



## Outline

### Overview

One binary covariate

One categorical (non binary) covariate

One continuous covariate

Multiple regression: two binary covariates

Multiple regression: one continuous and one binary covariates

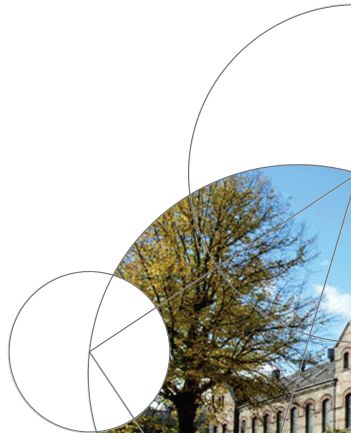
Multiple regression: interaction



## Day 6: logistic regression

Paul Blanche

Section of Biostatistics, University of Copenhagen



November 16, 2020

2 / 67

## Regression

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

### Quantitative (continuous) outcome

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

### 0-1 (binary) outcome

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

## Case: Framingham study

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).



- ▶ **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

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

### 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>1</sup>) of coronary heart disease was higher in males compared to females.*

5 / 67



6 / 67

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



## 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)



7 / 67

## 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 age.*

8 / 67



## 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 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(odds) per unit of  $X$

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

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



## 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>2</sup>

<sup>2</sup>Categorizing a continuous covariate can be useful when linearity does not hold.



## 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:

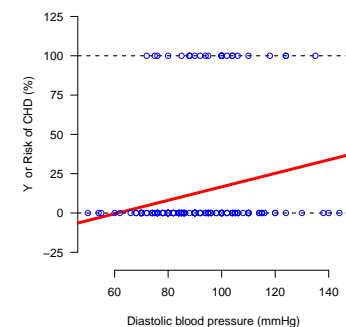
$$Y_i = \begin{cases} 1 & \text{if } i \text{ is diseased} \\ 0 & \text{if } i \text{ is not diseased} \end{cases}$$

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.



## (Univariate) logistic regression

We model the probability of the event  $Y_i = 1$  for a subject with predictor variable  $X_i$ .

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

The idea is to use the **logit** function  $p \mapsto \log\{p/(1-p)\}$ . Instead of using a linear regression for  $p_i$ , which is bounded between 0 and 1, we apply **linear regression to log(odds)**:

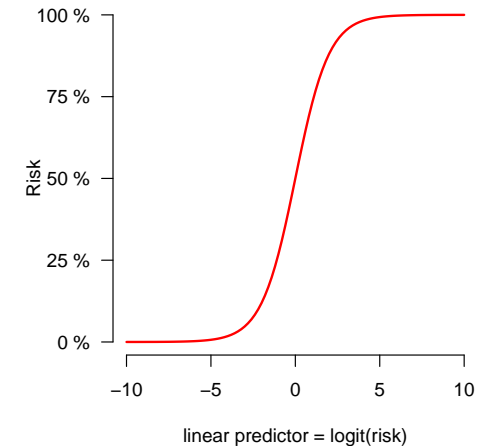
$$\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = a + bx_i$$

$\log\left(\frac{p_i}{1-p_i}\right)$  can take both negative and positive values. We will see that  $\exp(b)$  can be interpreted as an **odds ratio**.

Equivalently, the (univariate) logistic model is:

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

►  $a + bx_i$ : **linear predictor**



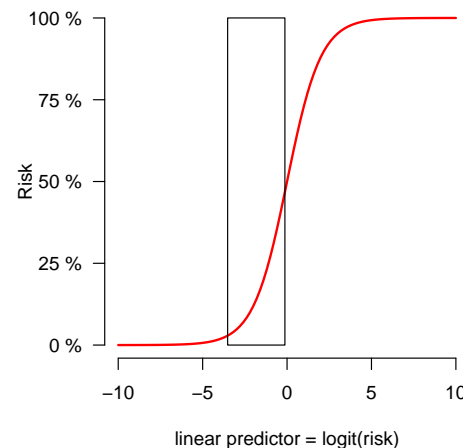
11 / 67

12 / 67

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

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

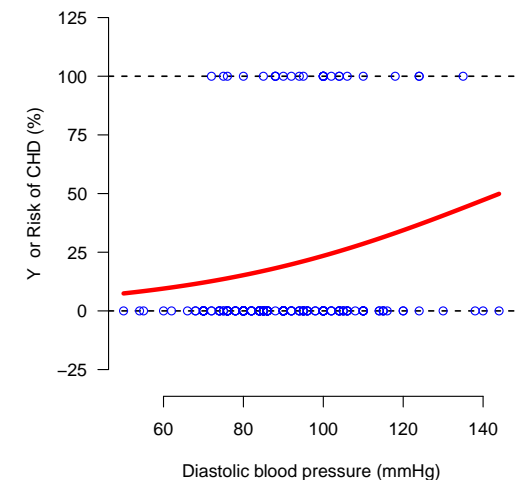
Here the **linear predictor** ranges from  $-3.86 + 0.027 \cdot 50 = -3.52$  to  $-3.86 + 0.027 \cdot 144 = -0.13$  because the pressure ranges from 50 to 144.



