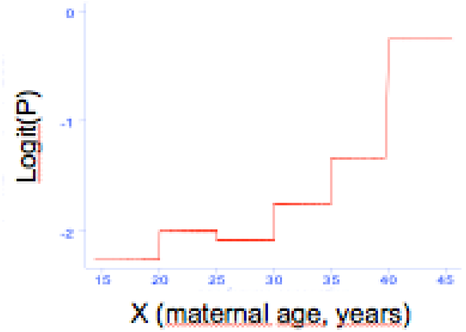
Regression Modeling in Epidemiology and Clinical Research **(3)**



# Fundamentals of **Logistic** Regression Modelling

## with applications to Epidemiology and Clinical Research

**Marion Ouidir & Rémy Slama**

Inserm and University Grenoble Alpes

IAB, Team of Environmental Epidemiology, Grenoble

# Let us start with a good news

He who knows linear regression will quickly master logistic regression.

# Overview

I. Motivation

II. Writing the logistic model and predicting disease risk

III. Interpreting the parameters of a logistic regression model

IV. Estimating the model’s parameters: the Maximum Likelihood method

V. Interaction (effect modification) in the logistic regression model

VI. Summary and perspectives

# I. Motivation

# When logistic regression is a relevant option…

- Linear regression applies to quantitative health parameters (i.e. represented by a variable taking many possible values).
- Sometimes, the health parameter corresponds to a binary outcome.
  - This is typically the case when data have been collected through a case-controls design
  - Information on the presence of a disease in subjects can also sometimes be assessed through a cross-sectional survey, and related to factors assessed at the same time

In this case, logistic regression is the right option.
Logistic regression allows to characterize the association between variables of various types (continuous, categorical…) and a binary outcome.

(Actually, logistic regression and its extensions can also be used if the outcome can be assessed on a categorical scale, i.e. through a variable taking on few values)

# Logistic regression is not the most relevant option as soon as the disease is assessed by a binary variable…

- Suppose now that the data come from a **cohort study**,

  Then the duration of follow-up (and, for those who developed the disease, the duration before the occurrence of the disease) likely varies between subjects

  There are also possibly subjects lost to follow-up

  - Excluding them may cause a selection bias
  - Consider them as remaining disease-free may cause misclassification in Y (some may have developed the disease the week after they were lost to follow-up)
  - It would therefore be useful to take the duration of follow-up in consideration, what logistic regression does not easily allow…

In any case, logistic regression is not the right option here.

# II. Writing the logistic regression model and predicting the disease risk

# II.1) Logistic regression: the dependent variable

We will assume that we are interesting in modeling the risk of occurrence of a health event or disease coded by a binary variable Y (e.g., disease Yes/No) with values

$$Y \begin{cases} 0\text{: no disease} \\ 1\text{: disease} \end{cases}$$

For example, imagine that we would like to see how a woman’s age influences the risk of a pregnancy ending with a spontaneous abortion.

- Unit of observation: the pregnancy
- Y is 0 if the pregnancy ends with a live birth, 1 for a spontaneous abortion
- X: maternal age at the pregnancy start (years).

# The health event is assessed on a binary scale: intuitive approach (1)

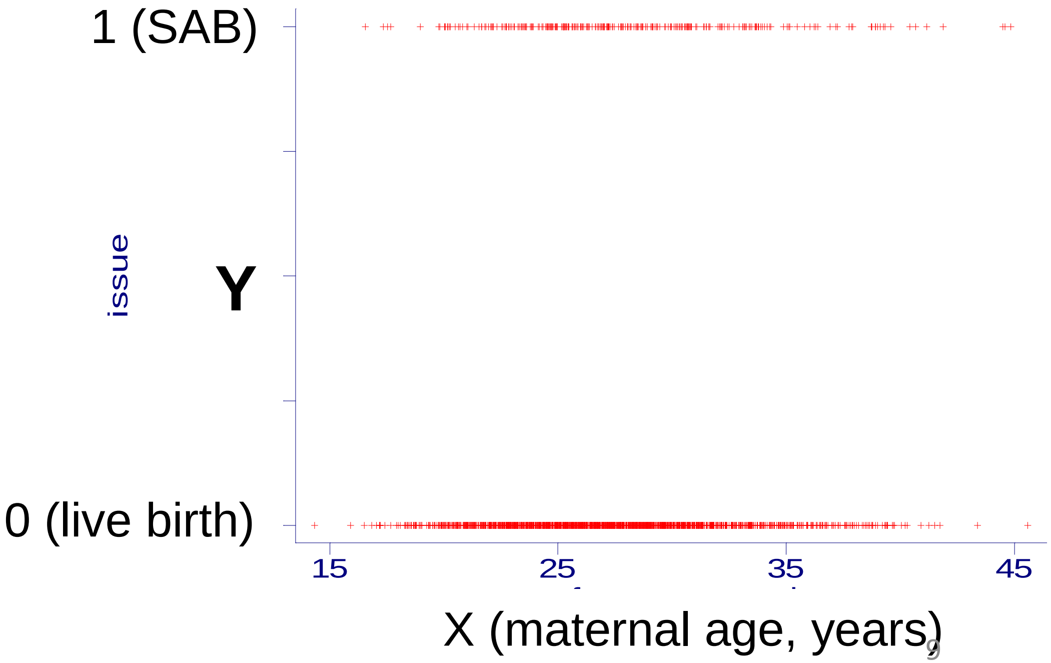
For linear regression (Y: quantitative variable), our first approach was to plot Y as a function of X.

For a binary explained variable, the scatterplot of Y according to X is usually not as informative as when Y is a continuous variable.

*Example:*

Y: occurrence of a spontaneous abortion (SAB) during the first 24 weeks of pregnancy (No=0/Yes=1)

X: Age of the woman (years)



# Estimating the frequency of disease from Y

*An interesting property of Y:*

If Y is a binary variable coded such as

$Y_i$ = 0 for healthy subjects (no disease, D-),

$Y_i$ = 1 for subjects who became sick (D+)

Then:

$\boldsymbol{E(Y) = Pr\,(D^{+})}$ = Probability(disease)

”The average value of Y in a subgroup of the population is an estimate of the probability of disease in this group.”

*Indeed, in the source population* (n observations):

$$E(Y) = \frac{\sum_{i=1}^{n} Y_i}{n} = \frac{\sum_{\text{D}^+\ \text{subjects}} Y_i + \sum_{D^-\ (healthy)\ subjects} Y_i}{n}$$

Where $\Sigma_{\text{D- subjects}}$ $(Y_i)$= 0 because Y=0 among healthy subjects.

and $\Sigma_{\text{D+ subjects}}$$(Y_i)$=(number of diseased subjects) x 1 because Y=1 among diseased subjects.

Therefore

$$E(Y / X = x) = \frac{\text{Number of diseased subjects (among those } X = x)}{\text{Total number of subjects with } X = x}$$

$$= P(disease / X = x) = P(Y = 1 / X = x)$$

**Notation:**

We will write indifferently E(Y) or P(Y=1) or Pr(disease) or P(x) to indicate E(Y / X=x)

# Describing the relation between X and Y (binary)- Intuitive approach (2)

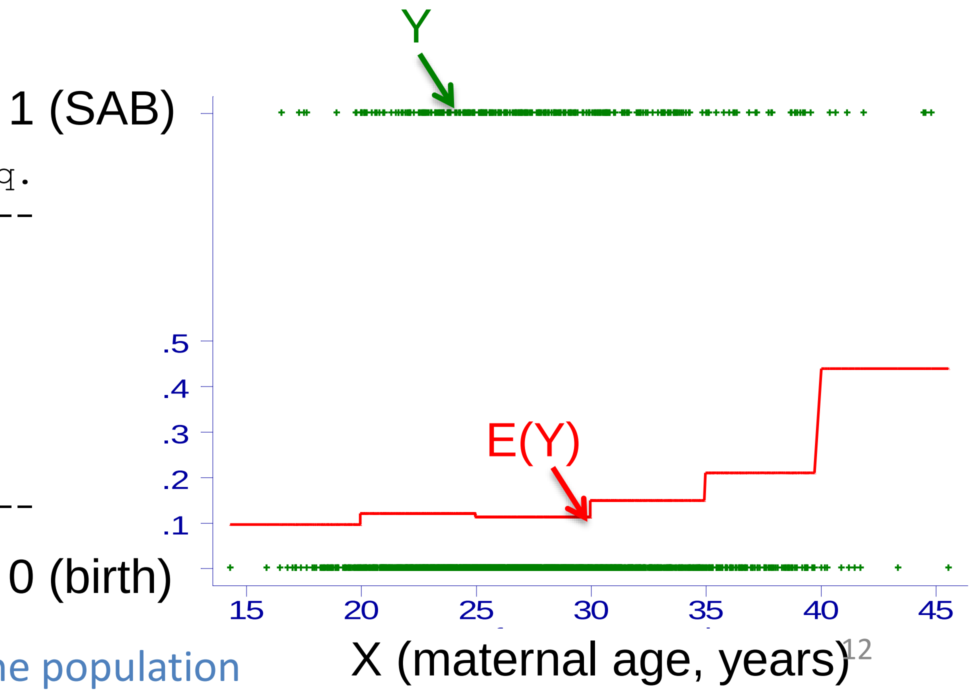
*Let us now make use of this property:*

If Y represents the occurrence of a spontaneous abortion (No=0/Yes=1), then the average value of Y in a group is an estimate of the frequency of spontaneous abortions in this group.

X has been categorized in 6 groups:

```
  age- years|   Summary of Probability SAB
  (categ.)  |        Mean      Std. Dev.    Freq.
------------+-----------------------------------
        <20 |     .0945946           0        74
      20-24 |    .11887073           .       673
      25-29 |    .11156186           .       986
      30-34 |    .14726841           0       421
      35-39 |    .20833333           .       120
       >=40 |        .4375           0        16
------------+-----------------------------------
      Total |    .12707424        .035    2290
```

Overall frequency of spontaneous abortion in the population



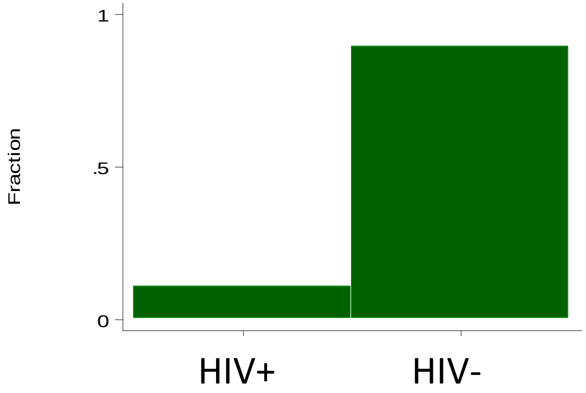
# *II.2) Several **types** of covariates*

Nothing new compared to what we have seen for linear regression…

- **Binary covariates**

  These only can take 2 values

  *e.g.:* sex = Female or male
  or HIV serology : HIV+/HIV-



- **Categorical covariates**

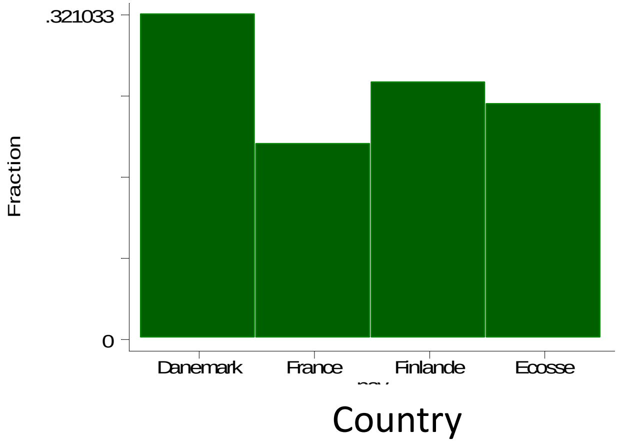
  These variables can take a finite (<<n) number of possible values, strictly greater than 2. They are either

  *-ordered*

  If there is a natural 'scale'
  e.g.: Number of children: 1, 2, 3, …,10.

  *-not ordered*

  If there is no obvious scale or ranking among values
  E.g.: City of recruitment (in a multicentric study)



- **Continuous (or continuous quantitative) covariates**

# II.3) Writing the logistic model

In the previous graph, we got closer to the situation corresponding to a linear regression model E(Y)=α + β.X

A difference is that, since 0 ≤ Y≤1, its expectation E(Y) will also be between 0 and 1.

However, α + β.X can vary over a large range when X varies (if β≠0).

➔ A way to handle this would be to transform E(Y) (now written P) so as to obtain an expression that varies over a broad range of values (ideally, from –∞ to +∞).

➔ One solution to do so is to apply the logistic function to P:

$$\boxed{\ln\left(\frac{E(Y)}{1-E(Y)}\right)=\ln\left(\frac{P}{1-P}\right)=\operatorname{logit}(P)=a+bX}$$

# Mathematical properties of the *logit* (or logistic) function

*If P is close to 0*, then 1-P is close to 1 and P/(1-P) is close to 0.

