagogical purpose, to illustrate the concepts.

<sup>6</sup>Note: we pooled the data of men and women.

<sup>7</sup>It also works when we "adjust" for other variables.

## Logistic regression: categorical variable with 4 levels:

$$\log\left(\frac{p_i}{1-p_i}\right) = \begin{cases} a & \text{age } 45-48 \\ a+b_1 & \text{age } 49-52 \\ a+b_2 & \text{age } 53-56 \\ a+b_3 & \text{age } 57-62 \end{cases}$$

Reference category 45-48

$$a = \log(\text{odds}(45-48))$$

$$b_1 = \log\left(\frac{\text{odds}(49-52)}{\text{odds}(45-48)}\right)$$

$$b_2 = \log\left(\frac{\text{odds}(53-56)}{\text{odds}(45-48)}\right)$$

$$b_3 = \log\left(\frac{\text{odds}(57-62)}{\text{odds}(45-48)}\right)$$

- Equivalent to making 3 times the 2x2 table analysis for the group 45-48 versus each of the three others.

## Results: one categorical predictor variable

```
framingham$AgeCut <- cut(framingham$AGE,
                        c(40,48,52,56,99),
                        labels=c("45-48","49-52","53-56","57-62"))
fit3 <- glm(disease~AgeCut, data=framingham, family=binomial)
publish(fit3)
```

Variable	Units	OddsRatio	CI.95	p-value
AgeCut	45-48	Ref		
	49-52	1.24	[0.82;1.85]	0.30425
	53-56	1.52	[1.02;2.28]	0.04151
	57-62	2.36	[1.61;3.46]	< 0.0001

Remarks:

- Not all (six) comparisons are directly available from the "summary" of the model fit, for example the odds ratio for group 57-62 vs 53-56 is not.
- $\widehat{OR} = (92 \times 308) / (51 \times 235) = 2.36$  and all estimates match those of each corresponding 2 x 2 table.
- Running a similar code after **changing the reference group** is a convenient **"trick"** to obtain any OR estimate, with corresponding 95% CI and p-value.



## Equivalent Results

```
framingham$AgeCutb <- relevel(framingham$AgeCut, "53-56")
fit3b <- glm(disease~AgeCutb, data=framingham, family=
  binomial)
publish(fit3b)
```

Variable	Units	OddsRatio	CI.95	p-value
AgeCutb	53-56	Ref		
	45-48	0.66	[0.44;0.98]	0.04151
	49-52	0.81	[0.55;1.20]	0.29468
	57-62	1.55	[1.08;2.24]	0.01798

### As expected:

- ▶  $0.66=1/1.52$ , i.e.  $OR(45-48 \text{ vs } 53-56)=1/OR(53-56 \text{ vs } 45-48)$
- ▶  $1.55=2.36/1.52$ , i.e.  $OR(57-62 \text{ vs } 53-56)=OR(57-62 \text{ vs } 45-48)/OR(53-56 \text{ vs } 45-48)$

28 / 68

## Appendix: all pairwise comparisons

### Statistical methods:

Comparisons between groups were made using a logistic model. P-values and 95% confidence intervals were adjusted for multiple testing using the min-P (aka max-t test) method as implemented in the multcomp-package [ref.<sup>8</sup>] of the statistical software R [ref.<sup>9</sup>] and described in [ref.<sup>10</sup>].

### Results (adjusted for multiple testing):

Comparison	Est. OR	95% CI	p-value
49-52 - 45-48	1.24	[0.7;2.1]	0.7329
53-56 - 45-48	1.52	[0.9;2.6]	0.1736
57-62 - 45-48	2.36	[1.4;3.9]	0.0001
53-56 - 49-52	1.23	[0.7;2.0]	0.7207
57-62 - 49-52	1.91	[1.2;3.1]	0.0028
57-62 - 53-56	1.55	[1.0;2.5]	0.0836

### Note:

- ▶ Significant association between CHD and age groups, p-value= 0.0001 (i.e. the **minimum**)

<sup>8</sup> Hothorn, Bretz & Westfall (2008). Simultaneous Inference in General Parametric Models. Biometrical Journal 50(3), 346–363.

<sup>9</sup> R Core Team (2023). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.

<sup>10</sup> Bretz, Hothorn, & Westfall (2016). Multiple comparisons using R. CRC Press.

## Outline

### Overview

ILO: to outline what the (univariate) logistic model is about

### One binary covariate

ILO: to interpret the model fit when using only one binary covariate

### One categorical (non binary) covariate

ILO: to interpret the model fit when using only categorical binary covariate

ILO: to use the model to perform a powerful multiple testing adjustment

### One continuous covariate

ILO: to interpret and check the model, when using only one continuous covariate

### Multiple regression: two binary covariates

ILO: to interpret the fit of a multiple regression (i.e. an adjusted model)

### Multiple regression: one continuous and one binary covariate

ILO: to interpret the fit of a multiple regression (i.e. an adjusted model)

### Multiple regression: interaction

ILO: to interpret interactions and explain their meaning to others

30 / 68

### Research questions:<sup>11</sup>

Is age associated with the risk of coronary heart disease?

How does age relate to the risk of coronary heart disease?

31 / 68

<sup>11</sup> A bit made up, just for pedagogical purpose, to illustrate the concepts.

