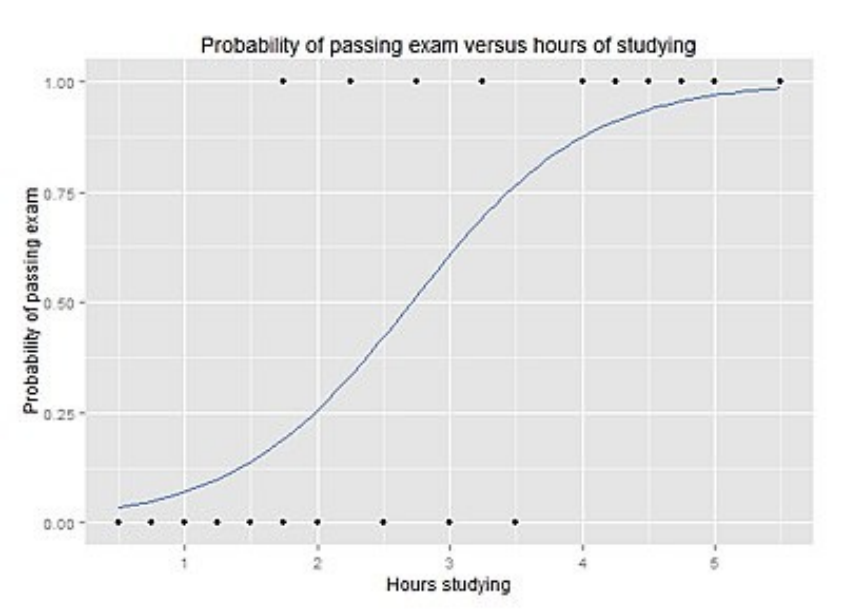
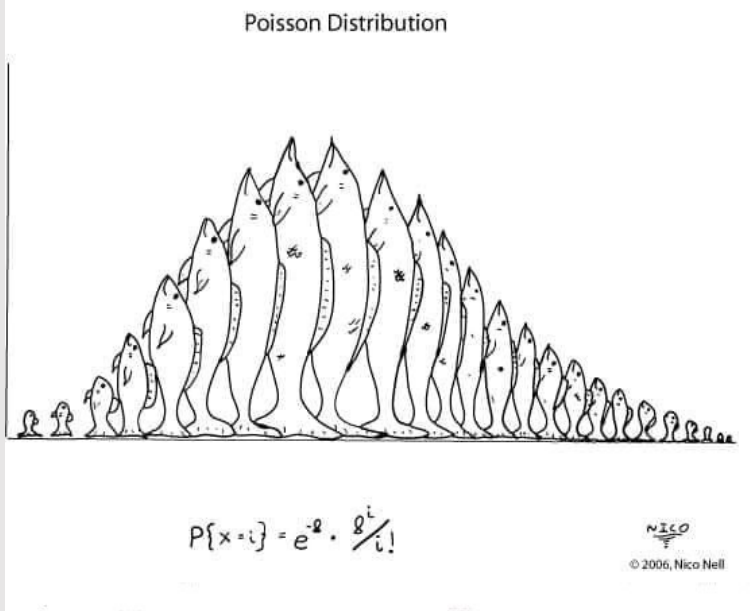
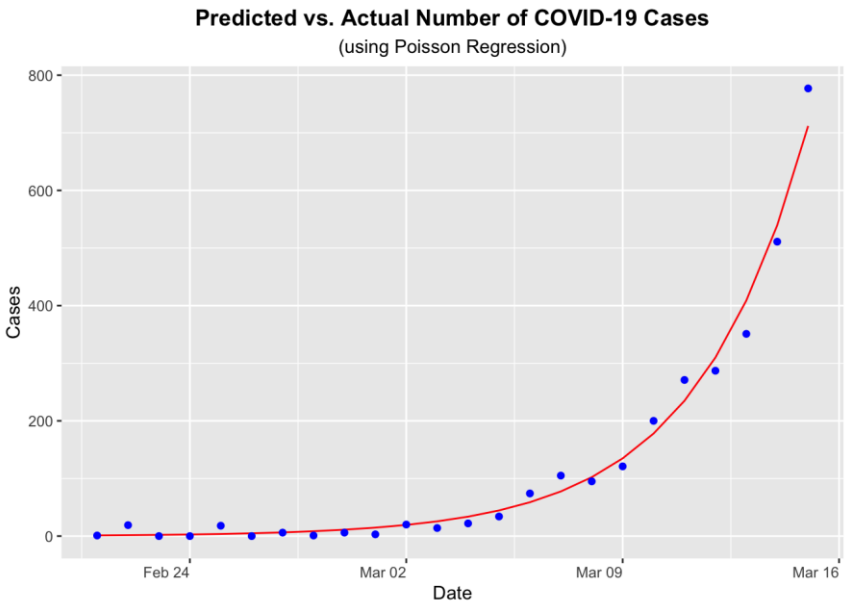
Regression models
(& their link to the study design)