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.



13 / 67

13 / 67

# Outline

## Overview

### One binary covariate

### One categorical (non binary) covariate

### One continuous covariate

### Multiple regression: two binary covariates

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

### Multiple regression: interaction

### Research question:<sup>3</sup>

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

14 / 67



15 / 67

<sup>3</sup>A bit made up, just for pedagogical purpose, to illustrate the concepts.



## A binary explanatory variable

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

$$Z_i = \begin{cases} 1 & \text{subject } i \text{ is a man} \\ 0 & \text{if subject } i \text{ a woman} \end{cases}$$

Univariate (Simple) logistic regression for  $p_i = P(Y_i = 1|Z_i = z_i)$ :

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

16 / 67



## A binary explanatory variable

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

$$Z_i = \begin{cases} 1 & \text{subject } i \text{ is a man} \\ 0 & \text{if subject } i \text{ a woman} \end{cases}$$

Univariate (Simple) logistic regression for  $p_i = P(Y_i = 1|Z_i = z_i)$ :

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

That means,

$$\begin{aligned} b &= (a+b) - a = \log(\text{odds for males}) - \log(\text{odds for females}) \\ &= \log\left(\frac{\text{odds for males}}{\text{odds for females}}\right) \end{aligned}$$

and  $-b = a - (a+b) = \log(\text{odds for females}/\text{odds for males})$ .

16 / 67

Note: remember that  $\exp(-b) = 1/\exp(b)$ .



## 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 variable Y and Sex.
- ▶ `glm`: means “**g**eneralized **l**inear **m**odel”.
- ▶ `family=binomial`: tells R that the outcome is **binary** and the **logit** link function should be used.

17 / 67



## R code: only sex variable

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

Call:

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

Deviance Residuals:

	Min	1Q	Median	3Q	Max
	-0.7674	-0.7674	-0.5586	-0.5586	1.9672

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 ***

---

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

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1351.2 on 1362 degrees of freedom  
 Residual deviance: 1324.9 on 1361 degrees of freedom  
 AIC: 1328.9

Number of Fisher Scoring iterations: 4

18 / 67

**Note:** pay attention to the default reference group.



## 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 = OR =  $(p1/(1-p1))/(p2/(1-p2)) = 0.4931$

Standard error = SE.OR =  $\sqrt{(1/a+1/b+1/c+1/d)} = 0.1394$

And we can see the same results:

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

For this simple case with only one binary predictor variable, logistic regression is no different from what we have seen last week.

19 / 67



## 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 :  $0.49 = \exp(-0.71)$

20 / 67



## 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 :  $0.49 = \exp(-0.71)$

*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).*

20 / 67



21 / 67

## 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 :  $2.03 = \exp(0.71)$

*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).*

21 / 67



22 / 67

## Outline

Overview

One binary covariate

One categorical (non binary) covariate

One continuous covariate

Multiple regression: two binary covariates

Multiple regression: one continuous and one binary covariates

Multiple regression: interaction



## Model with only one categorical explanatory variable

Assume that we want to **compare several groups**, e.g. four **age** groups.<sup>5</sup>

Research questions:<sup>4</sup>

Is age associated to 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 5).
- or **logistic regression** to make all-pairwise comparisons (via OR) and use the "modern" **min-P approach to efficiently account for multiple testing**.<sup>6</sup>

<sup>23/67</sup> <sup>4</sup>A bit made up, just for pedagogical purpose, to illustrate the concepts.

<sup>24/67</sup> <sup>5</sup>Note: both males and females.

<sup>6</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.  $OR(57-62 \text{ vs } 53-56)=OR(57-62 \text{ vs } 45-48)/OR(53-56 \text{ vs } 45-48)$

27 / 67



## All pairwise comparisons: min-P approach

### 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 method as implemented in the multcomp-package [ref.<sup>7</sup>] of the statistical software R [ref.<sup>8</sup>] and described in [ref.<sup>9</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**)
- ▶ Similarly, we can use the method for the “**many-to-one**” setting (as in Lecture 4).