## Quantitative explanatory factor

It is **sometimes more natural or better** to include the a continuous variable (e.g. age) as a quantitative predictor in the model (i.e., *No grouping*)<sup>12</sup>

$$\log\left(\frac{p_i}{1-p_i}\right) = a + b \cdot \text{age}_i$$

$$a = \log(\text{odds}(\text{age}=0))$$

$$b = \log\left\{\text{odds}(\text{age}=\textcolor{red}{x} + 1)\right\} - \log\left\{\text{odds}(\text{age}=\textcolor{red}{x})\right\}$$

**Interpretation:** for each year, the factor by which odds for CHD increases with each **one unit** increase of age (here 1 year) is

$$\exp(b) = \text{odds ratio}$$

<sup>12</sup>sometimes better but not always, due to the linearity assumption or similar.



33 / 68

## Appendix: details on inference (Est., 95% CI & p-values)

- ▶ We **estimate the parameters (a and b)** by '**maximum likelihood estimation**': giving them values that makes the observations of the outcome of our data the "most likely" to be observed (again).
- ▶ We compute the **standard error** for each parameter by looking at how much the likelihood to observe the outcome of our data is sensitive to the parameter values. Intuition: high sensitivity= a small range of parameter values makes the data "most likely"= small standard error.
- ▶ **95 % confidence interval** for parameters:

$$\text{estimate} \pm 1.96 \cdot \text{standard error}.$$

- ▶ **p-value** for the null hypothesis  $H_0$ : "parameter=0":

$$z = \frac{\text{estimate}}{\text{standard error}} \quad \text{and} \quad \text{p-value} = P(|Z| > |z|),$$

with  $Z$  being a random variable with a standard normal distribution. It works well, but software can also do something slightly more precise (called "profile likelihood" inference).



## Raw results

```
fit5 <- glm(disease~AGE,data=framingham,family=binomial)
summary(fit5)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-4.88431	0.77372	-6.313	0.000000000274 ***
AGE	0.06581	0.01446	4.550	0.000005374208 ***

▶  $\widehat{OR} = \exp(0.06581) = 1.07$



## Good reporting practice

### 1-year change in age (not very good)

```
fit5 <- glm(disease~AGE,data=framingham,family=binomial)
publish(fit5)
```

Variable	Units	OddsRatio	CI.95	p-value
AGE		1.07	[1.04;1.10]	< 0.0001

### 10-year change in age (probably better)

```
framingham$age10 <- framingham$AGE/10
fit5b <- glm(disease~age10,data=framingham,family=binomial)
publish(fit5b)
```

Variable	Units	OddsRatio	CI.95	p-value
age10		1.93	[1.45;2.56]	< 0.0001



## Good reporting practice

### 1-year change in age (not very good)