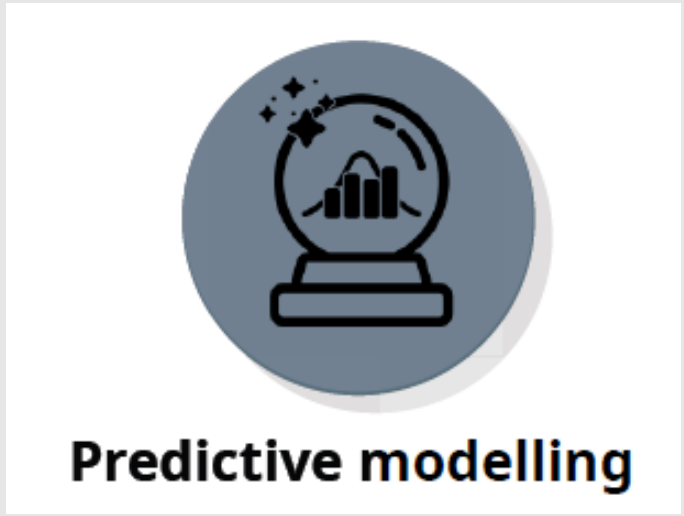




Different scientific aims

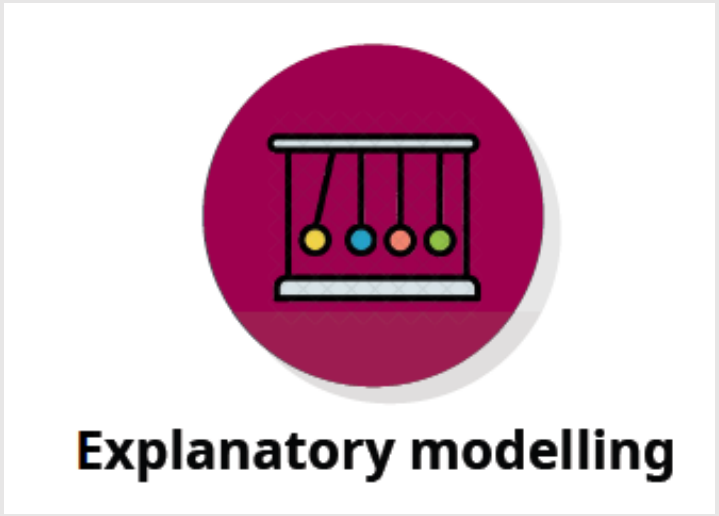
Describe the outcome of interest:
which factors *affect* it and how? (**association**)

Estimate a *prevalence* in function of age and sex



Accurate **predictions** of future observations.
No concern about causality and confounding (association)

Risk of developing CVD in the next x years



Testing and comparing existing **causal** theories.

Effect of LDL on CVD risk

Policy recommendations (statins)

(there is some overlap but a key is to define the PRIMARY goal)

Independently from the primary scientific aim, a **basic statistical tool** is the REGRESSION model approach.

The key difference is in **the scope of** the REGRESSION model according to the main aim...

The **Descriptive** Aim (The "What")

Regression as a tool for **parsimonious** summarization.

Describing the Population: How regression describes the *average* individual.

Trend Identification: Using splines or polynomials to describe non-linear patterns

Data Reduction: Using regression to simplify complex datasets into understandable trends.

The Predictive Aim (The "Who")

Focus on accuracy and generalizability of predictions.

The **Bias-Variance** Tradeoff: Why "more variables" isn't always better.

Feature Selection: we don't care about the specific role of a variable (if it is a "confounder" or "independent predictor"..) only if *it adds signal*.

Validation: The necessity of Cross-Validation and External Validation.

The Causal Aim (The "Why")

Focus on unbiased estimation of an **effect** of an intervention/exposure.

Counterfactual Framework: Potential Outcomes.

Directed Acyclic Graphs (DAGs): How to choose variables based on theory, not p-values...

Identifiability: assumptions of exchangeability, positivity, and consistency.

Type of Outcome Data
and Choice of Model (*classical* ones)

Type of outcomes	**Possible Model**
• **Continuous** measurement	• **Linear** regression (normal outcomes)
• **Count** data	• **Poisson** regression
• **Binary** data	• **Logistic** regression
• Ordered categorical	• *Proportional odds*
• **Censored lifetimes**	• **Proportional hazards Cox** regression
• Multinomial	• *Log-linear*

What models do we *typically* see in health research?