Hence $\ln\left(\frac{P}{1-P}\right) = \mathrm{logit}(P)$ goes towards $-\infty$.

*If P is close to 1*, then 1-P is close to 0 and P/(1-P) is close to +∞.

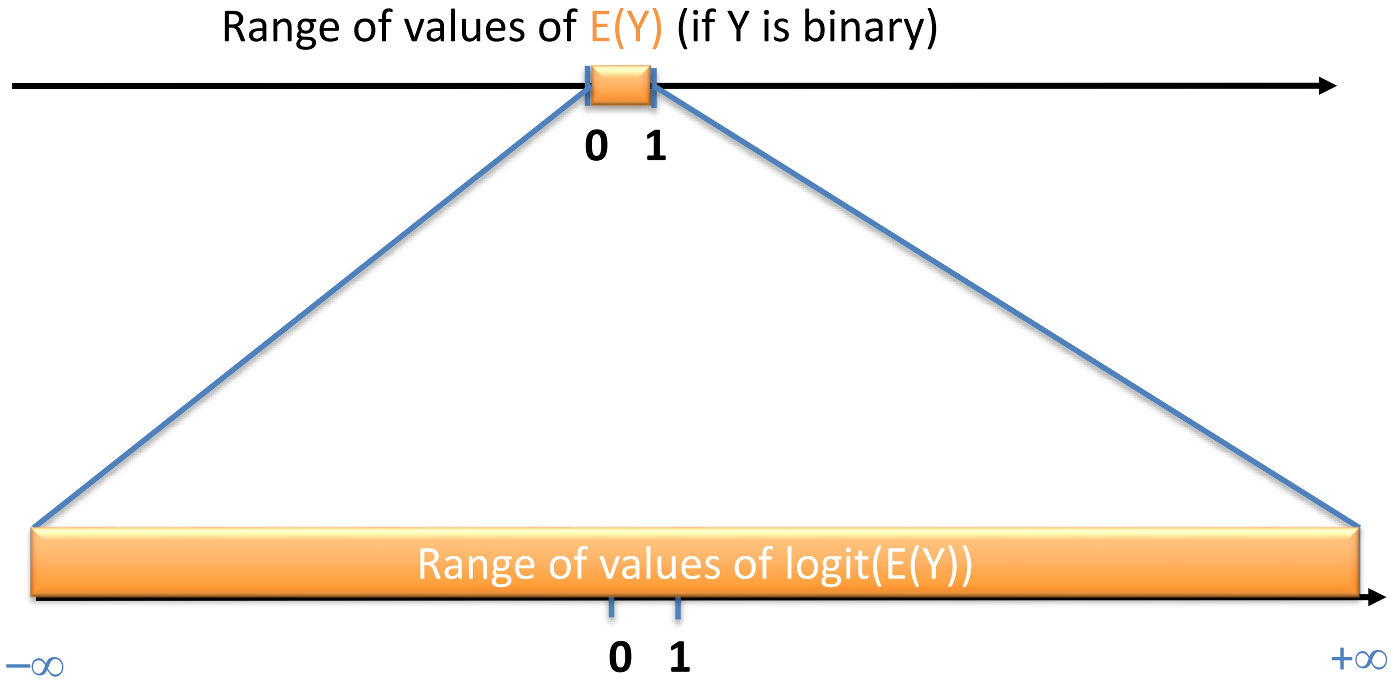
Hence $\ln\left(\frac{P}{1-P}\right) = \mathrm{logit}(P)$ goes towards $+\infty$.

To summarize:

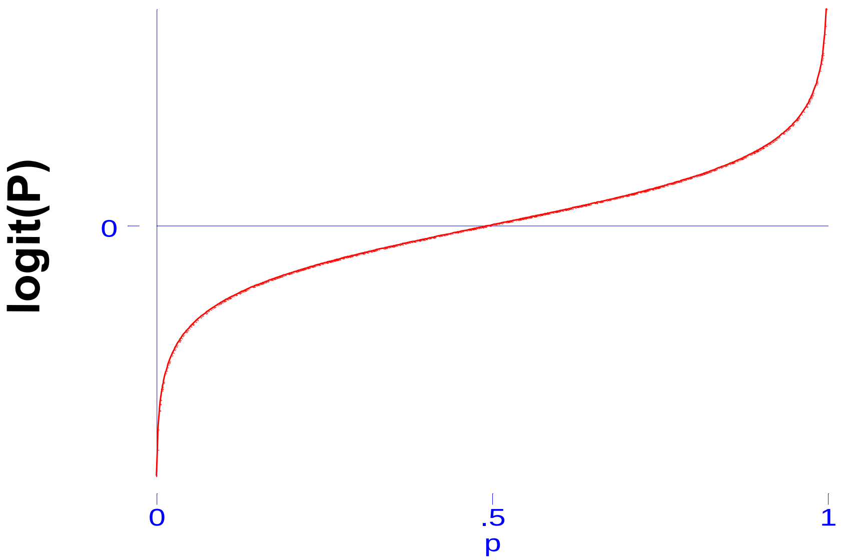
**logit(P)** is a continuous function,
defined for values of P in the interval ]0 ; 1[,
that takes all values between -∞ and +∞ when P varies from 0 to 1
and has a one to one equivalence with P.

By applying the logistic function to E(Y), we transformed it into a variable that takes value over a large range

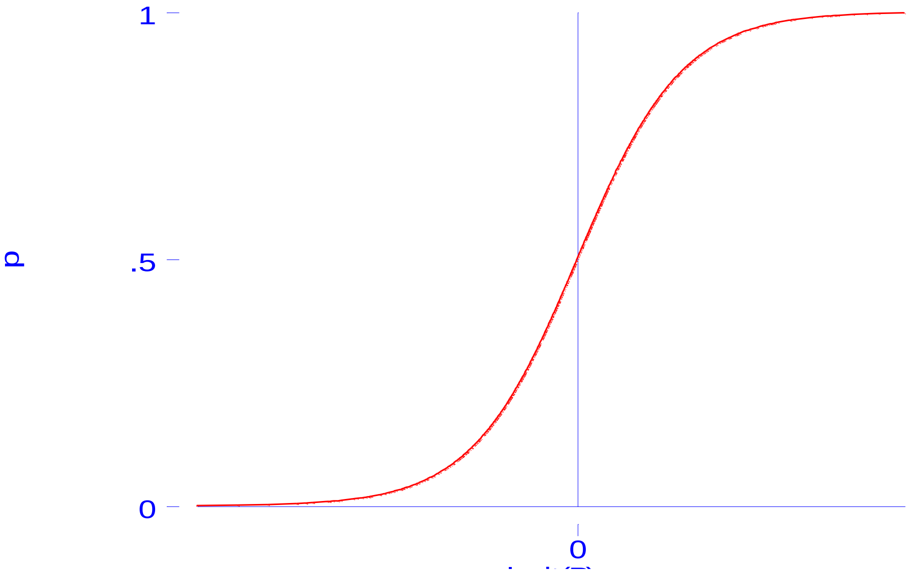
…so that we have gotten closer to the situation of linear regression.


Range of values of E(Y) (if Y is binary)
0
1
Range of values of logit(E(Y))
0
1
−∞
+∞

# Graphical representation of the *logit* (or logistic) function



The logistic function



Its reciproqual (α+βX →P)

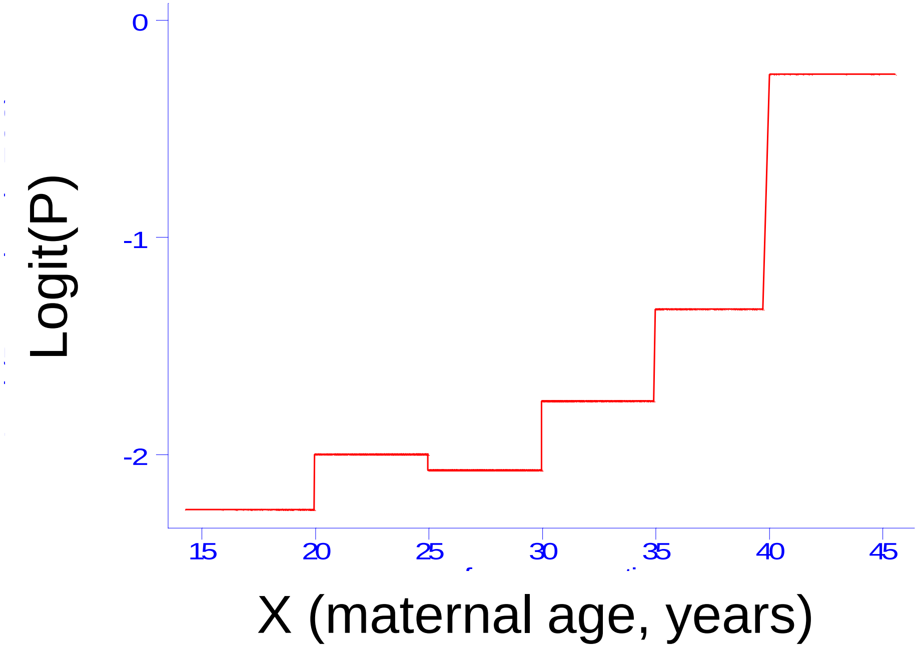
# Describe the relation between X and binary Y – Moving further with the intuitive approach (3)

Let us go back to the example of maternal age and spontaneous abortions:

Instead of plotting the average value of Y (or P, the probability of spontaneous abortion) in each age group, Let us plot logit[E(Y)]=logit(P):

```
.  gen logitp=ln((issuem)/(1-issuem))
. graph logitp agf
. tab agfc, sum(logitp)
```

```
  age (yrs) |  Summary of Logit(Probability SAB)
  (categor.)|        Mean   Std. Dev.       Freq.
------------+------------------------------------
        <20 |  -2.2587824           0          74
      20-24 |  -2.0031679           .         673
      25-29 |  -2.0748858           .         986
      30-34 |  -1.7561879           0         421
      35-39 |  -1.3350011           .         120
       >=40 |  -.25131443           0          16
------------+------------------------------------
      Total |  -1.9496487   .23987161        2290
```



# II.4) Other ways to write the logistic regression model

So far, we used this formulation:

$$\ln\left(\frac{P}{1-P}\right) = \text{logit}(P / X = x) = \alpha + \beta x$$

If we apply the exponential function to both terms:

$$\exp\left\{\ln\left[\frac{P}{1-p}\right]\right\} = \exp(\alpha + \beta x)$$

$$\frac{P}{1-P} = \exp(\alpha + \beta x)$$

$$P = (1-P).\ \exp\ (\ \alpha + \beta.x\ )$$

$$P = \exp\ (\ \alpha + \beta.x\ ) - P.\ \exp(\ \alpha + \beta.x\ )$$

$$P\ [1+\exp(\ \alpha + \beta.x\ )] = \exp(\ \alpha + \beta.x\ )$$

*Finally:*

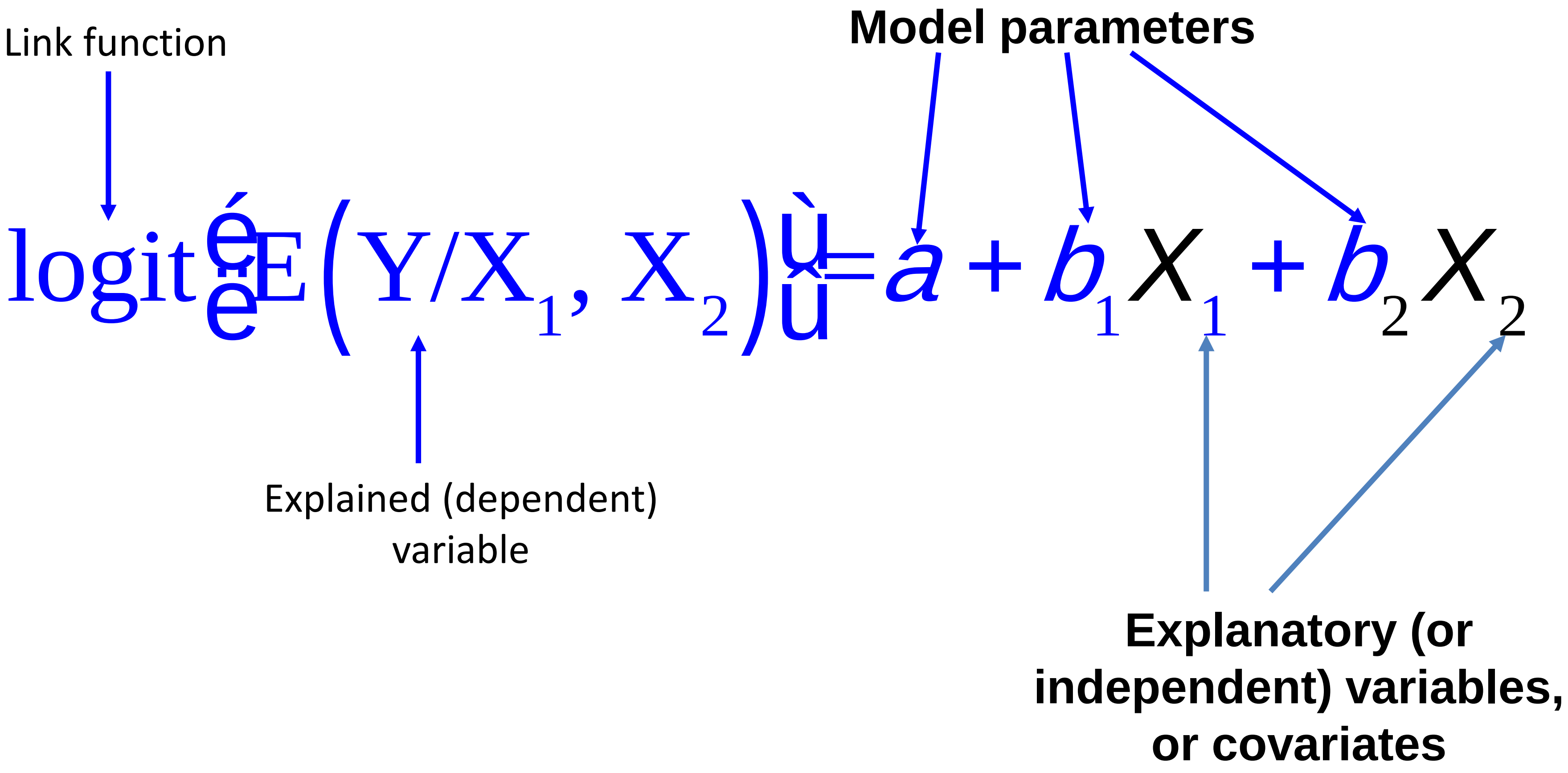
$$P = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)} = \frac{e^{\alpha + \beta x}}{1 + e^{\alpha + \beta x}} = \frac{1}{1 + \exp\left[-(\alpha + \beta x)\right]}$$