```
fit5 <- glm(disease~AGE,data=framingham,family=binomial)
publish(fit5)
```

Variable	Units	OddsRatio	CI.95	p-value
AGE		1.07	[1.04;1.10]	< 0.0001

### 10-year change in age (probably better)

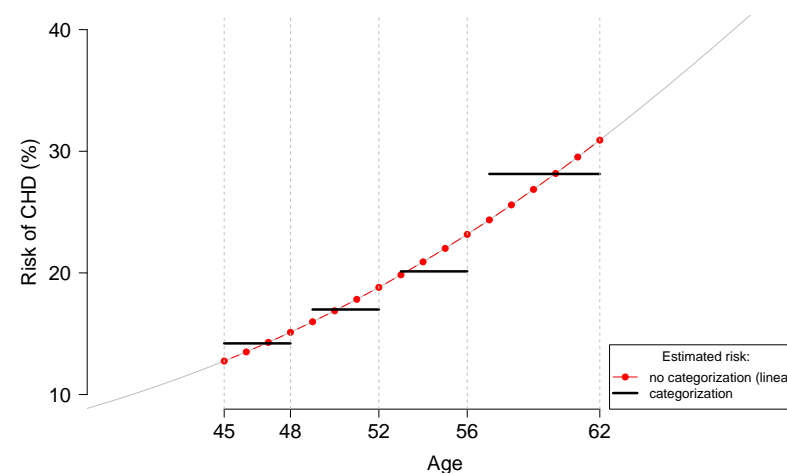
```
framingham$age10 <- framingham$AGE/10
fit5b <- glm(disease~age10,data=framingham,family=binomial)
publish(fit5b)
```

Variable	Units	OddsRatio	CI.95	p-value
age10		1.93	[1.45;2.56]	< 0.0001

These results are completely equivalent:  $1.93 = 1.07^{10}$ . The fitted models are the same, but the "default" way of presenting the results is different.

35 / 68

## Visualizing and checking the linearity assumption



- We compare the "flexible" model which uses the **categorized variable** to the "less flexible" model (but "nicer" if correct!) which uses the **continuous variable** (together with a "linearity" assumption).

36 / 68

## Outline

### Overview

ILO: to outline what the (univariate) logistic model is about

### One binary covariate

ILO: to interpret the model fit when using only one binary covariate

### One categorical (non binary) covariate

ILO: to interpret the model fit when using only categorical binary covariate

ILO: to use the model to perform a powerful multiple testing adjustment

### One continuous covariate

ILO: to interpret and check the model, when using only one continuous covariate

### Multiple regression: two binary covariates

ILO: to interpret the fit of a multiple regression (i.e. an adjusted model)

### Multiple regression: one continuous and one binary covariate

ILO: to interpret the fit of a multiple regression (i.e. an adjusted model)

### Multiple regression: interaction

ILO: to interpret interactions and explain their meaning to others

37 / 68

## Multiple logistic regression

Additive effects of several explanatory variables:

$$\log\left(\frac{p_i}{1-p_i}\right) = a + b_1 z_i + b_2 x_i + \dots$$

with  $p_i = P(Y_i = 1 | X_i = x_i, Z_i = z_i, \dots)$ .

- Multiple logistic regression is a way to control for **confounding / unbalanced design**.
- Makes it possible to estimate odds ratios to **compare the risks of two groups of subjects who are similar with respect to all predictor variables except one**.

38 / 68

## Multiple logistic regression

Additive effects of several explanatory variables:

$$\log\left(\frac{p_i}{1-p_i}\right) = a + b_1 z_i + b_2 x_i + \dots$$

with  $p_i = P(Y_i = 1 | X_i = x_i, Z_i = z_i, \dots)$ .

- ▶ Multiple logistic regression is a way to control for **confounding / unbalanced design**.
- ▶ Makes it possible to estimate odds ratios to **compare the risks of two groups of subjects who are similar with respect to all predictor variables except one**.
- ▶ We often say that the effect (via the odds ratio) on the outcome of each predictor variable under study (e.g. “exposure”), is **adjusted** for the other explanatory variables (e.g. age, sex, comorbidity).
- ▶ **Without interaction**, the **model assumes** that **the effect (odds ratio) of  $z$  on  $Y$  is the same for all values of  $x$** .

39 / 68



Research question:<sup>13</sup>

Are smokers more at risk of coronary heart disease than non-smokers?

Background (that we need to take into account):

It is known that men smoke more than women.

Hence the aim of the statistical analysis:

We want to compare the risk of two subjects, one smokes, the other doesn't, who are **similar with respect to sex** (i.e. either both men or both women).

<sup>13</sup>A bit made up, just for pedagogical purpose, to illustrate the concepts.



## Example of two binary variables

$$Z_i = \begin{cases} 1 & \text{if male} \\ 0 & \text{female} \end{cases} \quad \text{and} \quad V_i = \begin{cases} 1 & \text{if smokes} \\ 0 & \text{otherwise} \end{cases}$$

Data can be summarized as two 2 by 2 tables **in two ways**, but usually, one option is more interesting than the other for the research question.

	Males ( $Z=1$ )			Females ( $Z=0$ )	
	$Y=1$	$Y=0$		$Y=1$	$Y=0$
Smoker: $V=1$	107	288	$V=1$	27	192
Non Smoker: $V=0$	57	191	$V=0$	77	423

Here it is less interesting to look at the two 2 by 2 tables showing the association between  $Y$  (disease) and  $Z$  (Sex) given  $V$  (Smoking) because it is less related to our research question.



40 / 68

## Model with two binary variables (without interaction)

$$\begin{aligned} \log\left(\frac{p_i}{1-p_i}\right) &= a + b_1 Z_i + b_2 V_i \\ &= \begin{cases} a & \text{Female non-smoker} \\ a + b_1 & \text{Male non-smoker} \\ a + b_2 & \text{Female smoker} \\ a + b_1 + b_2 & \text{Male smoker} \end{cases} \end{aligned}$$

$$\text{Note: } b_1 = (a + b_1) - a \quad (\text{comparison among non-smoker})$$

$$= (a + b_1 + b_2) - (a + b_2) \quad (\text{comparison among smoker})$$

$$= \log OR(\text{♂ vs ♀ for given smoking status})$$

$$\text{and } b_2 = (a + b_2) - a \quad (\text{comparison among female})$$

$$= (a + b_1 + b_2) - (a + b_1) \quad (\text{comparison among male})$$

$$= \log OR(\text{smokers vs. non-smokers for given sex})$$



41 / 68

## Logistic regression results

```
fit2 <-glm(disease~sex+Smoke,data=framingham,family=binomial)
summary(fit2)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-1.09215	0.12717	-8.588	< 2e-16 ***
sexFemale	-0.69521	0.14635	-4.750	2.03e-06 ***
SmokeYes	0.03296	0.14457	0.228	0.82

42 / 68



## Extracting odds ratios with confidence intervals

```
publish(fit2)
```

Variable	Units	OddsRatio	CI.95	p-value
sex	Male	Ref		
	Female	0.50	[0.37;0.66]	<1e-04
Smoke	No	Ref		
	Yes	1.03	[0.78;1.37]	0.8196

43 / 68



## Extracting odds ratios with confidence intervals

```
publish(fit2)
```

Variable	Units	OddsRatio	CI.95	p-value
sex	Male	Ref		
	Female	0.50	[0.37;0.66]	<1e-04
Smoke	No	Ref		
	Yes	1.03	[0.78;1.37]	0.8196