Continuous outcome, linear regression model: Y=Heart rate

$$E(Y_i) = \alpha + \beta_1 Age_i + \beta_2 Sex_i + \cdots + \beta_n Hypertension_i$$

LP_i Linear Predictor

Binary outcome, logistic regression model: p =Probability of CV hospitalization

$$logit(\hat{p_i}) = \alpha + \beta_1 Age_i + \beta_2 Sex_i + \cdots + \beta_n Hypertension_i$$

LP_i Linear Predictor

Time-to-event outcome, Cox regression model: h = «hazard» at time t of CV hospitalization

$$h_i(t) = h_0(t) exp(\beta_1 Age_i + \beta_2 Sex_i + \cdots + \beta_n Hypertension_i)$$

LP_i Linear Predictor

How do we (typically) estimate the ***coefficients*** of models?

Likelihood: probability of data **given** the model, interpreted as function of model parameters

$$L(\beta|X,Y)$$

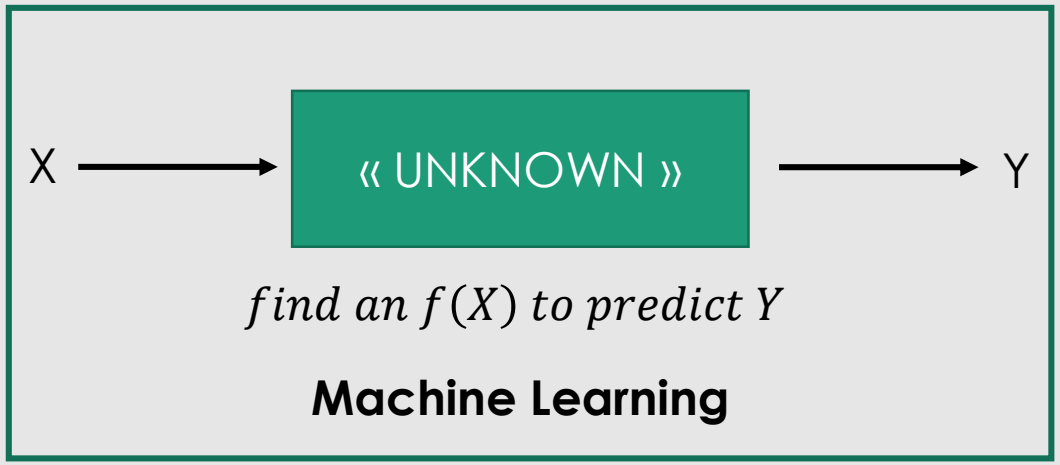
Ronald A.Fisher in 1913

Fisher (aged 22):

- **Maximum likelihood principle**:

$$find\ \beta\ such\ that\ L(\beta|X,Y)\ \rightarrow max$$

Assume a ***stochastic*** data model for the inside of the box:



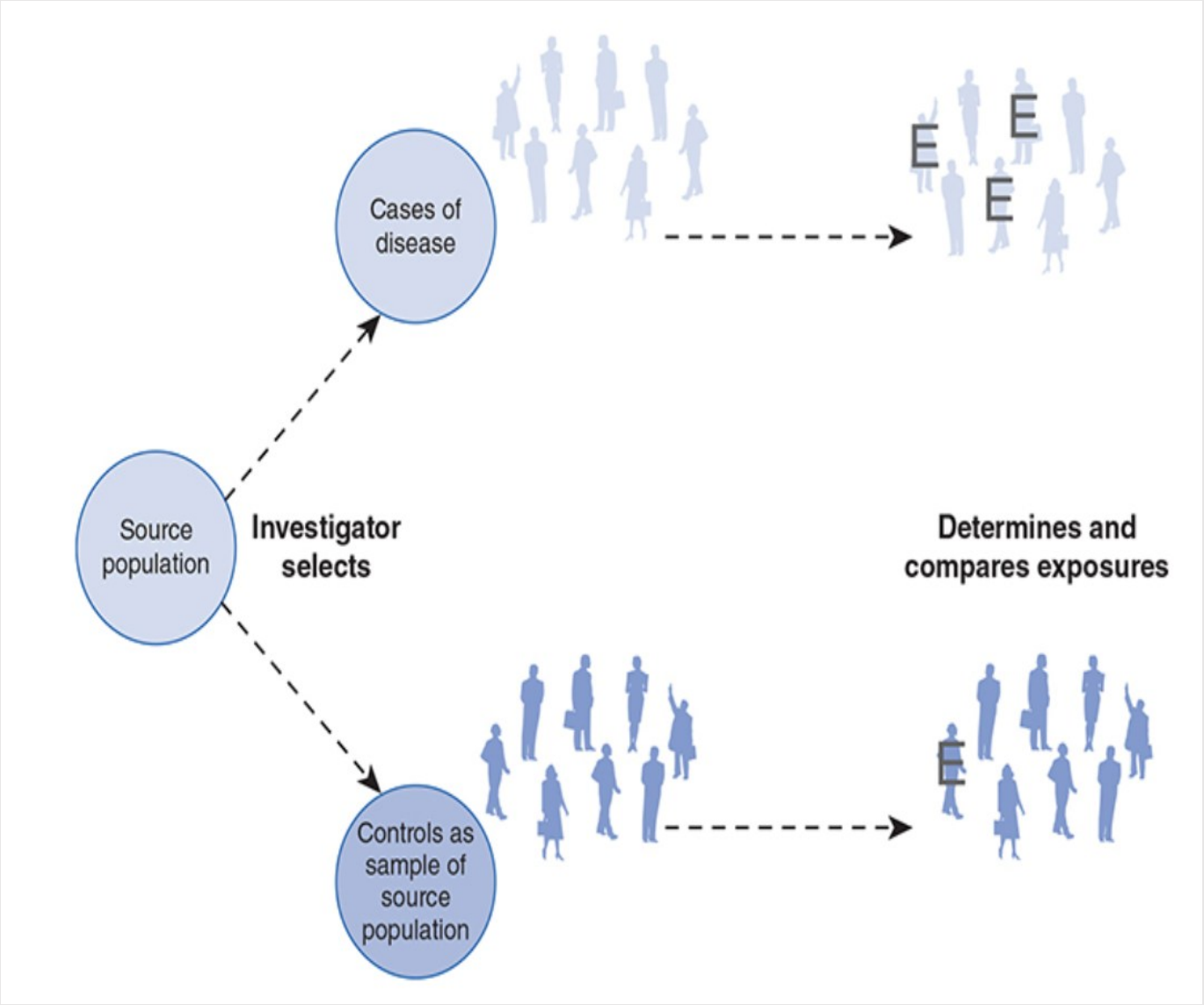
The **logistic regression model is particularly useful** to evaluate data obtained from a **case-control** study design.

The purpose of this design is to assess the **magnitude of the association** between an exposure and a specific disease or health-related event.

This is the most **cost-effective** study design and is recommended when the **incidence** of the disease or condition of interest is **rare** or has a **long latency.**

The aim could be:

- **describing** associations
- making ***predictions***, for example *prediction of diagnosis* could be a target
- establishing a **causal** effect for one exposure of interest.



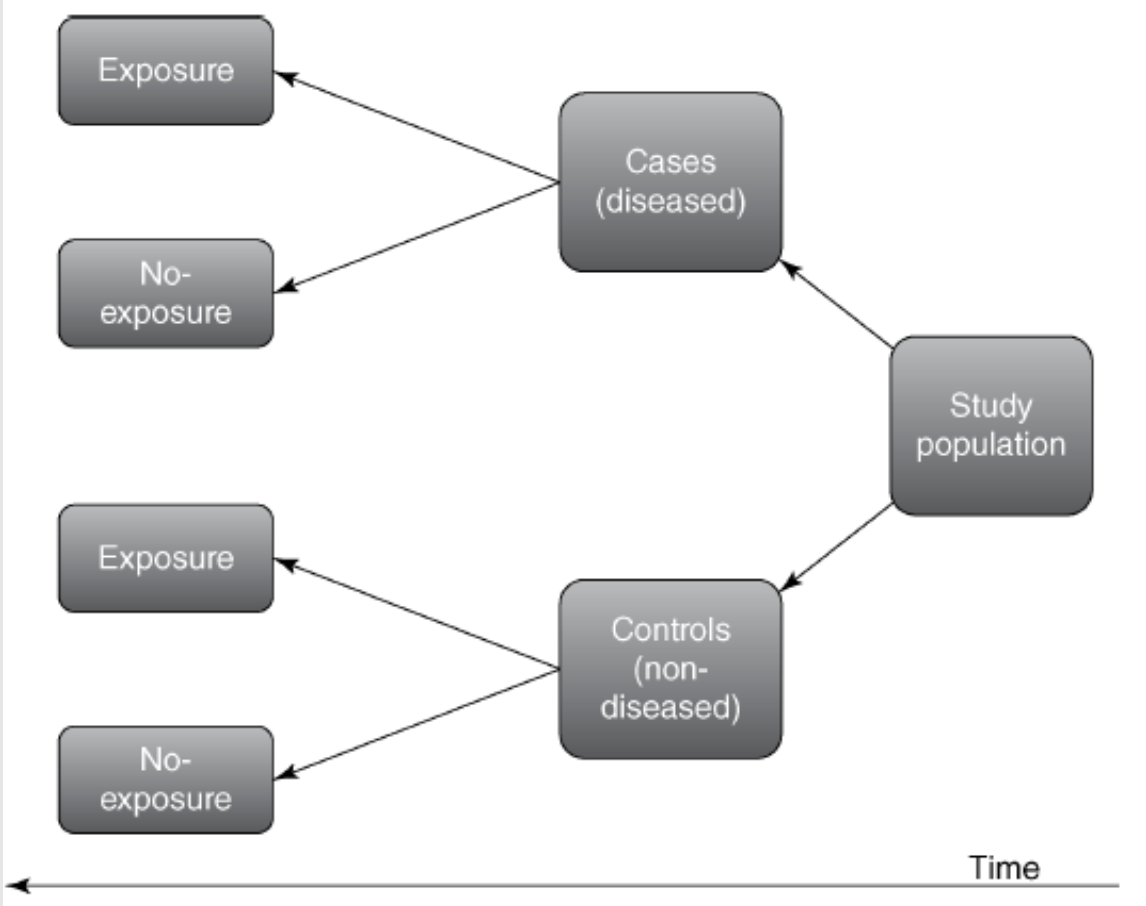
Specific Objectives (related to **descriptive** purposes):