$$\boxed{\ln\left(\frac{P}{1-P}\right) = \operatorname{logit}(P / X = x) = \alpha + \beta x} \qquad (1)$$

$$\boxed{P \quad = \quad \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)} \quad = \quad \frac{e^{\alpha + \beta x}}{1 + e^{\alpha + \beta x}}} \qquad (2)$$

$$\boxed{P \quad = \quad \frac{1}{1 + e^{[-(\alpha + \beta x)]}} \quad = \quad \frac{1}{1 + \exp\left[-(\alpha + \beta x)\right]}} \qquad (3)$$

(1), (2) and (3) are equivalent



Link function

Model parameters

$$\text{logit}\,\text{é}\,\text{ë}\,E\left(Y/X_1,\ X_2\right)\text{ù}\,\text{û} = a + b_1 X_1 + b_2 X_2$$

Explained (dependent) variable

Explanatory (or independent) variables, or covariates

# II.5) Values predicted by a logistic model

Formulas 2 and 3 above can be used to estimate the probability of disease P predicted by the logistic model.

*For example:*

If we assume that the estimation of the models’ parameters yielded $\alpha$= -13 and $\beta$=0.5

Then the probability of disease among subjects for whom X = 30 years is:

$$P(Y=1/X=30) \quad = \quad \frac{1}{1+\exp\left[-(a+bx)\right]}=\frac{1}{1+\exp\left[-(-13+0.5\times 30)\right]}=0.88$$

If you want an estimation of the number of cases in the group of subjects with age 30, one solution is to multiply this probability by the number of subjects in this age category. For 100 subjects, the model expects 88 cases.

For a given subject, you may consider that the model predicts (s)he is a case by comparing his predicted probability of disease with 0.5.

# Predicting disease risk from a logistic model: *Caveats*

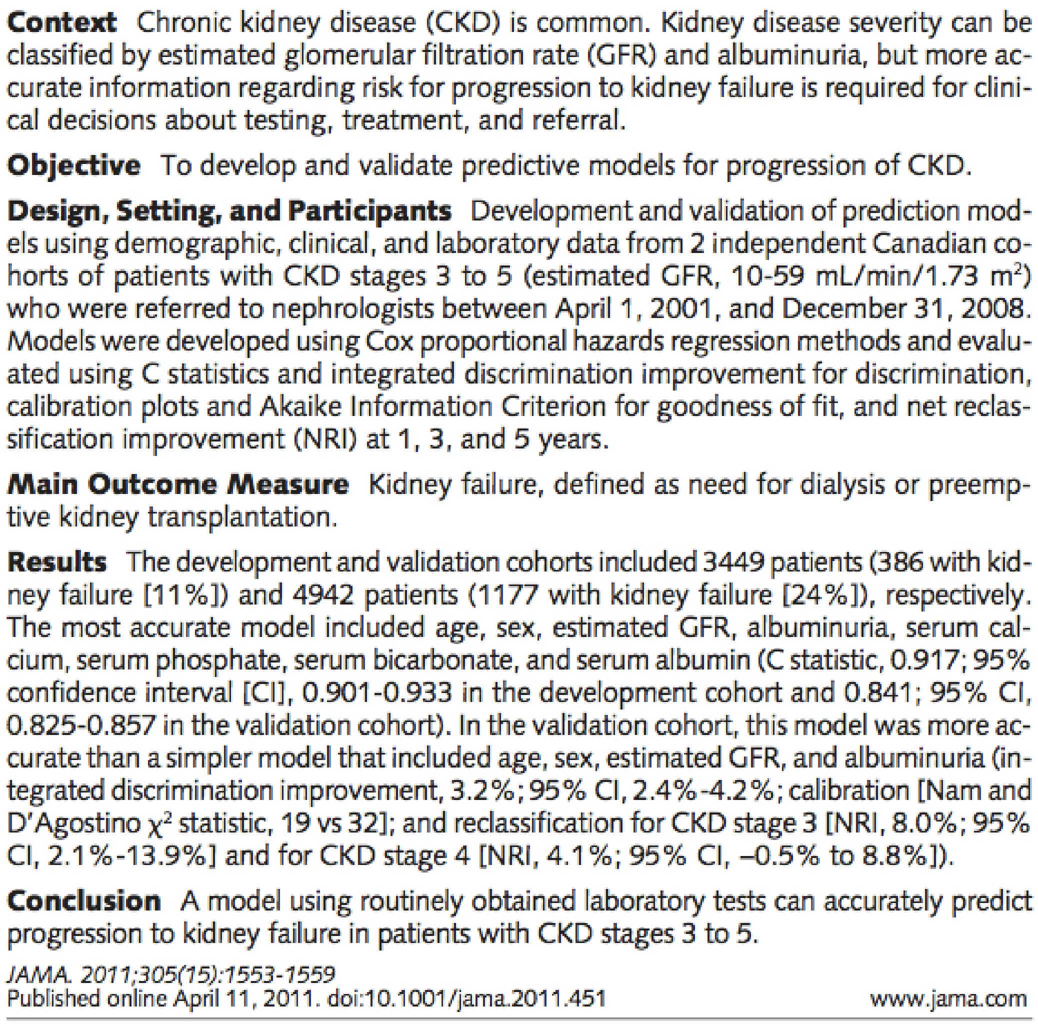
1) The probability of disease should not be predicted for subjects outside the range of the values of the covariates observed in the dataset used to estimate the model's parameters (e.g., do not predict the risk of SAB for a 20-year old woman if most women in the study are 25 years or older).

2) Case-control studies: the probability of disease in a case-control population is totally driven by the number of controls chosen for each case.

If there are 2 controls per case, then disease risk will be around 33%, but this is not an interesting information!

# Example

# A Predictive Model for Progression of Chronic Kidney Disease to Kidney Failure



**Context** Chronic kidney disease (CKD) is common. Kidney disease severity can be classified by estimated glomerular filtration rate (GFR) and albuminuria, but more accurate information regarding risk for progression to kidney failure is required for clinical decisions about testing, treatment, and referral.

**Objective** To develop and validate predictive models for progression of CKD.

**Design, Setting, and Participants** Development and validation of prediction models using demographic, clinical, and laboratory data from 2 independent Canadian cohorts of patients with CKD stages 3 to 5 (estimated GFR, 10-59 mL/min/1.73 $m^2$) who were referred to nephrologists between April 1, 2001, and December 31, 2008. Models were developed using Cox proportional hazards regression methods and evaluated using C statistics and integrated discrimination improvement for discrimination, calibration plots and Akaike Information Criterion for goodness of fit, and net reclassification improvement (NRI) at 1, 3, and 5 years.

**Main Outcome Measure** Kidney failure, defined as need for dialysis or preemptive kidney transplantation.

**Results** The development and validation cohorts included 3449 patients (386 with kidney failure [11%]) and 4942 patients (1177 with kidney failure [24%]), respectively. The most accurate model included age, sex, estimated GFR, albuminuria, serum calcium, serum phosphate, serum bicarbonate, and serum albumin (C statistic, 0.917; 95% confidence interval [CI], 0.901-0.933 in the development cohort and 0.841; 95% CI, 0.825-0.857 in the validation cohort). In the validation cohort, this model was more accurate than a simpler model that included age, sex, estimated GFR, and albuminuria (integrated discrimination improvement, 3.2%; 95% CI, 2.4%-4.2%; calibration [Nam and D'Agostino $\chi^2$ statistic, 19 vs 32]; and reclassification for CKD stage 3 [NRI, 8.0%; 95% CI, 2.1%-13.9%] and for CKD stage 4 [NRI, 4.1%; 95% CI, −0.5% to 8.8%]).

**Conclusion** A model using routinely obtained laboratory tests can accurately predict progression to kidney failure in patients with CKD stages 3 to 5.

*JAMA. 2011;305(15):1553-1559*
Published online April 11, 2011. doi:10.1001/jama.2011.451 www.jama.com

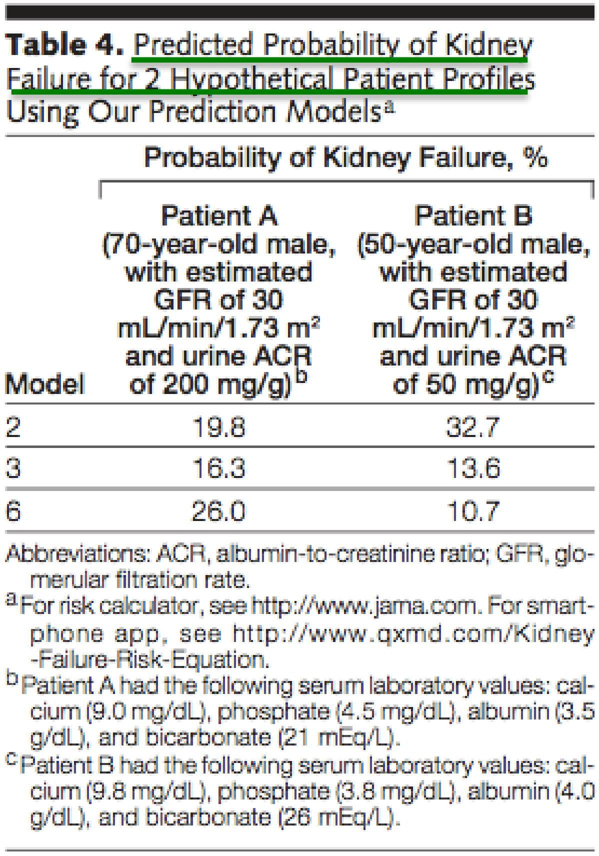


**Table 4.** Predicted Probability of Kidney Failure for 2 Hypothetical Patient Profiles Using Our Prediction Models[a]

| | Probability of Kidney Failure, % | |
|---|---|---|
| Model | Patient A (70-year-old male, with estimated GFR of 30 mL/min/1.73 $m^2$ and urine ACR of 200 mg/g)[b] | Patient B (50-year-old male, with estimated GFR of 30 mL/min/1.73 $m^2$ and urine ACR of 50 mg/g)[c] |
| 2 | 19.8 | 32.7 |
| 3 | 16.3 | 13.6 |
| 6 | 26.0 | 10.7 |

Abbreviations: ACR, albumin-to-creatinine ratio; GFR, glomerular filtration rate.

[a] For risk calculator, see http://www.jama.com. For smartphone app, see http://www.qxmd.com/Kidney-Failure-Risk-Equation.

[b] Patient A had the following serum laboratory values: calcium (9.0 mg/dL), phosphate (4.5 mg/dL), albumin (3.5 g/dL), and bicarbonate (21 mEq/L).