<sup>7</sup> Hothorn, Bretz & Westfall (2008). Simultaneous Inference in General Parametric Models. Biometrical Journal 50(3), 346–363.

<sup>8</sup> R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.

<sup>9</sup> Bretz, Hothorn, & Westfall (2016). Multiple comparisons using R. CRC Press.



## Outline

### Overview

### One binary covariate

### One categorical (non binary) covariate

### One continuous covariate

### Multiple regression: two binary covariates

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

### Multiple regression: interaction

### Research questions:<sup>10</sup>

Is age associated to the risk of coronary heart disease?

How age changes the risk of coronary heart disease?

29 / 67



30 / 67

<sup>10</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>11</sup>

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

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

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

**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>11</sup> sometimes better but not always, due to the linearity assumption.



32 / 67

## Inference with the logistic model (Estimates, 95% CI & p-values)

- ▶ We **estimate the parameters** by giving them values that makes the observations of the outcome of our data the “most likely” to be observed (again). This is called ‘**maximum likelihood estimation**’. **No simple formula**, except in very specific cases.
- ▶ We compute the **standard error** for each the parameter by looking at how much the likelihood to observe the outcome of our data is sensitive to the parameter values. Intuition: high sensitivity= small standard error. **No simple formula**, except in very specific cases.
- ▶ **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 (“profile likelihood”).



## 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)
```

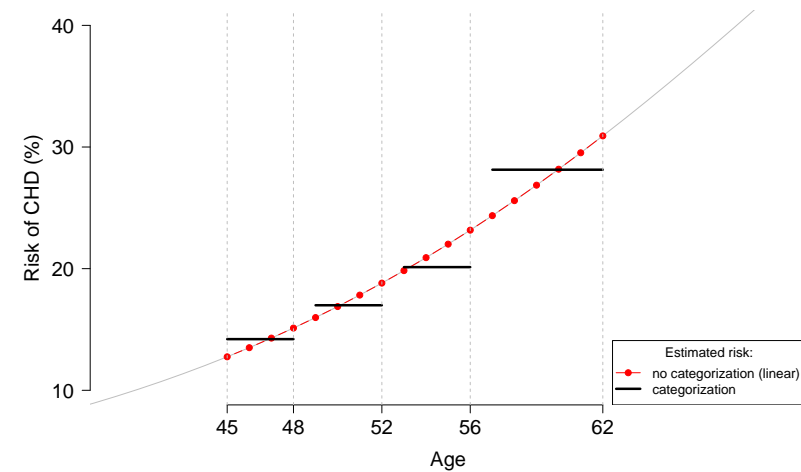
Variable	Units	OddsRatio	CI.95	p-value
age10		1.93	[1.45;2.56]	< 0.0001

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 / 67

## 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**.



## Outline

### Overview

### One binary covariate

### One categorical (non binary) covariate

### One continuous covariate

### Multiple regression: two binary covariates

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

### Multiple regression: interaction



36 / 67

## 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 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  $x$ .



37 / 67

## Example of two binary variables

Research question:<sup>12</sup>

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

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

Background:

It is known that men smoke more than women.

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.

Hence the aim of 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).

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

<sup>12</sup>A bit made up, just for pedagogical purpose, to illustrate the concepts.

## Model with two binary variables (without interaction)

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

$$\begin{aligned} \text{Note: } b_1 &= (a + b_1) - a && \text{(non-smoker)} \\ &= (a + b_1 + b_2) - (a + b_2) && \text{(smoker)} \\ &= \log OR \text{ (males vs. females for given smoking status)} \end{aligned}$$

$$\begin{aligned} \text{and } b_2 &= (a + b_2) - a && \text{(female)} \\ &= (a + b_1 + b_2) - (a + b_1) && \text{(male)} \\ &= \log OR \text{ (smokers vs. non-smokers for given)} \end{aligned}$$

## 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

## 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

42 / 67



## Extracting odds ratios with confidence intervals