- Evaluate/interpreting **associations** of multiple factors/variables to the probability of the outcome of interest

- Assess possible ***effect modification [interaction]***

Remember: in a standard case-control study design, it is not possible to estimate **disease incidence** in those who are exposed and those who are unexposed, since participants are selected according to the disease status, not on the basis of their exposure status.

However, it is possible to calculate the **odds of exposure** among cases and controls, and then the exposure/disease **odds ratio**.



Estimating Odds Ratios via Logistic Regression

The logistic regression model allows us to estimate the OR to assess the magnitude of the association between a specific factor and the disease under study taking into account **multiple** covariates **with possibly different scales** of measurements.

LR estimates* the probability of disease in the exposed and unexposed groups as follows:

$$OR = \frac{P_{exposed}/(1 - P_{exposed})}{P_{unexposed}/(1 - P_{unexposed})}$$

p_i

probability of having the disease for the subject *i*

$$p_i = \frac{1}{1 + e^{-\left(\beta_0 + \sum_{j=1}^{J} \beta_j X_{ij}\right)}}$$

Factors «X» in whatever scale of measure: binary, numerical, categorical..

*estimate β by *maximising the likelihood*, i.e. probabilities to observe the data in hand get maximal.

The probability of observing a control (non diseased person) through the LR model is:

$$1 - p_i = 1 - \frac{1}{1 + e^{-\left(\beta_0 + \sum_{j=1}^{J} \beta_j X_{ij}\right)}}$$

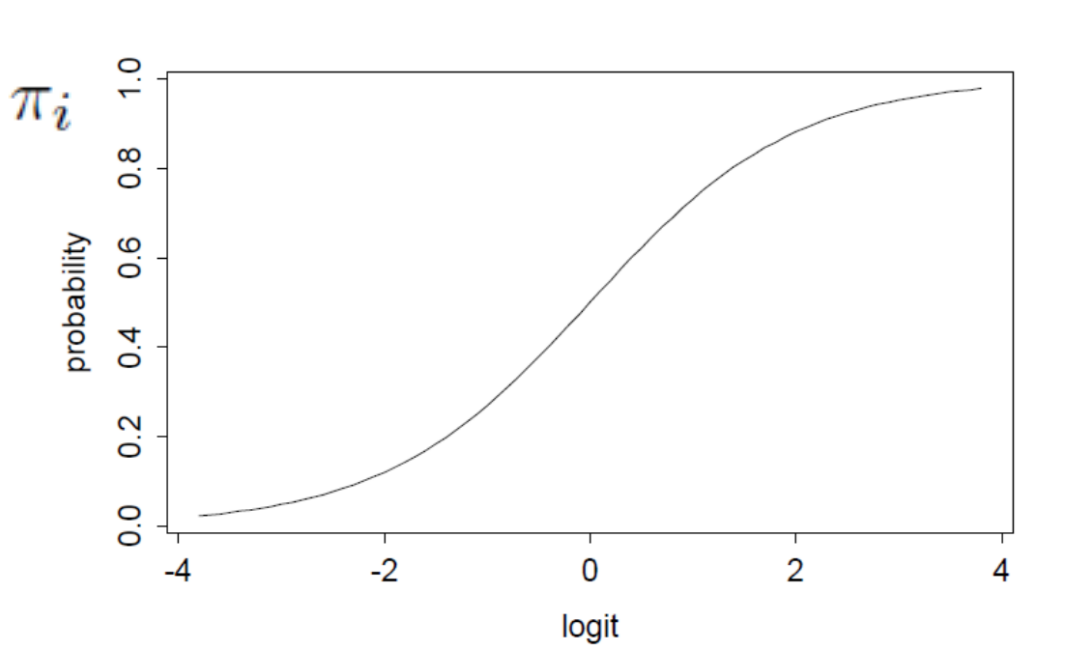
As a result, the odds of disease can be defined by:

$$\frac{p_i}{1 - p_i} = e^{\left(\beta_0 + \sum_{j=1}^{J} \beta_j X_{ij}\right)}$$

On a logarithmic scale, the odds of disease would be:

$$log\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \sum_{j=1}^{J} \beta_j X_{ij}$$

Logit function [**link** function]



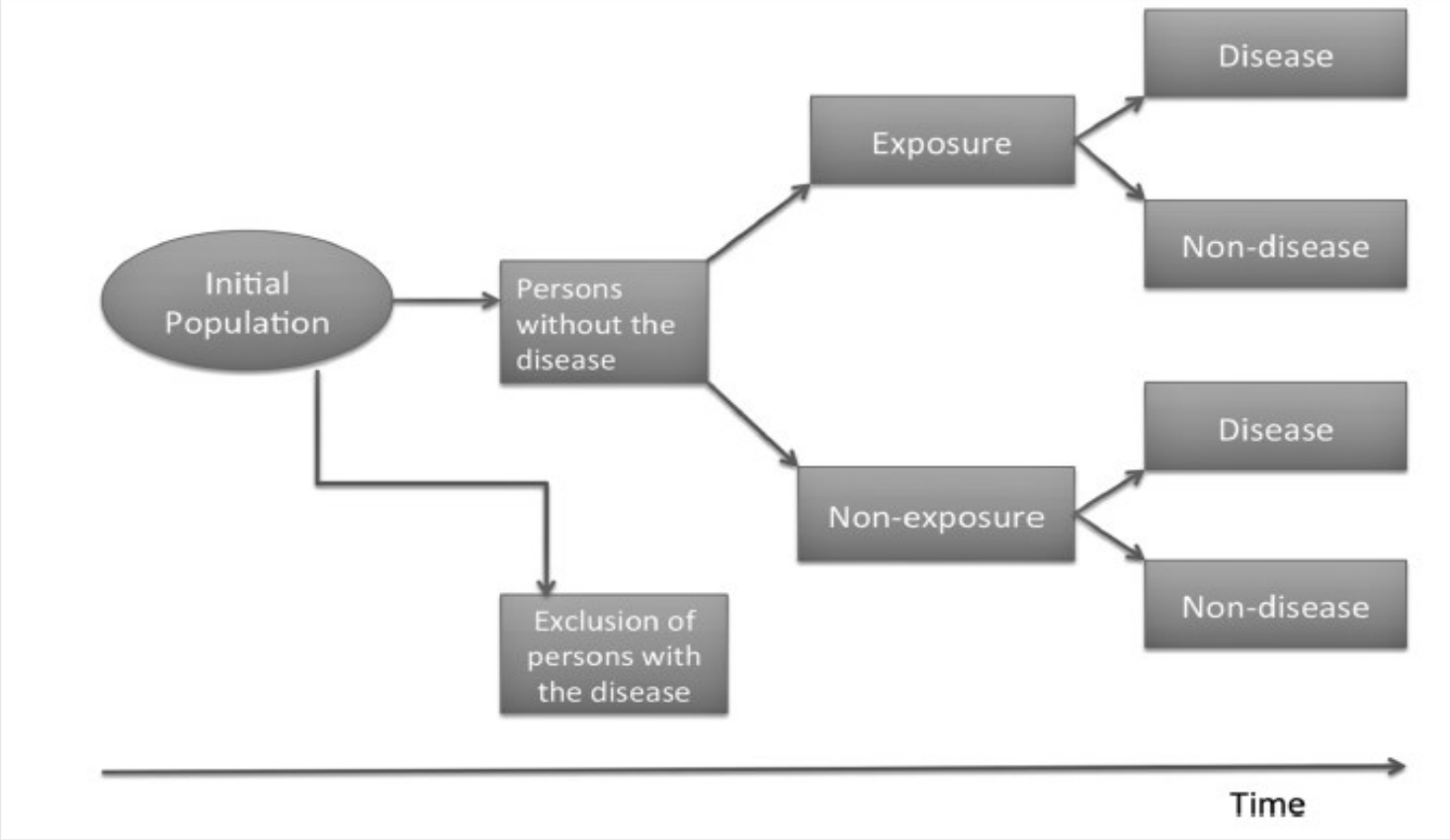
$$\eta_i = \text{logit}(\pi_i) = \log \frac{\pi_i}{1 - \pi_i}$$