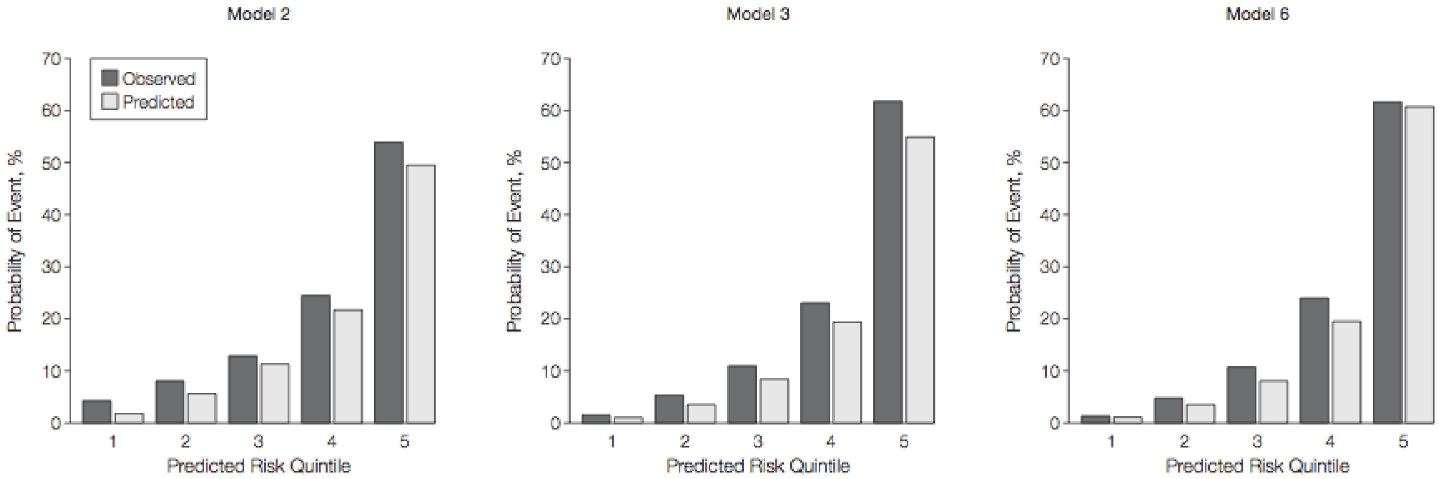
[c] Patient B had the following serum laboratory values: calcium (9.8 mg/dL), phosphate (3.8 mg/dL), albumin (4.0 g/dL), and bicarbonate (26 mEq/L).

(Tangri et al., JAMA, 2011)

# A Predictive Model for Progression of Chronic Kidney Disease to Kidney Failure

**Figure.** Observed vs Predicted Probability of Kidney Failure at 3 Years Using Models 2, 3, and 6 in the Validation Cohort

Probability of event



The predicted and observed event probability estimates represent the mean predicted probability from the Cox proportional hazards regression model and the mean observed probability from the population (Kaplan-Meier estimate) divided into quintiles of predicted probability. Predicted risk categories for quintiles 1 through 5 correspond with 0% to 4.3%, 4.4% to 8.1%, 8.2% to 12.9%, 13.0% to 24.5%, and 24.6% to 53.9%, respectively, for model 2; 0% to 1.6%, 1.7% to 5.3%, 5.4% to 11.0%, 11.1% to 23.1%, 23.2% to 61.7%, respectively, for model 3; and 0% to 1.4%, 1.4% to 4.8%, 4.9% to 10.7%, 10.8% to 24.0%, 24.1% to 61.6%, respectively, for model 6. Nam and D'Agostino $\chi^2$ statistic is 37, 32, and 19 for models 2, 3, and 6, respectively.

(Tangri et al., *JAMA*, 2011)

# III. Interpretation of the parameters of the logistic regression model

We will consider a multiple logistic regression model, assuming the existence of 2 covariates $X_1$ and $X_2$.

- For example, Y can be the outcome of the pregnancy (spontaneous abortion  Yes/No), $X_1$ age (years) and $X_2$ tobacco smoking.

$$\boxed{\ln\left(\frac{P}{1-P}\right)=\operatorname{logit}(P)=\operatorname{logit}\left[E(Y)\right]=\alpha+\beta_1 X_1+\beta_2 X_2}$$

# The parameter α (constant value)

*If we write the model in subjects for whom all covariates have a value of 0* ($X_1$=$X_2$=0)

Let $P_0$ be the probability of disease in these subjects:

$$\ln\left(\frac{P_0}{1-P_0}\right)=\text{logit}(P_0)=a+b_1.X_1+b_2.X_2=a+b_1.0+b_2.0=a \qquad \text{(III.2)}$$

$$\frac{P_0}{1-P_0}=\exp(a),\ \text{soit } P_0=(1-P_0)\exp\left(a\right),$$

If we exponentiate the formula above:

$$P_0[1+\exp(a)]=\exp(a)$$

Finally:

$$P_0=\frac{\exp(a)}{1+\exp(a)} \qquad (3)$$

Hence α allows estimating **$P_0$**, the disease **probability among subjects for whom all covariates have value 0.**

$$P_0 = \frac{\exp(\alpha)}{1+\exp(\alpha)}$$ can be called the « baseline disease risk ».

## *Caution (1): Extrapolation*

This value does not always have an interpretation.

In particular, this expression has no interpretation if subjects with covariate value 0 ($X_1=X_2=0$) do not correspond to a group of the population.

*Example:* If X is age and Y is infecondity, $P_0$ would correspond to infecondity risk at age 0.

## *Caution (2): Case-controls studies*

In a case controls study, the absolute value of the disease risk is of little interest.

Indeed, it mostly depends on the number of controls chosen for each case (for example, the average disease risk is 33% if 2 controls have been chosen for each case). It does not bear real information on the disease risk in the source population.

# Interpretation of parameter β

- *For subjects with $X_1=0$* and $X_2=x_2$

We write $P_{X1=0}$ the disease probability in this group of subjects:

$$\ln\left(\frac{P_0}{1-P_0}\right)=\mathrm{logit}(P_0)=a+b_1.X_1+b_2.\mathrm{X}_2=a+b_2.\mathrm{x}_2 \qquad \text{(III.4)}$$

- *For subjects with $X_1$**=1**, (*$X_2$ still having the same value $x_2$)

We write $P_{X1=1}$ the disease probability in this group of subjects:

$$\ln\left(\frac{P_1}{1-P_1}\right)=\mathrm{logit}(P_1)=a+b_1.X_1+b_2.\mathrm{X}_2=a+b_1+b_2.\mathrm{x}_2 \qquad \text{(III.5)}$$

$\beta_1$ can be expressed by subtracting (4) from (5):

(III.5)-(III.4) $\ln\left(\frac{P_1}{1-P_1}\right) - \ln\left(\frac{P_0}{1-P_0}\right) = \alpha + \beta_1 + \beta_2.X_2 - (\alpha + \beta_2.X_2) = \beta_1$

Note that $\ln\left(\frac{P_1}{1-P_1}\right) - \ln\left(\frac{P_0}{1-P_0}\right) = \ln\left(\frac{P_1}{1-P_1}\right) + \ln\left(\frac{1-P_0}{P_0}\right) = \ln\left[\left(\frac{P_1}{1-P_1}\right)\left(\frac{1-P_0}{P_0}\right)\right]$

and that $\left(\frac{P_1}{1-P_1}\right)\left(\frac{1-P_0}{P_0}\right) = \mathrm{OR}$

We recognize here the odds-ratio of disease associated with $X_1$. (The relative risk would simply be $P_1/P_0$ )

Hence: $\beta_1 = \ln\left[\left(\frac{P_{(X_1=1)}}{1-P_{(X_1=1)}}\right)\left(\frac{1-P_{(X_1=0)}}{P_{(X_1=0)}}\right)\right] = \ln\left(\frac{P_{(X_1=1)}/(1-P1)}{P0/(1-P0)}\right) = \ln\left(\mathrm{OR}_{(X_1=1\text{ vs.}X_1=0)}\right)$

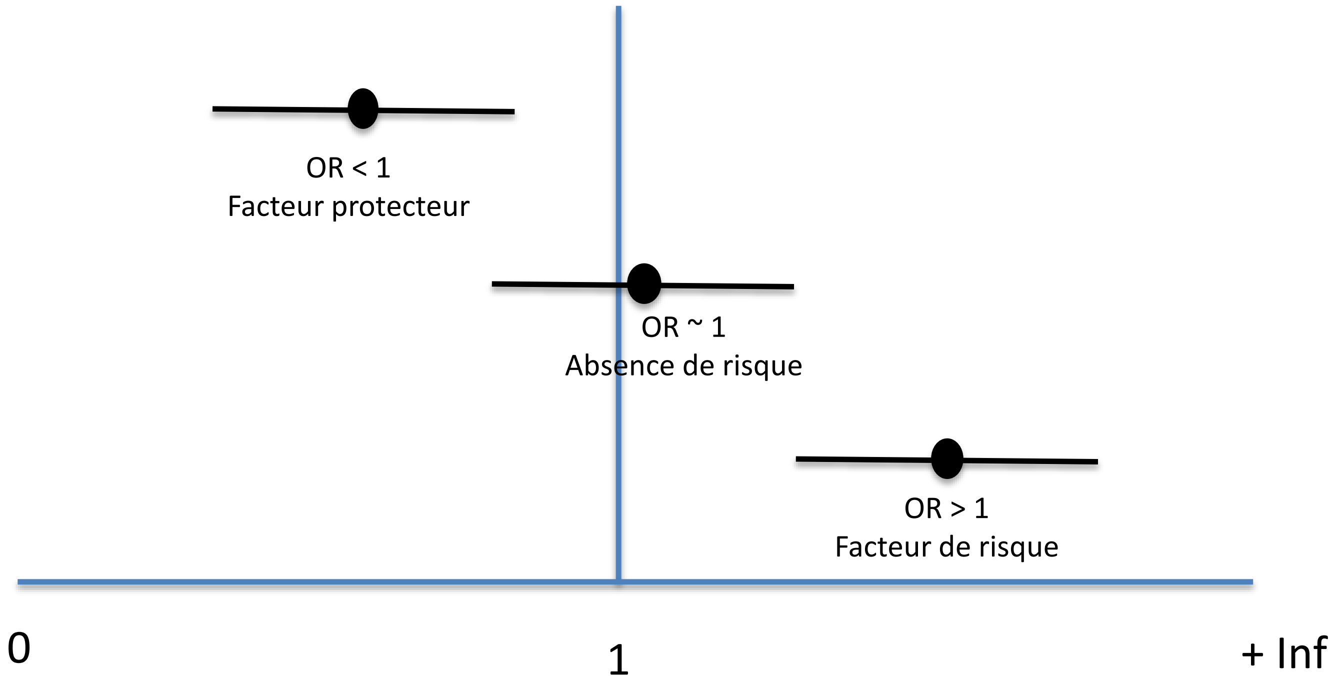
Equivalently: $$\boxed{\exp(\hat{\beta}_1) = \hat{\mathrm{O}}\mathrm{R}_{(X_1=1\text{ versus }X_1=0)}}$$

**exp(β$_1$)** is thus an estimate of the odds-ratio of disease comparing subjects in which $X_1$=1 and $X_1$=0, when $X_2$ has a given value $x_2$.

It is an estimate of the OR of disease associated with $X_1$ adjusted for $X_2$.

# ODDS RATIO

- Nombre sans unité [0, +Inf[
- Prévalence faible (<10%) OR ~ RR



$$\boxed{\exp(\hat{b}_1) = \hat{O}R_{(X_1 = 1 \text{ versus } X_1 = 0)}}$$

# Illustration

In a cross sectional study among pregnant women, time from end of contraception use to the start of the pregnancy has been collected retrospectively.

***Y: lack of pregnancy within 12 months after end of contraceptive use.***

***X: Previous history of gynecological disorder.***

**Y**: month12=0 if the pregnancy starts 12 months or less after the end of contraception use
month12=1 if the pregnancy starts more than 12 months after the end of contraceptive use.

**X**: gynec=0 without history of gynecological disorder
gynec=1 if the woman already had such a disorder.

```
. logit month12 gynec
(…)
Logit estimates                                   Number of obs   =        926
                                                  LR chi2(1)      =       4.53
                                                  Prob > chi2     =     0.0333
Log likelihood = -277.09783                       Pseudo R2       =     0.0081

------------------------------------------------------------------------------
  month12 |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
----------+-------------------------------------------------------------------
  gynec   |   .5673402   .2576911     2.202   0.028       .0622749    1.072405
   _cons  |  -2.451881   .1356798   -18.071   0.000      -2.717809   -2.185954
------------------------------------------------------------------------------
```

$P_0 =$ OR =

# Reminder: Principle of the case-controls design

We have mentioned that the logistic regression model is particularly suited to the case-control design. Its principles are that:

- Cases of the disease and controls are sampled separately
- Exposure is assessed retrospectively at inclusion
- **Cases** can either be all (or a given proportion) of ***incident*** **cases** (those newly diagnosed/identified after the start of the study)
- or of ***prevalent*** **cases** (already present/treated in a given clinical department when the study starts)
- Controls should be free of the disease at the time of inclusion of cases and should correspond to the population who would be recruited as cases, should they contract the disease

**Thus, controls need not (always) be representative of the general population**

# Measure of the association between exposure and health in a case-control study

- **Incident case-control study**

  The OR can be interpreted as the ratio of the **hazard rates** of the health outcome across exposure categories (just like in a cohort study).

  If the OR associated with a binary covariate X is 2.0, at each time, subjects exposed to X have twice the instantaneous risk of developing the disease of unexposed subjects.