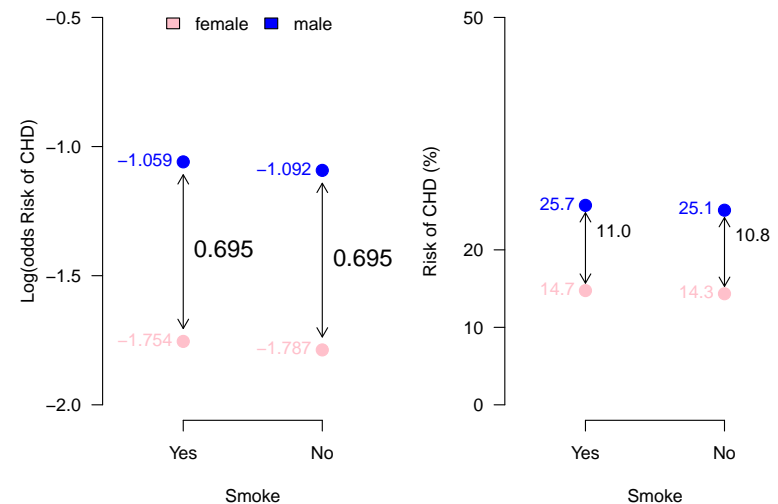
"Typical"/possible conclusion sentence:

Logistic regression adjusted for sex did not show an increase in odds of CHD in smokers compared to non-smokers (OR=1.03, 95% CI: [0.78;1.37],  $p=0.82$ ).

43 / 68



## Visual interpretation



**Note:** additivity is on the logit scale (i.e.  $\log(\text{odds})$ ), not on the risk scale

44 / 68



# Outline

## Overview

ILO: to outline what the (univariate) logistic model is about

## One binary covariate

ILO: to interpret the model fit when using only one binary covariate

## One categorical (non binary) covariate

ILO: to interpret the model fit when using only categorical binary covariate

ILO: to use the model to perform a powerful multiple testing adjustment

## One continuous covariate

ILO: to interpret and check the model, when using only one continuous covariate

## Multiple regression: two binary covariates

ILO: to interpret the fit of a multiple regression (i.e. an adjusted model)

## Multiple regression: one continuous and one binary covariate

ILO: to interpret the fit of a multiple regression (i.e. an adjusted model)

## Multiple regression: interaction

ILO: to interpret interactions and explain their meaning to others

45 / 68



## Research question:<sup>14</sup>

Do men and women have the same risk of coronary heart disease?

## Background:

It is known that aging increases the risks of coronary heart disease. We could not collect the data in a way that necessarily makes the distribution of age similar for men and women.

## Hence the aim of statistical analysis:

We want to compare the risk of two subjects, one is a man, the other a woman, both are similar with respect to age.

<sup>14</sup>A bit made up, just for pedagogical purpose, to illustrate the concepts.

46 / 68



## Another multiple regression example

Additive model (no statistical interactions)

$$\log \left( \underbrace{\frac{p_i}{1-p_i}}_{=\text{odds}_i} \right) = a + b_1 z_i + b_2 x_i$$

Effect of **sex**  $z_i$  (0 = female, 1 = male) adjusted for **age** ( $x_i$ )

$$\begin{aligned} \frac{\text{odds}(\text{age}=50, \text{male})}{\text{odds}(\text{age}=50, \text{female})} &= \frac{\exp(a + b_1 + b_2 50)}{\exp(a + b_2 50)} \\ &= \exp(a + b_1 + b_2 50 - a - b_2 50) \\ &= \exp(b_1). \end{aligned}$$

The result is the same for age 46 and age 61 and all other ages.



47 / 68

Effect of age ( $x_i$ ) for **males**:

$$\begin{aligned} \frac{\text{odds}(\text{age}=51, \text{male})}{\text{odds}(\text{age}=50, \text{male})} &= \frac{\exp(a + b_1 + b_2 51)}{\exp(a + b_1 + b_2 50)} \\ &= \exp(a + b_1 + b_2 51 - a - b_1 - b_2 50) \\ &= \exp(b_2). \end{aligned}$$

The result is the same for **females**:

$$\begin{aligned} \frac{\text{odds}(\text{age}=51, \text{female})}{\text{odds}(\text{age}=50, \text{female})} &= \frac{\exp(a + b_2 51)}{\exp(a + b_2 50)} \\ &= \exp(a + b_2 51 - a - b_2 50) \\ &= \exp(b_2). \end{aligned}$$

Linearity means that the result is **the same for** a comparison of age 63 and age 62 and **all other one year differences**.



48 / 68

## Results (raw)

```
fit6 <- glm(disease ~ AGE + sex, family = binomial,
            data = framingham)
summary(fit6)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )	
(Intercept)	-4.59208	0.78019	-5.886	3.96e-09 ***	
AGE	0.06672	0.01458	4.575	4.75e-06 ***	
sexFemale	-0.71613	0.14052	-5.096	3.46e-07 ***	

49 / 68



## Results (formatted for publication)