On the logit scale we come back to a **linear** model

Remind: Population based / Cohort Studies / [RCTs]

In a population based/cohort/[RCT] study, a population is selected on the basis of their exposure/treatment and **followed over a period of time** to determine the occurrence of a disease or any other health-related event.

The **incidence** of disease could then compared in exposed and unexposed groups.



In the **descriptive** setting the main goal is to estimate **disease incidence among exposed** and **unexposed** individuals and then quantifying the **magnitude** of the association between exposure and disease.

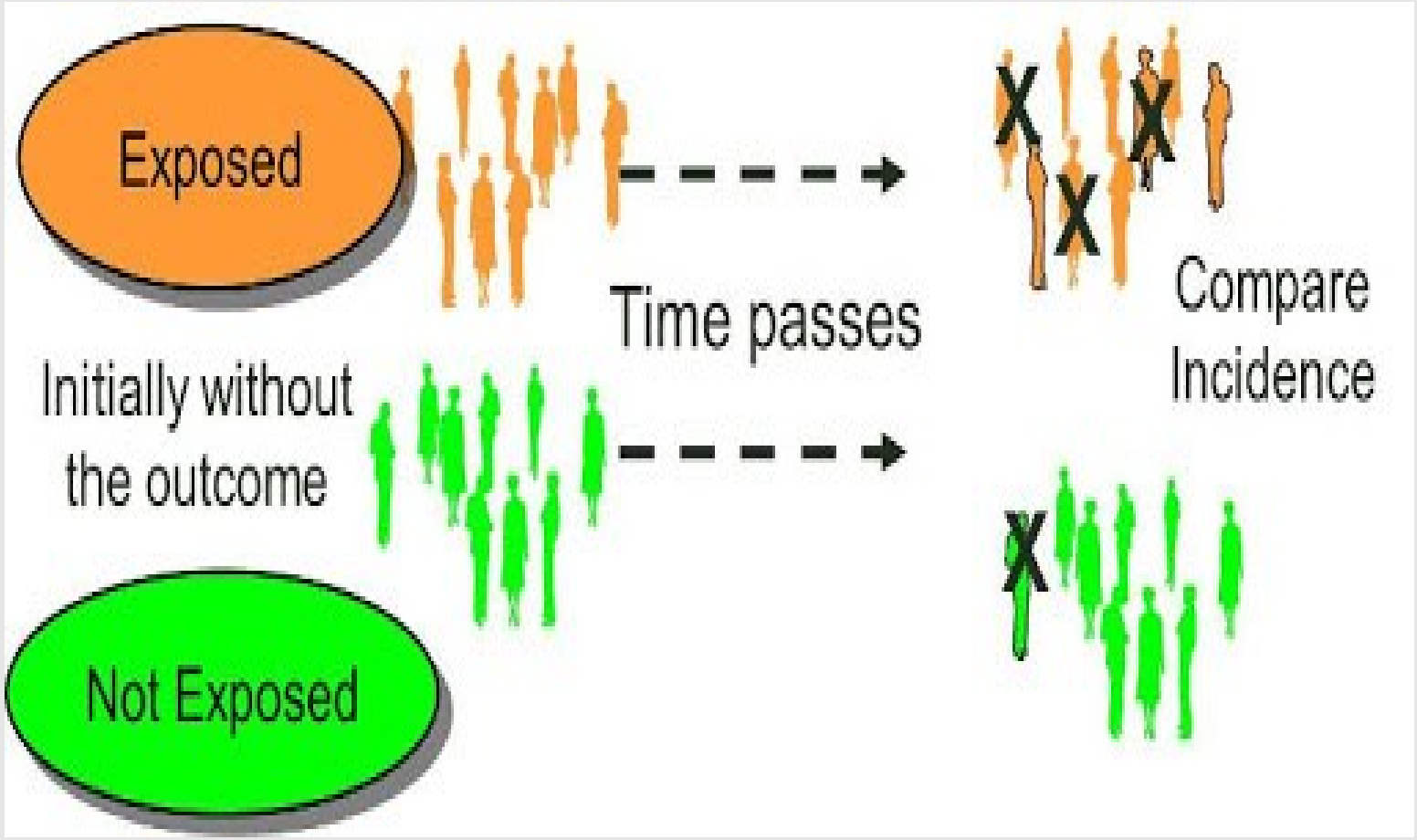
In the prognosis/prediction setting the main goal is to estimate **the probability to develop the outcome in a certain time interval, according to subject's characteristics.**