```
publish(fit2)
```

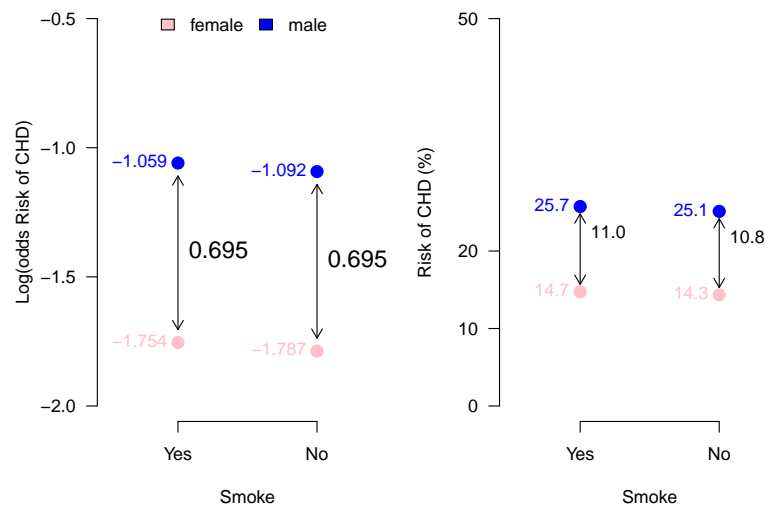
Variable	Units	OddsRatio	CI.95	p-value
sex	Male	Ref		
	Female	0.50	[0.37;0.66]	<1e-04
Smoke	No	Ref		
	Yes	1.03	[0.78;1.37]	0.8196

42 / 67



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],  $p=0.82$ ).

## Visual interpretation



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

43 / 67



## Outline

Overview

One binary covariate

One categorical (non binary) covariate

One continuous covariate

Multiple regression: two binary covariates

Multiple regression: one continuous and one binary covariates

Multiple regression: interaction

44 / 67



Research question:<sup>13</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.

## Another multiple regression example

Additive model (no statistical interactions)

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

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

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

<sup>45 / 67</sup> A bit made up, just for pedagogical purpose, to illustrate the concepts.

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

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**.

## 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	***

## 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

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 ( $p < 0.0001$ ).

49 / 67



## Predicted risks based on logistic regression model

A logistic regression model can be used to predict personalized risks, since

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

is equivalent to

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

The risks (and risk ratios) depend on all predictor variables simultaneously.

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>14</sup>

<sup>14</sup>However, utmost 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. Same remark in Lecture 3.



## Visualization of predicted risks

► For men:

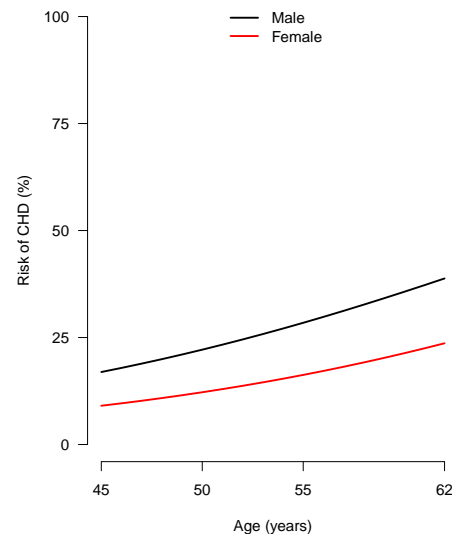
$$\frac{\exp(-4.59208 + 0.06672 \cdot \text{age})}{1 + \exp(-4.59208 + 0.06672 \cdot \text{age})}$$

► 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:**  $\widehat{OR}(\text{male vs female}) = \exp(-0.71613) = 0.489$  but  $\widehat{RR}$  varies from 0.535 to 0.610.

51 / 67



## Computation of predicted risks with R

```
dnew <- expand.grid(AGE=seq(from=45,to=62,by=1),
  sex=c("Female","Male"))
dnew$risk <- predict(fit6,newdata=dnew,type="response")
head(dnew)
```

	AGE	sex	risk
1	45	Female	0.09063347
2	46	Female	0.09628451
3	47	Female	0.10224827
4	48	Female	0.10853706
5	49	Female	0.11516303
6	50	Female	0.12213808

52 / 67



## Outline

Overview

One binary covariate

One categorical (non binary) covariate

One continuous covariate

Multiple regression: two binary covariates

Multiple regression: one continuous and one binary covariates

Multiple regression: interaction

## Statistical interaction = Effect modification

The effect of  $X$  on  $Y$  depends on  $Z$

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

53 / 67



54 / 67



## 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  $X$ .

**Research question:**<sup>15</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 / 67



56 / 67





## 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.,

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

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

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

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

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

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

**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).

57 / 67



58 / 67



## Statistical interaction in R

First option (more transparent):

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

Shorter syntax (less transparent):