```
fit6 <- glm(disease ~ AGE + sex, family = binomial, data =
            framingham)
publish(fit6)
```

Variable	Units	OddsRatio	CI.95	p-value
AGE		1.07	[1.04;1.10]	<1e-04
sex	Male	Ref		
	Female	0.49	[0.37;0.64]	<1e-04

### Possible conclusion sentences:

*Logistic regression was used to investigate gender differences in odds (risks) of CHD adjusted for age.*

*The age adjusted odds ratio was 0.49 (95%-CI: [0.37;0.64]) showing that the risks of CHD were significantly lower for women compared to men ( $p < 0.0001$ ).*

50 / 68



## Predicted risks based on logistic regression model

A logistic regression model can be used to predict  
**“personalized”/conditional risks**, since

$$\log\left(\frac{p_i}{1-p_i}\right) = a + b_1 z_i + b_2 z_i + \dots$$

is equivalent to

$$p_i = \frac{\exp(a + b_1 z_i + b_2 x_i + \dots)}{1 + \exp(a + b_1 z_i + b_2 x_i + \dots)}$$

We can predict a risk for any value of the covariates  $Z, X, \dots$  once we have estimated the model parameters. We just need to plug the estimated parameter values into the equations.<sup>15</sup>

**Note:** the risks (and risk ratios) depend on all predictor variables simultaneously.

<sup>15</sup>However, upmost caution is needed when using covariate values beyond the range of those observed (e.g. age=110). Usually we do not want to extrapolate beyond observed data.

51 / 68



## Visualization of predicted risks

► For men:

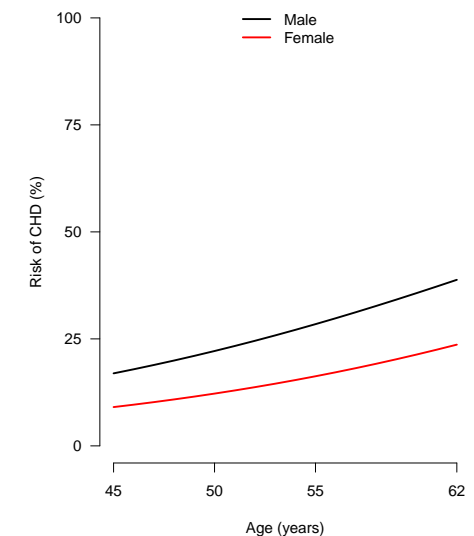
$$\frac{\exp(-4.59208 + 0.06672 \cdot \text{age})}{1 + \exp(-4.59208 + 0.06672 \cdot \text{age})}$$

► For women:

$$\frac{\exp(-4.59208 - 0.71613 + 0.06672 \cdot \text{age})}{1 + \exp(-4.59208 - 0.71613 + 0.06672 \cdot \text{age})}$$

Because we have seen:

	Estimate
(Intercept)	-4.59208
AGE	0.06672
SexFemale	-0.71613



**Note:**  $\widehat{OR}(\text{male vs female given age}) = e^{-0.71613} = 0.489$  but  $\widehat{RR}$  varies from 0.535 to 0.610.

52 / 68



# Outline

## Overview

ILO: to outline what the (univariate) logistic model is about

## One binary covariate

ILO: to interpret the model fit when using only one binary covariate

## One categorical (non binary) covariate

ILO: to interpret the model fit when using only categorical binary covariate

ILO: to use the model to perform a powerful multiple testing adjustment

## One continuous covariate

ILO: to interpret and check the model, when using only one continuous covariate

## Multiple regression: two binary covariates

ILO: to interpret the fit of a multiple regression (i.e. an adjusted model)

## Multiple regression: one continuous and one binary covariate

ILO: to interpret the fit of a multiple regression (i.e. an adjusted model)

## Multiple regression: interaction

ILO: to interpret interactions and explain their meaning to others

53 / 68



54 / 68



Statistical interaction = Effect modification

The effect of the exposure  $X$  on the outcome  $Y$  depends on some covariate  $Z$

Example: the effect of age ( $X$ ) on coronary heart disease ( $Y$ ) depends on the sex ( $Z$ ).

# Effect modification

## Setting: 3 variables.

- ▶ two predictor variables  $X$  and  $Z$
- ▶ one outcome  $Y$

## Meaning

In **logistic regression**, an interaction means that the **odds ratio** which describes the effect of  $X$  on the odds of  $Y = 1$  depends on the value of  $Z$ .

## Symmetry

If the effect of variable  $X$  on  $Y$  is modified by  $Z$  then also the effect of  $Z$  on  $Y$  is modified by  $X$ .

## Research question:<sup>16</sup>

What are the risk of coronary heart disease for men and women at any age?

How different is the consequence of aging on the risk of coronary heart disease between men and women?

55 / 68



56 / 68

<sup>16</sup>A bit made up, just for pedagogical purpose, to illustrate the concepts.





## Interaction between a continuous and a binary variable

To model the **interaction** we add “ $b_3x_i \cdot z_i$ ” in the model, i.e.,

$$\log\left(\underbrace{\frac{p_i}{1-p_i}}_{=\text{odds}_i}\right) = a + b_1z_i + b_2x_i + b_3x_i \cdot z_i$$

► The effect of **sex**  $z_i$  (0 = female, 1 = male) depends on **age** ( $x_i$ ).

$$\frac{\text{odds}(\text{age}=50, \text{male})}{\text{odds}(\text{age}=50, \text{female})} = \frac{\exp(a + b_1 + b_2\text{50} + b_3\text{50})}{\exp(a + b_2\text{50})} = \exp(b_1 + b_3\text{50}).$$

When  $\begin{cases} b_3 > 0 \\ b_3 < 0 \end{cases}$ , then OR(♂ vs ♀ given age)  $\begin{cases} \text{increases} \\ \text{decreases} \end{cases}$  with age.

57 / 68

► The effect of **age** ( $x_i$ ) depends on **sex**  $z_i$ .

$$\frac{\text{odds}(\text{age}=50, \text{male})}{\text{odds}(\text{age}=45, \text{male})} = \frac{\exp(a + b_1 + b_2\text{50} + b_3\text{50})}{\exp(a + b_1 + b_2\text{45} + b_3\text{45})} = \exp(b_2\text{5} + b_3\text{5}).$$

$$\frac{\text{odds}(\text{age}=50, \text{female})}{\text{odds}(\text{age}=45, \text{female})} = \exp(b_2\text{5}).$$

**Note:**  $\exp(b_2)$  describes the odds ratio for age in the **reference group** for sex (female) only, while it is  $\exp(b_2 + b_3)$  in the other group (male).

58 / 68

## Statistical interaction in R

First option (more transparent):