No “rare disease” assumption necessary

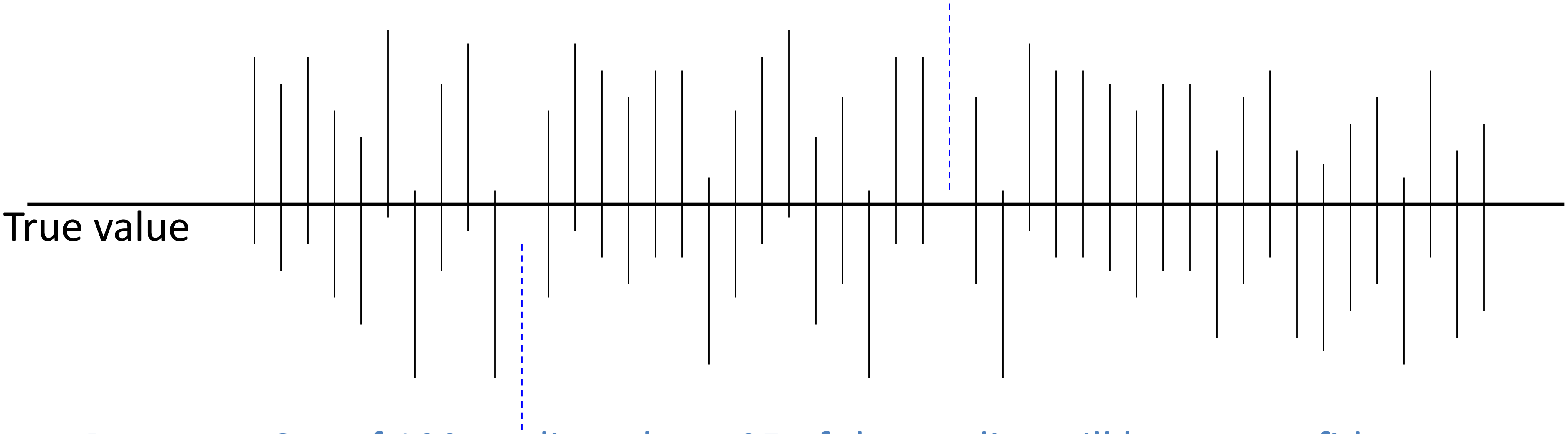
- **Prevalent case-control study**

  The OR varies in the same direction as the incidence rate ratio that would be estimated in the corresponding cohort study (but is generally not equal to it).

# Side remark: Confidence Intervals

Assume that a study with a given design is conducted a large number of times: random sampling of study participants, estimation of the average of a given quantity (e.g., triglyceride levels) in this population, together with its 95% confidence intervals.

The confidence intervals estimated in each of these studies are shown below (in dotted line if the true unmeasured value of the parameter is not included in the confidence interval).



**Property:** Out of 100 studies, about 95 of the studies will have a confidence interval containing the true value.

This is why it is sometimes said that a 95% confidence interval has “95% probability to contain the true value”.

This property only holds in the hypothesis of a lack of bias. If there is confounding or selection bias, it may be that not a single CI contains the true value…

CI take into account random fluctuations, but not the other sources of error.

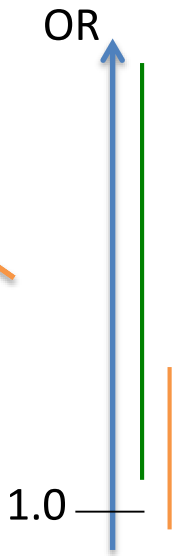
# Side remark: **Confidence Intervals**

- Confidence intervals allow to assess how informative a study is.

**Which study is more informative?**

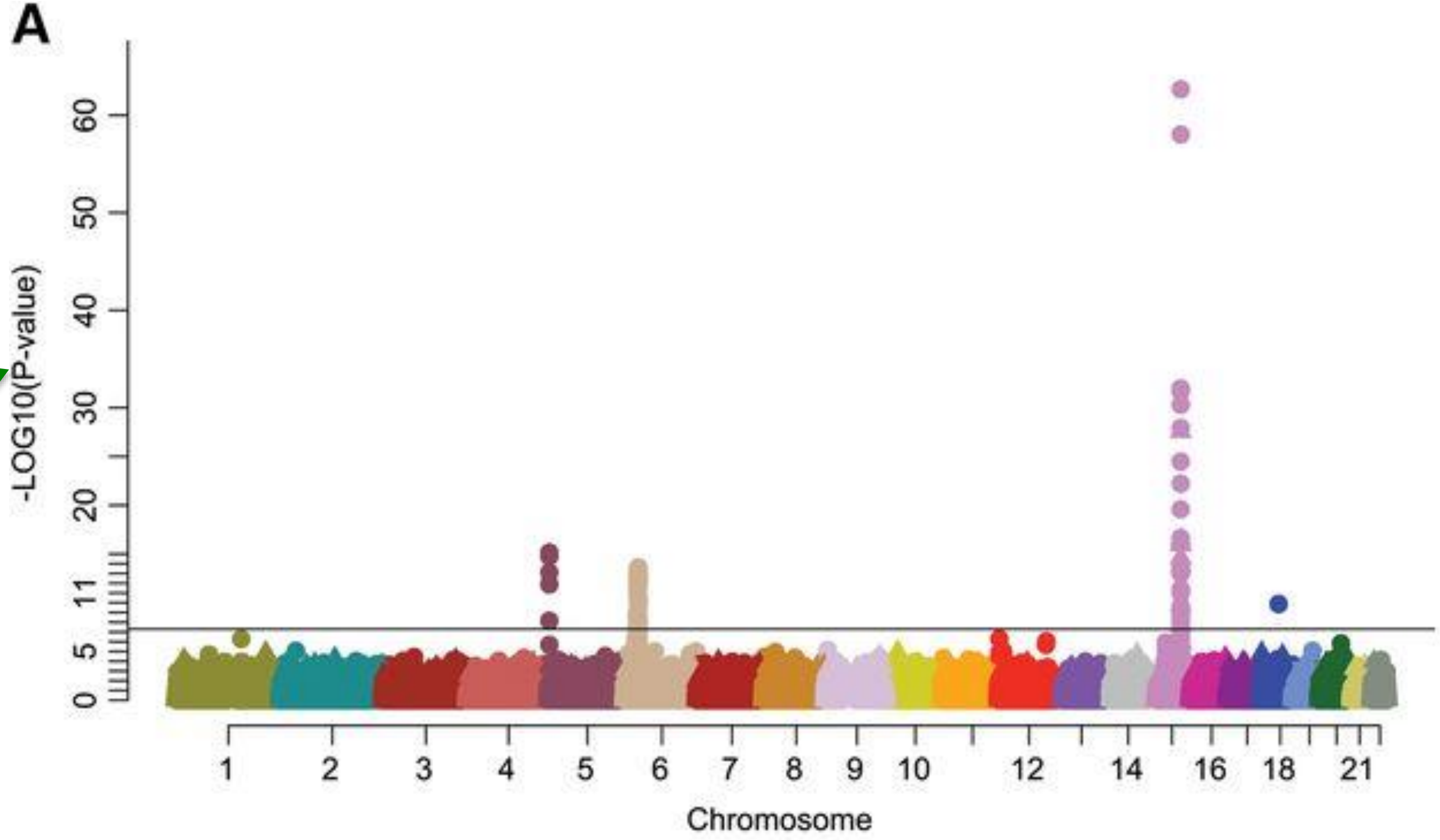
~~1. A study in which the confidence interval of the parameter of interest (Odds-Ratio…) **does not contain the value 1.0** ?~~

2. A study in which the confidence interval of the parameter of interest (Odds-Ratio…) is **narrow** (and possibly includes 1.0)?



# Example: Logistic Regression.
# Identification of genetic risk factors for lung cancer through a genome wide association study (GWAS)

A meta-analysis ; 14,900 lung cancer cases and 29,500 controls



This is the p-value of the association between each single genetic polymorphism considered separately and lung cancer risk, adjusted for sex, age, cohort…

(Timofeeva, *Hum Mol Genetics*, 2012)

# *Interpreting $\beta$ if $X_1$ has more than 2 levels*

**Property:** If $X_1$ is an ordered categorical or continuous covariate, exp(β) is an estimate of the **OR of disease associated with an increase by 1 in $X_1$**, all other covariates remaining constant

$$\boxed{\exp(\hat{b})=\hat{O}R_{(X1=a+1\text{ versus }X1=a)}}$$

*Exercise*: prove it.

*Example:*

Assume $X_1$ is the subjects' age.

If the model estimates a value of the parameter associated with age of

$\beta_1=0.095$

Then the corresponding odds-ratio is $\exp(\beta_1)=1.1$

This means that, on average, the odds of disease is multiplied by 1.1 each time that age increases by 1 (an increase by about 10% in disease risk).

*In practice, the OR associated with an increase by 1 year of age is not very meaningful. It would be more explicit to give the OR associated with an increase by 10 years.*

This other OR can easily be expressed from β.

**Property:** The parameter associated with an increase by p in $X_1$ is $p\beta_1$. The corresponding OR is the OR associated with an increase by 1 in $X_1$ at the power p

**Proof:**

- *For subjects in whom $X_1$ has a value **a*** ($X_2$ having a given value $x_2$)

  We write $P_a$ the disease probability in this group:

$$\ln\left(\frac{P_a}{1-P_a}\right)=\mathrm{logit}(P_a)=\alpha+\beta_1.X_1+\beta_2.\mathrm{X}_2=\alpha+\beta_1.a+\beta_2.\mathrm{x}_2 \qquad \text{(III.6)}$$

- *For subjects in whom $X_1=a+p$* ($X_2$ still having a given value $x_2$)

  We write $P_{a+p}$ the disease probability in this group:

$$\ln\left(\frac{P_{a+p}}{1-P_{a+p}}\right)=\mathrm{logit}(P_{a+p})=\alpha+\beta_1.X_1+\beta_2.\mathrm{X}_2=\alpha+\beta_1.(a+p)+\beta_2.\mathrm{x}_2 \qquad \text{(III.7)}$$

By subtracting (6) from (7):

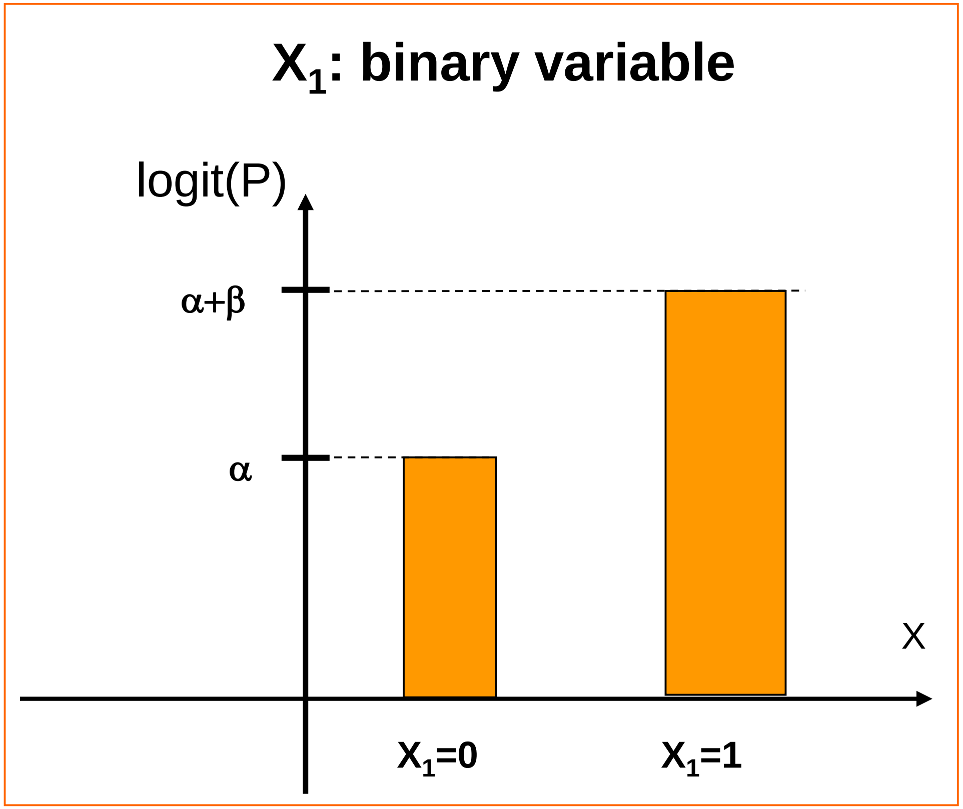
$$\ln\left(\frac{P_{a+p}}{1-P_{a+p}}\right)-\ln\left(\frac{P_a}{1-P_a}\right)=\ln(OR_{\mathrm{X1}=a+p\ \mathrm{vs.\ X1=a}})=\alpha+\beta_1.(a+p)+\beta_2.\mathrm{x}_2-\left(\alpha+\beta_1.a+\beta_2.\mathrm{x}_2\right)=p\beta_1 \quad \text{(7)-(6)}$$