```
glm(Y ~ AGE * SEX, family = binomial, data = framingham)
```

## 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	.

59 / 67



60 / 67



## 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

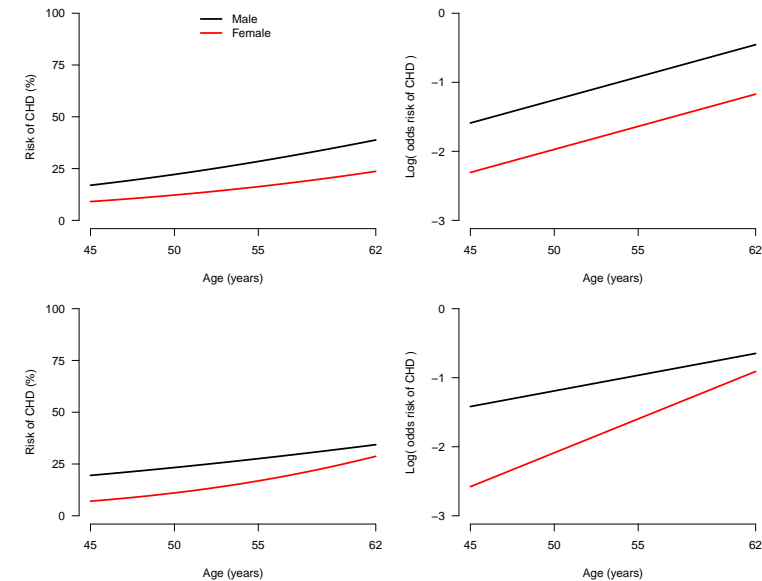
- ▶ One year more in age increases the odds by 5% (95% CI=[1;9]) in males and by 10% (95% CI=[5;10]) in females.
- ▶ However, note that the difference in the increase in odds between men and women is not significant (p-value=0.076).

61 / 67



62 / 67

## Predicted risk with or without interaction



## When using models with interaction?

- ▶ When it makes sense in the context of your study<sup>16</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>16</sup>But you should have enough data... the more flexible the model the more data you need to estimate it accurately.

63 / 67



## 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

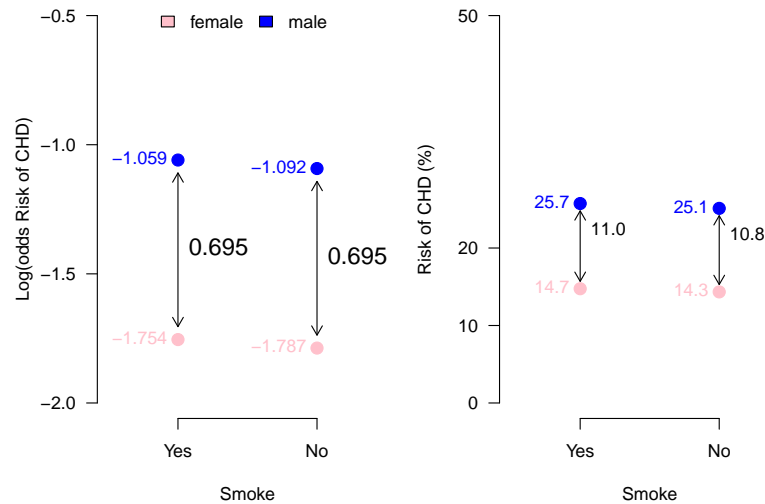
```
publish(fit8)
```

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



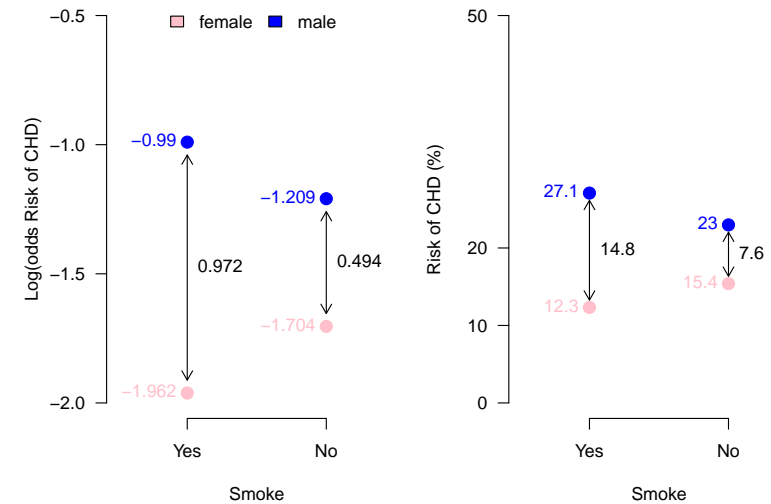
64 / 67

## Reminder: results without interaction



65 / 67

## Reminder: results with interaction



- Estimates are simply those obtained by stratifying, i.e. they match those of the two 2x2 tables of slide 39, e.g.  $27.1\% = 107/(107+288)$ .

66 / 67

## 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.
- Analysis of an exposure of interest can be adjusted for potential confounders.
- In an additive model (no interactions), odds ratios do 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).

67 / 67