```
glm(disease ~ AGE + sex + AGE:sex, family = binomial,
    data = framingham)
```

Shorter syntax (less transparent):

```
glm(disease ~ AGE * sex, family = binomial,
    data = framingham)
```

59 / 68

## Raw R output

```
fit7 <- glm(disease ~ AGE + sex + AGE:sex, family = binomial,
            data = framingham)
summary(fit7)
```

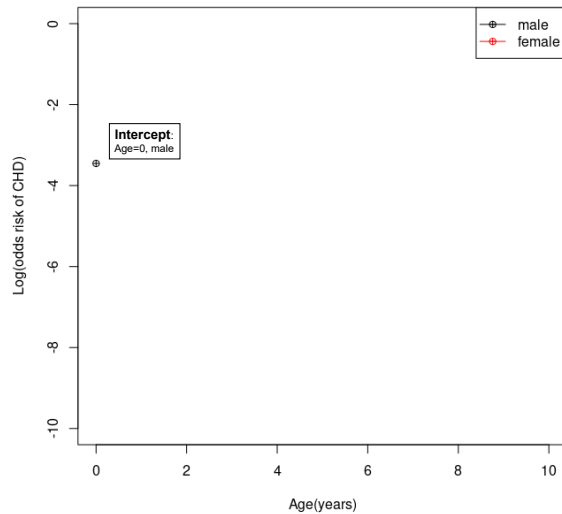
Coefficients:

	Estimate	Std. Error	z value	Pr(> z )	
(Intercept)	-3.45290	1.00008	-3.453	0.000555	***
AGE	0.04523	0.01883	2.402	0.016288	*
sexFemale	-3.54459	1.60431	-2.209	0.027146	*
AGE:sexFemale	0.05297	0.02987	1.773	0.076194	.

60 / 68

## Interpretation

$$\log\left(\frac{p_i}{1-p_i}\right) = a + b_1 z_i + b_2 x_i + b_3 x_i \cdot z_i$$



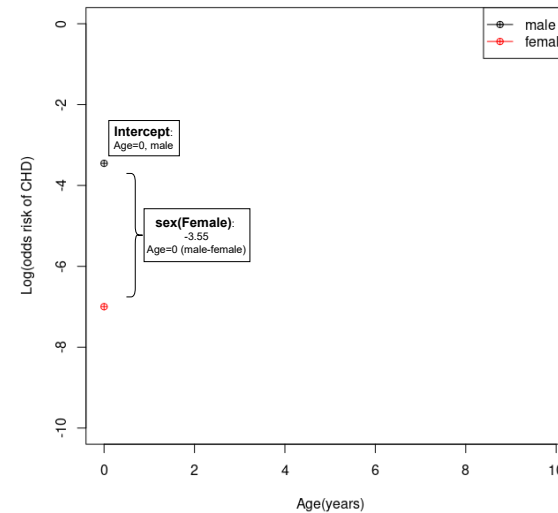
61 / 68

- Intercept  $a$ : est. log of the odds in the reference group (male, AGE=0).  
**Not meaningful here!**



## Interpretation

$$\log\left(\frac{p_i}{1-p_i}\right) = a + b_1 z_i + b_2 x_i + b_3 x_i \cdot z_i$$



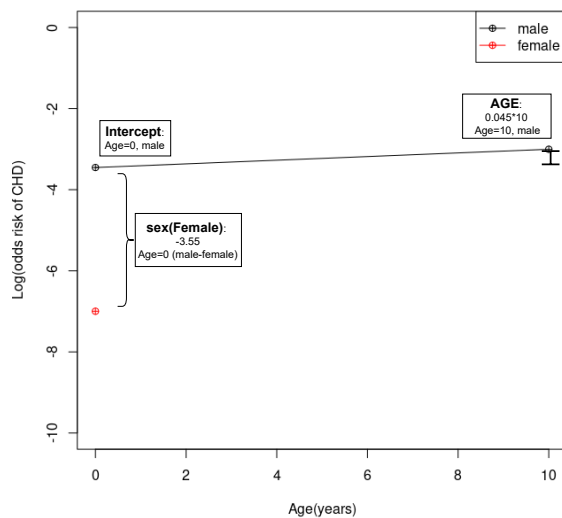
61 / 68

- sexFemale  $b_2$ : est. log of the OR, when comparing the risk of a female to that of a male at AGE=0.  
**Not meaningful here!**



## Interpretation

$$\log\left(\frac{p_i}{1-p_i}\right) = a + b_1 z_i + b_2 x_i + b_3 x_i \cdot z_i$$



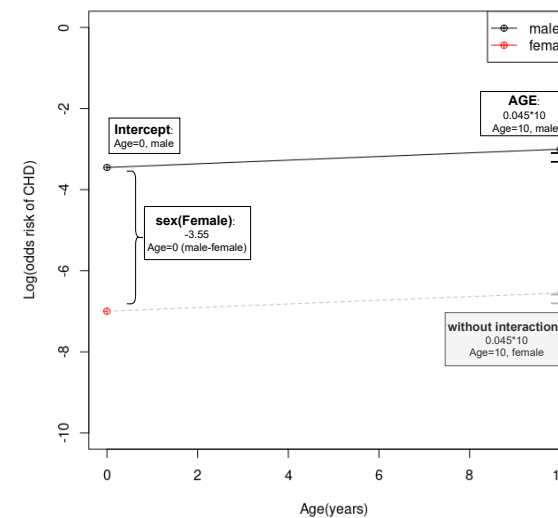
61 / 68

- Age  $b_1$ : est. log of the OR, when comparing the risks of two **males** (the reference group for sex), one being 1 year older than the other.  
Because of linearity, at 10 years =  $0.045 \cdot 10$



## Interpretation

$$\log\left(\frac{p_i}{1-p_i}\right) = a + b_1 z_i + b_2 x_i + b_3 x_i \cdot z_i$$



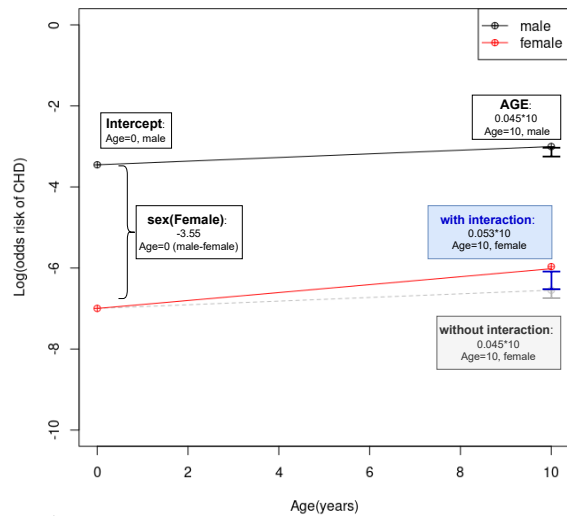
61 / 68

- Without interaction:**  $b_1$  would also be the estimated log of the OR when comparing the risk of two **females** one being 1 year older than the other



## Interpretation

$$\log\left(\frac{p_i}{1-p_i}\right) = a + b_1 z_i + b_2 x_i + b_3 x_i \cdot z_i$$



- ▶ **With Interaction**  
(AGE:sexFemale)  $b_3$ :est.  
log of the ratio of two ORs: the OR to compare the risks of two **males** with 1 year age difference, and the OR to compare the risks of two **females** with one 1 year difference.  
For 10 years, the value is  $10 * 0.05297$ .

## Formatted results