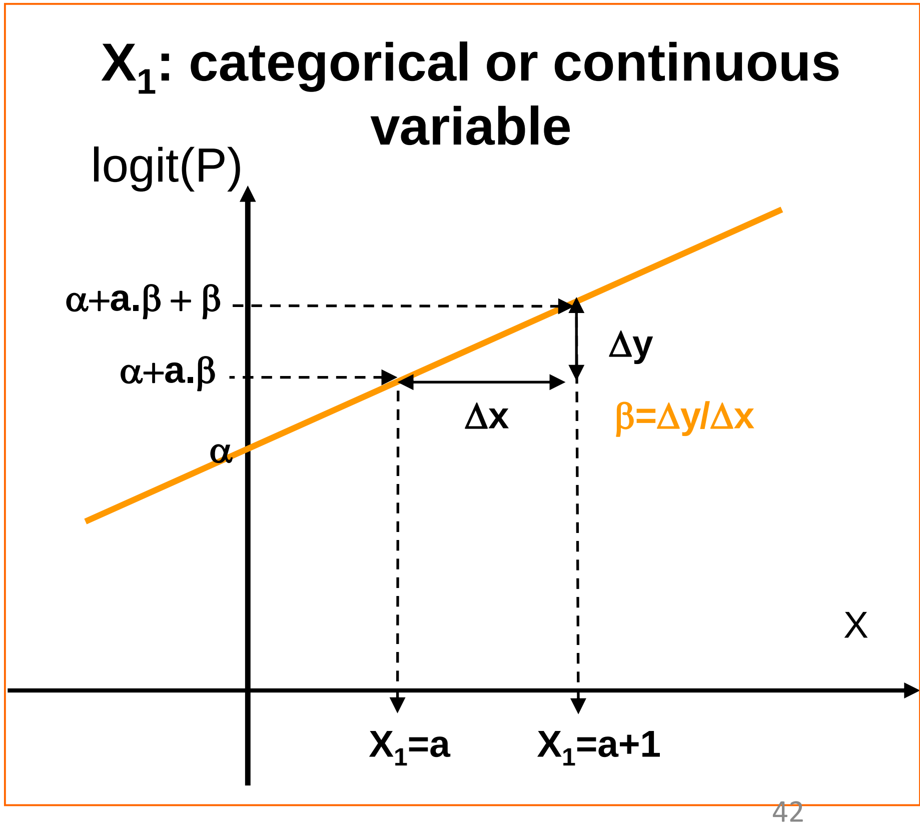
$$\boxed{OR_{X=a+p\ \mathrm{versus}\ X=a}=\exp\left(p\beta_1\right)=\left[\exp\left(\beta_1\right)\right]^p=\left[OR_{X=a+1\ \mathrm{versus}\ X=a}\right]^p}$$

# Interpretation of the parameters of the logistic model: graphical summary

$$\ln\left(\frac{P}{1-P}\right) = \operatorname{logit}(P) = \operatorname{logit}\left[E(Y)\right] = \alpha + \beta_1 X_1 + \beta_2 X_2$$

**$\exp(\beta_1) = OR_{(Xa+1 \text{ versus } Xa)}$, $X_2$ remaining constant.**





# $X_1$ can take on many different levels: *illustration*

## Relation between the risk of 12-month involuntary infertility and sperm morphology

Study among partners of pregnant women in which semen has been collected and the morphology of spermatozoa assessed

*Y:* **month12,** 12-month involuntary infecundity (No/Yes)

*X:* **typ**, proportion of spermatozoa with normal morphology, in % (continuous, from 0 to 100).

```
. logit  month12 typ
(…) Iteration 3 :   log likelihood =  -239.7109

Logit estimates                                   Number of obs   =        858
                                                  LR chi2(1)      =       5.56
                                                  Prob > chi2     =     0.0184
Log likelihood =  -239.7109                       Pseudo R2       =     0.0115

  mois12 |      Coef.   Std. Err.       Z      P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     typ |   -.018667   .0078502     -2.378   0.017       -.034053   -.0032809
   _cons |    -1.5223   .3839849     -3.964   0.000      -2.274897   -.7697036
```

OR of 12-month involuntary infecundity if **typ** increases by 1 (and 95%CI): 0.98 (0.97; 1.00)

OR of 12-month involuntary infecundity if **typ** increases by 10%:

$OR_{10\%\ typ}$=exp(10x(-0.018667))=exp(-0.18667)=0.83 (0.71; 0.97)

# Model’s assumptions

The model

$$\ln\left(\frac{P}{1-P}\right) = \mathrm{logit}(P) = a + b_1.X_1 + b_2.X_2 \quad (1)$$

assumes:

**1) That, for any given value of $X_2$, logit(P) is a linear function of $X_1$**

➔ According to this model, the change in the odds of disease is the same when age increases from 20 to 21 years, and when it increases from 40 to 41 years

**2) That the effect of $X_1$ is the same whatever the value of $X_2$**

i.e. that there is not effect measure modification of $X_1$ by $X_2$ (otherwise it would not make sense to estimate the effect of $X_1$ *adjusted* for $X_2$).

We will see later how these default assumptions can be avoided.

HHH

# The Hazards of Hidden Hypotheses

**One should not conclude from such an estimate that the risk (or the log odds) of disease regularly increases with the covariate.** On the contrary, this is a hypothesis done while writing the model, and whose plausibility needs to be checked.

Such hypotheses are in some cases not realistic:

If the outcome is the probability of involuntary infertility, it is known that the risk of involuntary infertility does not increase by the same amount from 25 and 35 years and from 35 and 45 years.

# Model’s residuals

As for linear regression, the residuals correspond to the difference between the observed value of Y for subject i and the value predicted by the model for this subject.

Since Y only takes on values of 0 or 1:

$\varepsilon_i = 1 – P(x_i)$ if $Y_i=1$

$\varepsilon_i = 0 – P(x_i)$ if $Y_i=0$

**Average of residuals:** 0

**Variance of residuals:** $P(x_i).[1-P(x_i)]$

This corresponds to the variance of a binomial function with parameter $P(x_i)$.

Thus the variance of residuals depends depends on the subject’s characteristics x. It is not the same for all subjects, contrarily to the situation in linear regression for which all residuals have the same distribution.

The logistic model thus assumes that residuals $\varepsilon$ follow a binomial distribution. For this reason, it is said that the logistic model belongs to the family of binomial regression models.

The estimation of the model’s parameters further assumes that all residuals are independent.

If this hypothesis is not true (e.g., if some observations are statistically dependent one from another, as would be the case, then the models’ estimates may not be valid.

# IV Estimating the parameters of a logistic regression model: *the Maximum Likelihood (ML) method*

# Estimating the model’s parameters

$$\ln\left(\frac{P}{1-P}\right)=\alpha+\beta_1 x_1$$

- The estimation of the parameters of a logistic regression model is (usually) done using the **maximum-likelihood (ML) method**.

- Just like for the sum of squares with linear regression, the principle is to find the values of parameters $\alpha$, $\beta$ that maximize a function (the likelihood function) quantifying how close the model is from the observed values.

- The maximum likelihood method allows obtaining an unbiased estimate of $\alpha$, $\beta_1$; it relies on the maximisation of what is called a likelihood function written l($\alpha$, $\beta_1$), or in a simpler way l($\boldsymbol{\beta}$).

# The likelihood function l

The likelihood ξ of an observation ($x_i$,$y_i$) is defined as the probability to observe it under the hypothesis that the model is true.

*If, for observation i, $y_i$ has a value 1, the likelihood is:*

$$\xi(x_i,y_i)=P(y_i=1/x_i)=\frac{\exp(\alpha+\beta x_i)}{1+\exp(\alpha+\beta x_i)}$$

*If, for observation j, $y_j$ has a value 0, the likelihood is:*

$$\xi(x_i,y_i)=P(y_i=0/x_i)=1-P(y_i=1/x_i)=\frac{1}{1+\exp(\alpha+\beta x_i)}$$

These 2 situations can be summarized by the following formula:

$$\xi(x_i,y_i)=P(y_i=1/x_i)^{y_i}.[1-P(y_i=1/x_i)]^{1-y_i}$$

(remember that $a^0=1$)

Finally, the likelihood function **l**, defined for the data as a whole, is written as the product of all individual likelihoods ξ($x_i$, $y_i$) :

$$l=\prod_{i=1}^{n}\xi(x_i,y_i)=\prod_{i=1}^{n}P(y_i=1/x_i)^{y_i}.[1-P(y_i=1/x_i)]^{1-y_i}$$

It is more convenient to consider the logarithm of this product, called the log-likelihood, and written L:

$$\mathrm{L} = \sum_{i=1}^{n} \ln\left[\xi(x_i, y_i)\right] = \sum_{i=1}^{n} \left\{ y_i \ln\left[P(y_i = 1/x_i)\right] + (1 - y_i)\ln\left[1 - P(y_i = 1/x_i)\right] \right\}$$

*(remember that ln(ab)=ln(a)+ln(b))*

***Simplified notation:***

$$\boxed{\mathrm{L} = \sum_{i=1}^{n} \{\, y_i \ln(P_i) \;+\; (1 - y_i)\ln(1 - P_i) \,\}}$$

L is in fact a function of parameters ($\alpha$, $\beta$) because $P_i$ is a function of ($\alpha$, $\beta$).

Note that the log-likelihood is always negative (because ln(P)<0 and ln(1-P)<0)

*Interpretation :*

The likelihood function depends both on the observed data ($y_i$) and on the predicted values $P_i$=$P(y_i=1/x_i)$.

**Table:** Likelihood $L_i$ of an observation i according to $y_i$ and $P_i$

| **Predicted value** $P_i$<br>**Observed value** $\mathbf{y_i}$ | $P_i \approx 1$ | $P_i \approx 0$ |
|---|---|---|
| $\mathbf{y_i}$=1 | $L_i=y_i\ln(P_i) \approx 0$ | $L_i=y_i\ln(P_i) << 0$ |
| $y_i$=0 | $L_i=(1-y_i)\ln(1-P_i) << 0$ | $L_i=(1-y_i)\ln(1-P_i) \approx 0$ |

*For a given observation i:*

- If the value $P_i$ predicted by the model is close to the observed one, the contribution of observation i to the log-likelihood will be close to 0
- If the value $P_i$ predicted by the model is very different from the observed one, the observation i will strongly decrease the log-likelihood

For 2 different possible levels of the model's parameter $\alpha$, $\beta$, the log-likelihood will be higher for the values of $\alpha$ and $\beta$ corresponding to a situation when the values $P_i$ predicted by the model are close to the observations $Y_i$.

The log-likelihood therefore constitutes an information on the fit of the model to the data.

This property is used by the maximum likelihood estimator:

It consists in identifying the values of $\hat{a}$ and $\hat{b}$ maximizing the log-likelihood and hence, in a way, the fit of the model to the data.

This estimator has the property of being unbiased and efficient (i.e., it is the estimator with the smallest variance) if the number of observations is large.

$\hat{a}$ and $\hat{b}$ are not observed values. They are estimates. They depend on the model's assumption and the validity thereof. You are never sure they are "true" – and if the model's assumptions are not verified, it is safer to assume that the model's estimates are not reliable.

*The maximum likelihood estimator in practice:*

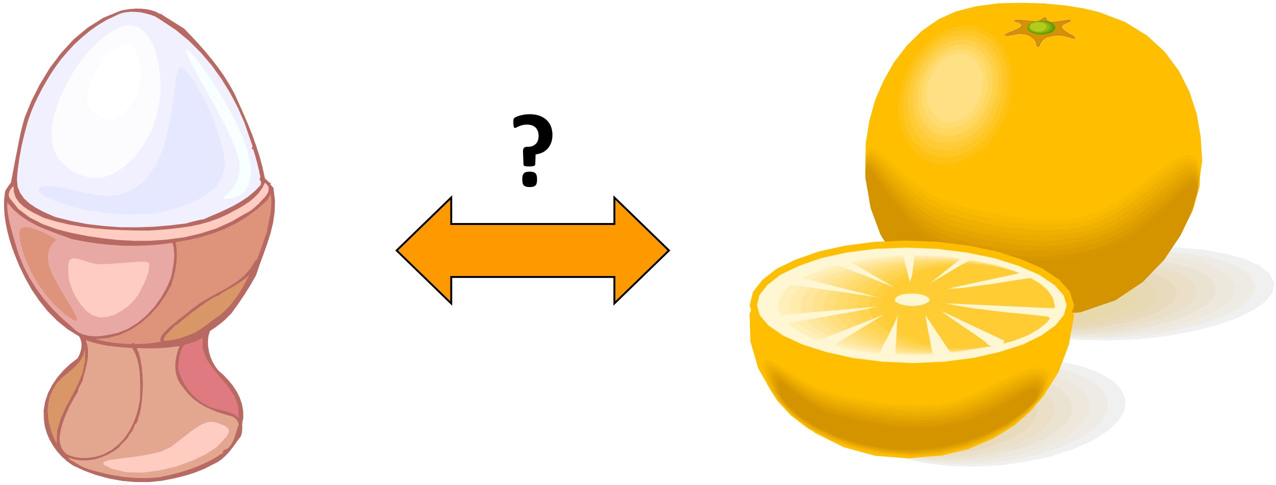
The maximum likelihood is an iterative approach.

At each step, new values of the model's parameters $\alpha$ and $\beta$ are tested to see if the log-likelihood can be increased. Once the log-likelihood has been maximized, the corresponding parameters' values are provided.

The model output generally provides the "final" log-likelihood of the model.

In some situations, the estimation process may not converge, or drop some of the covariates entered in the model. In this case, one should not consider the model's estimates (if any) as final but try understanding the causes of the problem.

# V. The Likelihood-Ratio (LR) test



Model 1

Model 2

# Principle of the likelihood-ratio (LR) test

This test allows to determine if adding one (or several) covariate(s) in a logistic model allows to improve the fit of the model to the data.

It provides a p-value associated with the covariate(s) added. If one considers adding several covariates (e.g., 4 dummy covariates corresponding to a variable with 5 categories), the likelihood-ratio test can provide a global significance test of all 4 covariates.

More generally, this test allows comparing 2 models in terms of goodness of fit, if one of these 2 models is **nested** within the other.

# Nested models

Two statistical models are said to be **nested** if one can be obtained by setting *constraints* on the values of the parameters of the other model.

*Constraint*: e.g., setting the value of a parameter to 0

In other words, if one of the model is a “simpler” form of the other model.

*For example*, model **(A)**

$$\mathrm{logit}(P)=\alpha + \beta_1 X_1 + \beta_2 X_2 \qquad \textbf{(A)}$$

is nested in model **(B)**

$$\mathrm{logit}(P)=\alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \qquad \textbf{(B)}$$

because if we set $\beta_3=0$ in (B), we obtain (A).

*BUT* model **(C)**

$$\mathrm{logit}(P)=\alpha + \beta_1 X_1 + \beta_4 X_4 \qquad \textbf{(C)}$$

is NOT nested in model **(B)**

because there is no way to obtain (C) by setting specific values to the parameters of (B), from which variable $X_4$ is absent.

Hence models (A) and (B) can be compared by a LR test, but not (B) and (C).

# Implementation of the LR test

We first consider the model:

$$\mathrm{logit}(P)=\alpha + \beta_1X_1 + \beta_2X_2 \qquad \textbf{(1)}$$

We would like to know if the introduction of covariate $X_3$ in the model could allow to improve the fit (model 2):

$$\mathrm{logit}(P)=\alpha + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 \qquad \textbf{(2)}$$

We assume that variables $X_1$, $X_2$ and $X_3$ are defined for exactly the same subjects, without missing data.

Since adding a covariate (with no missing data on the subjects of the previous model) automatically improves (even by a small amount) model fit, the log-likelihood of the model will increase (i.e., L(M2)>L(M1))

The principle of the LR test relies on the statistic G defined as:

**G=-2[L(model 1) – L(model 2)]**

= -2 x (log-likelihood of the *simpler* model – log-likelihood of the *richer* model)

## Property:

Under the assumption $H_0$ that the 2 models are equivalent

(i.e., the more elaborate model (2) does not really bring information, in other words, $X_3$ is not associated with the outcome after adjustment on $X_1$ and $X_2$)

Then G follows a $\chi^2$ distribution with one degree of freedom.

$$\boxed{G \sim \chi^2_{(1)} \qquad \text{if } \beta_3=0}$$

This means in practice that in order to perform the test, one should:

1) Write the 2 models to compare (and check that they are embedded)
2) Calculate G
3) Compare G to the $\chi^2$ distribution and calculate the corresponding p-value
4) If G is small (compatible with a $\chi^2$ distribution, or high p-value), then $H_0$ cannot be rejected (in other words, one cannot exclude that the simpler model is right, and that the new variables do not really bring information)
5) If G is larger and not compatible with a $\chi^2$ distribution (low p-value), then $H_0$ can be rejected. In other words, it is safer to prefer the more complex model.

# Likelihood-ratio test: illustration

Y = pregnancy outcome. (0=birth, 1=spontaneous abortion)

**Model 1:** 1 covariate (maternal age, agf).

```
. logit issue agf
```

```
Number of obs    =        2172
Log likelihood = -794.39076
```

```
------------------------------------------------------------------------------
   issue |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     agf |   .0374935   .0145041      2.585   0.010       .0090661    .0659209
   _cons |  -3.025602   .4097632     -7.384   0.000      -3.828723   -2.222481
------------------------------------------------------------------------------
```

**Model 2:** One considers adding the covariate corresponding to tobacco smoking during pregnancy (*tabag*, in cigarettes/day)

**$H_0$**: the parameter associated to tabag is 0

```
. logit issue agf tabag
Number of obs    =        2172
Log likelihood = -792.37743
------------------------------------------------------------------------------
   issue |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     agf |   .0397656   .0145416     2.735   0.006       .0112646    .0682665
   tabag |   .0307801   .0146871     2.096   0.036       .0019939    .0595664
   _cons |  -3.143189   .4141537    -7.589   0.000      -3.954915   -2.331463
------------------------------------------------------------------------------
```

Note that in this case, the p value of the LR test and that of the *Wald test* of the covariate are equivalent

**G =** -2(L(modèle(1) – L(modèle(2)]= -2(-794.39 + 792.38) = **4.02**

If the variable *tabag* did not bring anything, then G would follow a $\chi^2_{(1)}$ distribution.

The value of 4.02 corresponds to the 96th centile of $\chi^2_{(1)}$:

Hence p=0,04 and it is safer to reject $H_0$.

**One can therefore consider that the variable *tabag* should be kept in the model.**

# The Wald test

The Wald test allows to provide a p-value for each of the model's parameters.

It relies on the statistic

$$Z = \hat{\beta}/S\hat{E}(\beta)$$

**Property:**

Under the null hypothesis

$H_0 : \beta=0$

the square value of Z  follows a $\chi^2$ distribution with 1 degree of freedom.

Z=0.0398/0.0145=2.735.

$Z^2$=7.48, hence p

```
------------------------------------------------------------------------------
   issue |      Coef.   Std. Err.       z     P>|z|      [95% Conf. Interval]
---------+--------------------------------------------------------------------
     agf |   .0397656   .0145416     2.735   0.006       .0112646    .0682665
```

**Model 3:** Let us consider adding maternal alcohol consumption during pregnancy (variable *alcg* in glasses of alcohol/week) :

```
. logit issue agf tabag alcg
Number of obs    =        2172
Log likelihood =  -790.9586
------------------------------------------------------------------------------
   issue |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     agf |   .0391483   .0145099     2.698   0.007       .0107093    .0675873
   tabag |   .0307169   .0147236     2.086   0.037       .0018592    .0595746
    alcg |   .0390405   .0220661     1.769   0.077      -.0042082    .0822893
   _cons |  -3.176997   .4139216    -7.675   0.000      -3.988268   -2.365725
------------------------------------------------------------------------------
```

**G’** = -2[L(model(2) – L(model(3)]= -2(-792.38 + 790.96) = **2.84**

This value corresponds approximately to the 92th centile of a $\chi^2_{(1)}$ distribution.

**Interpretation:**

If $H_0$ were true (equivalently, if $\beta_2=0$) then one would observe such a value of G in about 8 studies out of 100 (the test p-value is about 8%). This means that the improvement in the model fit brought by the introduction of the alcohol variable in the model already including age and tobacco is more limited than the introduction of the tobacco variable in the initial model M1 was.

In this situation, one may question the relevance of the introduction of the variable alcohol in the model.

# LR-test interpretation and decision

Most epidemiologists would probably prefer to keep the variable *alcohol* in the model. Why?

- **If one follows a purely statistical decision framework** to build the model and choose the potential confounders to adjust for, it is considered safer to retain a p-value of about 25% as a cut-off to decide which covariates to include (i.e., a covariate is kept in the model if its p-value is below 25%)
- **If the framework chosen to build the model is based on a priori knowledge**, then, if there is some evidence from previous human or animal studies for an effect of alcohol on abortion risk, then, again, the variable alcohol should be retained (in this case, the LR test is actually useless because one chooses to rely on external information to build the model).

*Next step:* It will then be useful to define the most relevant coding for the alcohol variable.

# **Likelihood-ratio test**: Generalisation

We have considered the situation when the richer model had only 1 parameter more than the simpler model. In this case, the test statistic G is assumed to follow a $\chi^2$ distribution with 1 degree of freedom.

Let us now consider another situation in which model $M_2$ (the *richer* model) has q additional covariates ($X'_1$, $X'_2$,…, $X'_q$) compared to $M_1$.

In this case, the LR test can still be used to compare the impact of the addition of the q covariates altogether in the model.

$$\mathbf{G=-2[L(M_1) - L(M_2)]} \qquad \text{(L = log-likelihood)}$$

The only difference compared to the previous situation is that this time, G needs to be compared with a $\chi^2$ distribution with **q degrees of freedom** (and not 1).

Indeed, under the hypothesis $H_0$

$$H_0 : \beta'_1 = \beta'_2 = \ldots = \beta'_q = 0$$

G follows a distribution $\chi^2_{(q)}$

$$\mathbf{H_0 : G \sim \chi^2_{(q)}}$$

This is important, because in this situation in which one is interested in the global effect of several covariates, the Wald test cannot be used.

# LR test, degrees of freedom: illustration

If one wants to compare the model

$$\text{logit}(P)=\alpha + \beta_1X_1 + \beta_2X_2 \qquad \textbf{(1)}$$

with the model:

$$\text{logit}(P)=\alpha + \beta_1X_1 + \beta_2X_2 + \beta'_2X^2_2 + \beta_3X_3 + \beta_4X_4 \qquad \textbf{(2)}$$

then the test statistic G should be compared to a $\chi^2$ distribution with…

…3 degrees of freedom (because model (2) has 3 more parameters than model (1)).

# The LR test in practice

1) Write down the 2 compared models $M_1$ and $M_2$ and make sure that $M_1$ is nested within $M_2$. Write down the alternative hypotheses $H_0$ and $H_1$ implied by the comparison of $M_1$ and $M_2$ and calculate the difference in the number of *degrees of freedom* **q** between both models (i.e. the difference in the number of parameters to estimate between the 2 models)

2) Estimate the parameters of models and note their corresponding log-likelihoods $L(M_1)$ and $L(M_2)$.

3) Make sure that the number of subjects "used" to estimate both models is the same (and that these are the same subjects (in practice, some covariates have sometimes more missing values than other, which makes some subjects present in the simpler model disappear in the richer model)

4) Calculate the LR test statistic **$G=-2[L(M_1) – L(M_2)]$** and compare it to the distribution $\chi^2_{(q)}$ (you thus obtain the test's p-value)

**6) Interpretation:**

The lower the p-value, the higher the probability that the added variables ($X'_1, X'_2,\ldots, X'_q$) present in $M_2$ but absent from $M_1$ improve the model fit.

One can use this p-value to decide whether or not these covariates should stay in the model or not.

Of course, the p-value will depend on the covariates chosen.

# VI. Effect measure modification

# Allowing effect measure modification between covariates

- One speaks of effect measure modification when the apparent effect of a variable $X_1$ is different in subgroups defined by another covariate

  For example, the effect of smoking on cardiovascular diseases may be stronger (or weaker) in men than in women

  (this is sometimes called interaction, but this term has many different meanings and should be avoided)

- By default, the regression model

$$\text{logit}(E(Y))=\alpha+\beta_1 X_1+\beta_2 X_2$$

will not be able to tell you that the effect measure of X varies with $X_2$ if you do not ask him…

# A more complex model

- A solution is to add another variable in the model that simultaneously depends on $X_1$ and $X_2$.