```
fit7 <- glm(disease ~ AGE + sex + AGE:sex, family = binomial,
            data = framingham)
publish(fit7)
```

Variable	Units	OddsRatio	CI.95	p-value
AGE: sex(Male)		1.05	[1.01;1.09]	0.01629
AGE: sex(Female)		1.10	[1.05;1.15]	< 1e-04

### Interpretation

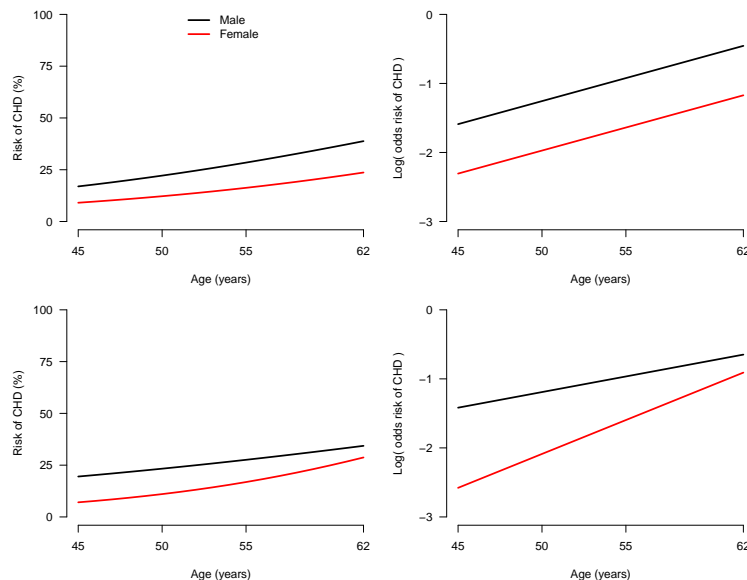
- ▶ One year more in age increases the odds by 5% (95% CI=[1;9]) in males and by 10% (95% CI=[5;15]) in females.<sup>17</sup>.
- ▶ However, note that the difference in the increase in odds between men and women is not significant (p-value=0.076).

61 / 68

62 / 68

$1.05 = \exp(0.04523)$  and  $1.10 = \exp(0.04523 + 0.05297)$

## Predicted risk with or without interaction



**Note:** without an interaction (top), the curves cannot cross.  
With interaction (bottom), they can.

63 / 68

## When using models with interaction?

- ▶ When it makes sense in the **context** of your study<sup>18</sup>.
  - ▶ Because of the research question.
  - ▶ To better “adjust”.
  - ▶ When subgroup analyses could be interesting.
- ▶ To check that the corresponding model without interaction seems “reasonable”, i.e. to challenge your modeling assumptions.

<sup>18</sup>But you should have enough data... the more flexible the model the more data you need to estimate it accurately.

64 / 68

## Two binary variables revisited: with interaction

```
fit8 <-glm(disease~sex*Smoke,data=framingham,family=binomial)
summary(fit8)
```

Coefficients:

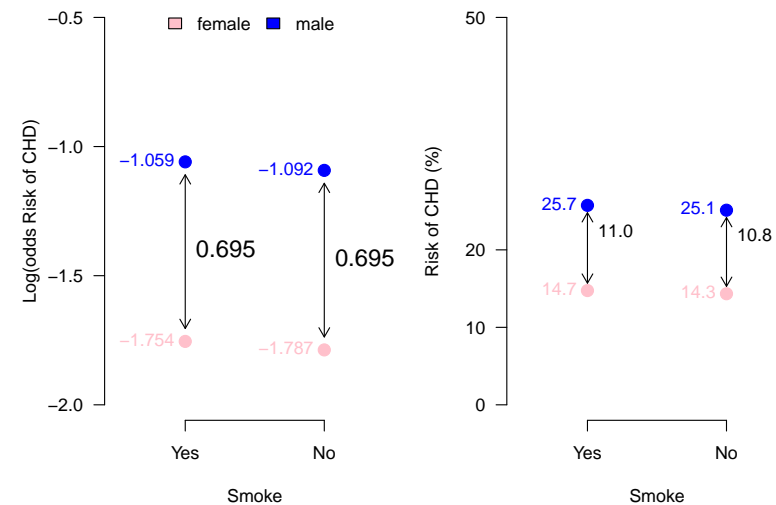
	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-1.2092	0.1509	-8.012	1.13e-15 ***
sexFemale	-0.4943	0.1953	-2.532	0.0114 *
SmokeYes	0.2191	0.1887	1.161	0.2456
sexFemale:SmokeYes	-0.4772	0.3053	-1.563	0.1180

```
publish(fit8)
```

Variable	Units	OddsRatio	CI.95	p-value
sex(Male): Smoke(Yes vs No)		1.24	[0.86;1.80]	0.24555
sex(Female): Smoke(Yes vs No)		0.77	[0.48;1.24]	0.28219
Smoke(No): sex(Female vs Male)		0.61	[0.42;0.89]	0.01135
Smoke(Yes): sex(Female vs Male)		0.38	[0.24;0.60]	< 1e-04

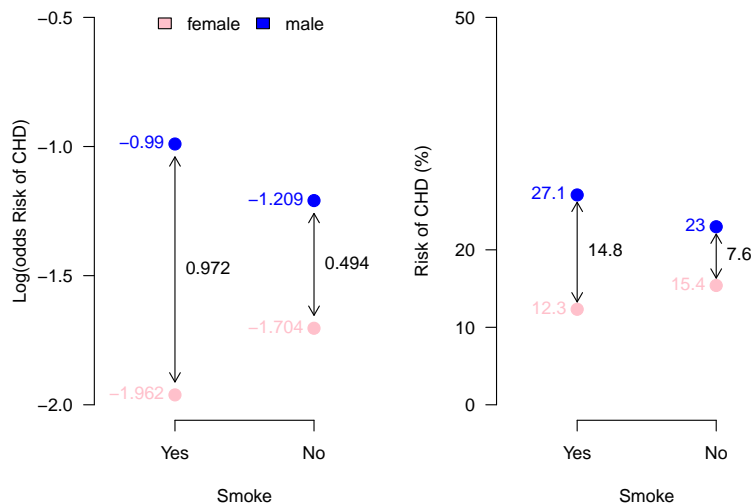
65 / 68

## Reminder: results without interaction



66 / 68

## Reminder: results with interaction



- Estimates are simply those obtained by stratifying, i.e. they match those of the two 2x2 tables of slide 40, e.g.  $27.1\% = 107/(107+288)$ .

67 / 68

## Take home messages

- (Multiple) logistic regression describes associations between one or several explanatory variables and the risk of an event (binary outcome), via odds ratio.
- The analysis of an exposure of interest can be adjusted for potential confounders.
- In an additive model (no interactions), the odds ratio for each explanatory variable does not depend on the other explanatory variables.
- Risks and risk ratios predicted by the model depend on the other explanatory variables.
- Linearity and absence of interaction are assumptions which might need to be checked.
- Models with interactions are flexible and useful but need more concentration to be interpreted correctly and more data to be fitted.
- Many models can be fitted from the same data, but some are more relevant than others for a given research question (e.g. in terms of adjustments and interactions).

68 / 68