  *For example*, the product of $X_1$ and $X_2$ can be used:

$$\text{logit}(E(Y))=\alpha+\beta_1 X_1 + \beta_2 X_2 + \gamma .\, X_{1.} X_2 \qquad (1)$$

**Just like for linear regression, the model**

$$\text{logit}(E(Y))=\alpha+\beta_1 X_1 + \beta_2 X_2 + \gamma .\ X_{1.} X_2$$

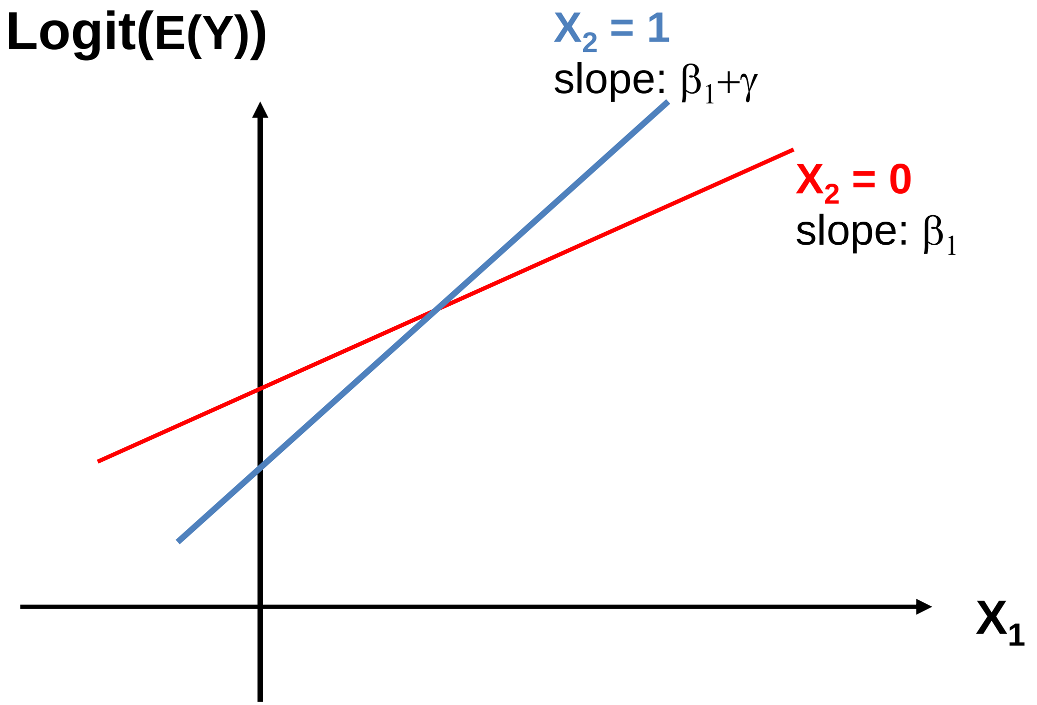
**allows us to obtain:**

- an estimate of the effect of $X_1$ in subjects not exposed to $X_2$: $OR(X_1/X_2=0)=\exp(\beta_1)$
- an estimate of the effect of $X_1$ in subjects not exposed to $X_2$ : $OR(X_1/X_2=0)=\exp(\beta_1)$
- an estimate of the effect of $X_1$ in subjects exposed to $X_2$ : $OR(X_1/X_2=1)=\exp(\beta_1+\gamma)$

In other words, we now allowed for the effect of $X_1$ to differ according to the value of $X_2$
(or, equivalently, *for the effect measure of $X_1$ to be modified by $X_2$*)

# *Effect measure modification*

## Plot of the predicted values with one continuous covariate ($X_1$) and a binary covariate ($X_2$)



The estimated effect of X is different for subjects with $X_2$=0 and for those with $X_2$=1.

# What if $X_1$ and $X_2$ are both binary variables?

- The same model can be used

$$\text{logit}(E(Y))=\alpha+\beta_1X_1 + \beta_2X_2 + \gamma.\ X_{1.}X_2 \qquad (1)$$

*Since* $X_1$ and $X_2$ are both binary variables, we will assume that they are both coded with the values 0 and 1 (this is usually the safest option). In this case, $X_1.X_2$ is an indicator variable with a value of 1 if and only if both $X_1$ and $X_2$ have a value of 1.

Let us write the Odds-Ratios of disease estimated by this model in the 4 categories defined by all the possible combinations of $X_1$ and $X_2$:

Odds-Ratio of disease

| | | $X_1$ | |
|---|---|---|---|
| | | 0 | 1 |
| $X_2$ | 0 | 1 | $\exp(\beta_1)$ |
| | 1 | $\exp(\beta_2)$ | $\exp(\beta_1+\beta_2+\gamma)$ |

# More interesting interactions…

1. If $X_1$ is a continuous variable and $X_2$ is a binary variable

   Simple interaction term (*see above*)

2. If $X_1$ and $X_2$ are binary

   Simple interaction term (*see above*)

3. If $X_1$ is a categorical variable and $X_2$ binary

   It is safer to code $X_1$ with dummy variables (k-1 if $X_1$ has k categories)

   And to add an interaction term with each of these dummy covariates

4. If $X_1$ and $X_2$ are continuous covariates

   A safe option is to transform them in categories, code them by dummy variables, and create interaction terms for all the pairwise combinations of the dummy covariates. This means (j-1)x(k-1) terms for the interaction if $X_1$ and $X_2$ are coded with j and k categories.

# *Coding effect measure modification: summary*

If $X_1$ and $X_2$ are quantitative covariates, $\beta_1$ and $\beta_2$ the associated parameters, and $\gamma$ the parameter associated with their interaction term:

$$\text{logit}(E(Y)) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \gamma .\ X_{1.} X_2$$

- **exp($\beta_1$)** is the OR of disease associated with **an increase by 1 in $X_1$** among subjects in whom $X_2=0$
- **exp($\beta_1+\gamma$)** is the OR of disease associated with **an increase by 1 in $X_1$** among subjects in whom $X_2=1$
- **exp($\beta_1+x_2\gamma$)** is the OR of disease associated with **an increase by 1 in $X_1$** among subjects in whom $X_2=x_2$
- **exp($\beta_2$)** is the OR of disease associated with **an increase by 1 in $X_2$** among subjects in whom $X_1=0$
- **exp($\beta_2+\gamma$)** is the OR of disease associated with **an increase by 1 in $X_2$** among subjects in whom $X_1=1$
- **exp($\beta_2+x_1\gamma$)** is the OR of disease associated with **an increase by 1 in $X_2$** among subjects in whom $X_1=x_1$

# How to test effect measure modification?

Testing if a variable $X_2$ modifies the effect measure of $X_1$ on the disease occurrence can be done by comparing the interaction-free model (1)

$$\text{logit}(E(Y)) = \alpha + \beta_1 X_1 + \beta_2 X_2 \qquad \textbf{(1)}$$

with the model

$$\text{logit}(E(Y)) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \gamma.\ X_{1.} X_2 \qquad \textbf{(2)}$$

Model (1) is nested in model (2) (because if one assumes $\gamma$=0 in (2), (1) is obtained)

Both models thus can be compared with a *likelihood ratio test*.

If the log-likelihood of models (1) and (2) is written $LL_1$ and $LL_2$, respectively:

$$\mathbf{G} = 2(LL_1 - LL_2)$$

follows a $\chi^2$ (1) distribution under the null hypothesis of a lack of modification of the effect measure of $X_2$ by $X_1$

This test can thus be seen as an “interaction” test and the corresponding p-value be used to decide if the interaction term is relevant or not

# Example: Effect of smoking on ovarian cyst occurrence according to Body Mass Index

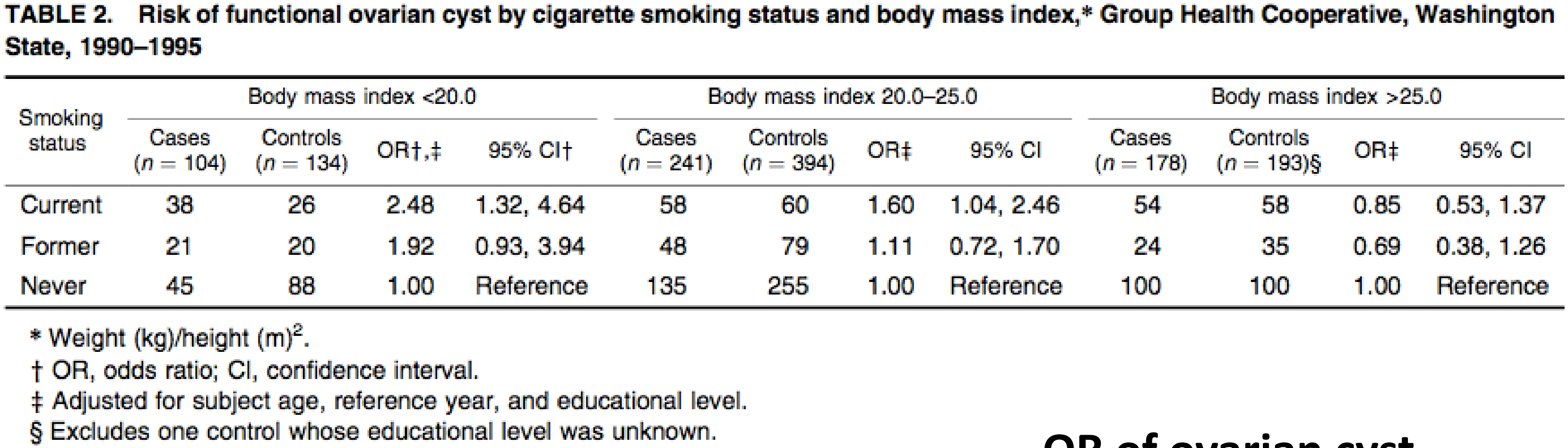


**TABLE 2. Risk of functional ovarian cyst by cigarette smoking status and body mass index,* Group Health Cooperative, Washington State, 1990–1995**

| Smoking status | Body mass index <20.0 | | | | Body mass index 20.0–25.0 | | | | Body mass index >25.0 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Cases (*n* = 104) | Controls (*n* = 134) | OR†,‡ | 95% CI† | Cases (*n* = 241) | Controls (*n* = 394) | OR‡ | 95% CI | Cases (*n* = 178) | Controls (*n* = 193)§ | OR‡ | 95% CI |
| Current | 38 | 26 | 2.48 | 1.32, 4.64 | 58 | 60 | 1.60 | 1.04, 2.46 | 54 | 58 | 0.85 | 0.53, 1.37 |
| Former | 21 | 20 | 1.92 | 0.93, 3.94 | 48 | 79 | 1.11 | 0.72, 1.70 | 24 | 35 | 0.69 | 0.38, 1.26 |
| Never | 45 | 88 | 1.00 | Reference | 135 | 255 | 1.00 | Reference | 100 | 100 | 1.00 | Reference |

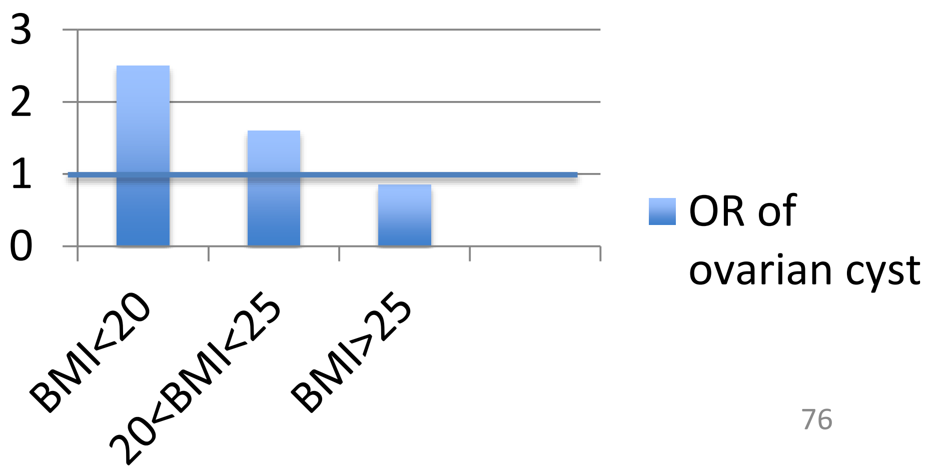
* Weight (kg)/height (m)$^2$.

† OR, odds ratio; CI, confidence interval.

‡ Adjusted for subject age, reference year, and educational level.

§ Excludes one control whose educational level was unknown.

**OR of ovarian cyst**



(Holt, *Am J Epidemiol*, 2005)

# VII. Summary and perspectives

# Comparing *Linear* and *logistic* regression

Y is a continuous variable

$E(Y)=\alpha+\beta_1X_1 + \beta_2X_2$

Y is binary

$\text{logit}(E(Y))=\alpha+\beta_1X_1 + \beta_2X_2$

## Estimation of model parameters

Least square method

Maximum likelihood method

## Interpretation of $\beta$

Association between X and E(Y)

$\beta_1$= increase in E(Y) when $X_1$ increases by 1

Association between X and Pr(Y=1)

$\beta_1$ is the log of the OR associated with an increase by 1 in $X_1$.

## Interaction terms

$\gamma.\ X_{1.}X_2$

$\gamma.\ X_{1.}X_2$

## Between-model comparisons

F Test

Likelihood-Ratio (LR) test.

BIC criterion

# Extensions of the logistic regression model

- In spite of its great robustness and flexibility, the logistic model has limitations. Two of these limitations relate to the type of explained variable that can be handled
- What to do if Y is not a binary variable but a categorical variable with 3 or more categories?
- What to do if Y is a binary variable (indicating for example the risk of occurrence of a disease) assessed among subjects who have been followed-up prospectively, with varying durations of follow-up (and possibly some subjects lost to follow-up)?

# Extensions of the logistic regression model

- In spite of its great robustness and flexibility, the logistic model has limitations. Two of these limitations relate to the type of explained variable that can be handled
- What to do if Y is not a binary variable but a categorical variable with 3 or more categories?

You could group some categories so as to end up with a new variable Y' with 2 categories (and use logistic regression)

Alternatively, **polytomic regression** is an approach allowing to handle such categorical outcomes (not detailed here)

- What to do if Y is a binary variable (indicating for example the risk of occurrence of a disease) assessed among subjects who have been followed-up prospectively, with varying durations of follow-up (and possibly some subjects lost to follow-up)?

Survival models (e.g. Cox, Accelerated Failure Time, Weibul… models) are the right option

# Perspectives

- In the future lectures we will
  - Further think about the coding of covariates
  - Discuss the impact of exposure misclassification (misclassification in the covariates)
  - Come back on the interpretation of the models estimates (parameters, p-values…)

# Important properties of a model

- Internal
  - “Aim”
  - Goodness of fit
  - Parsimony
- External
  - Robustness
  - Plausibility (e.g., with biological knowledge)

# Thank you for your attention