Advantages of the pop.based/cohort study/[RCTs]:

- **Temporal sequence** between the exposure under study and the disease or any other health-related event can be established
- Determination of disease incidence in each exposure group and investigation of the **association** of the exposure/s of interest on possible ***multiple* outcomes** (not all a-priori defined)

Disadvantages

- Susceptibility to (differential) **loss of subjects** during follow up, which may introduce **selection bias** and thus affect the internal validity of the study
- **Comparability** of subjects who remain in the study and those who are lost, **by exposure status**, must be determined in order to assess potential **confounders/bias**
- **Expensive** (time/costs) [if **primary** data collection]
- Not suitable for **rare** diseases



A regression model for counts/rates

The Poisson regression model estimates the incidence of a disease or health-related event ***under different conditions***.

To determine the incidence, it is necessary to compute the **number of new cases** during the **observation period** and identify the initial conditions of the study, such as the **type of exposure/s** at baseline.

The Poisson model ***establishes a relationship***/*[makes a prediction]* between the **expected** number of cases and the covariates included in the model.

Recap:
The Poisson probability distribution can be used when the random variable represents the number of cases (successes) under **3** conditions:

- in a very large number of independent ***Bernoulli*** trials [when the constant probability of success is small]
- for a **unit of time** (e.g., day, month, or year)
- on a unit area (e.g., square meter, square kilometer, or square mile) …

The Poisson regression model can be written as:

$$\mu_i = P_i e^{\beta_0 + \sum_{j=1}^{J} \beta_j X_{ij}}$$

μ_i **expected value of new cases** in *condition i* : a combination of the values of the covariates. [We assume that the number of new cases is a RV that has a Poisson distribution]

P_i **population** in the *i-th group* of exposure

- Person-time units → Incidence rate
- Population at baseline → Cumulative incidence

X_{ij} *j-th covariates*

β_0 Exp(β_0) : expected incidence of the number of new cases when the X variables take the value of zero.

We are **assuming** here that the response variable is a count of events ***occurring independently*** among different subgroups [number of newly diagnosed cases of kidney cancer at different hospitals every year] and that this random variable follows a Poisson distribution.

We are **assuming** that μ is linked to the **exponential** of a linear function of the candidate associated factors; so the changes in the incidence resulting from the combined effects of factors are multiplicative.

[incidence of events]

$$\frac{\mu_i}{P_i} = e^{\beta_0 + \sum_{j=1}^{J} \beta_j X_{ij}}$$

$$ln(\mu_i) = ln(P_i) + \beta_0 + \sum_{j=1}^{J} \beta_j X_{ij}$$

Since the model contains the variable $ln(P_i)$ there is no need to estimate the coefficient for this variable, referred to as an **offset**

When we have a binary variable E:

$$I_1 = \frac{\mu_1}{P_1} = e^{\beta_0+\beta_E}$$ Incidence in the exposed

$$I_0 = \frac{\mu_0}{P_0} = e^{\beta_0}$$ Incidence in the unexposed

$$\Rightarrow \quad RR = \frac{I_1}{I_0} = e^{\beta_E}$$

When we have a continuous variable X:

Let's say we want to estimate the impact of air temperature on the expected number of hospital admissions for respiratory issues.

P_i population size of the city (e.g., 500,000 people).

$$\frac{\mu_i}{P_i} = e^{\beta_0+\beta_1 X_i}$$

$$I_1 = e^{\beta_0+\beta_1(X_i+1)} \qquad I_0 = e^{\beta_0+\beta_1 X_i} \quad \Rightarrow \quad RR = \frac{I_1}{I_0} = e^{\beta_1}$$

rate of admissions changes for every 1 degree of increase in temperature.

There are **two** important assumptions for Poisson regression:

- Risk is ***homogeneous*** among person-[times] contributed by different subjects who have the same characteristics of interest (e.g. sex, age-group…) and the same period.

- Asymptotically, or as the sample size becomes larger and larger, the *mean* of the counts is equal to the *variance*.

Note here that the linear regression model (assuming constant variance & normal errors) is not appropriate for count data for **3** main reasons:

1. the model might lead to the prediction of negative counts

2. the variance of the response may increase with the mean

3. the errors will not be normally distributed

Poisson regression take into account a *crucial* issue not faced by other regression techniques (linear/logistic).

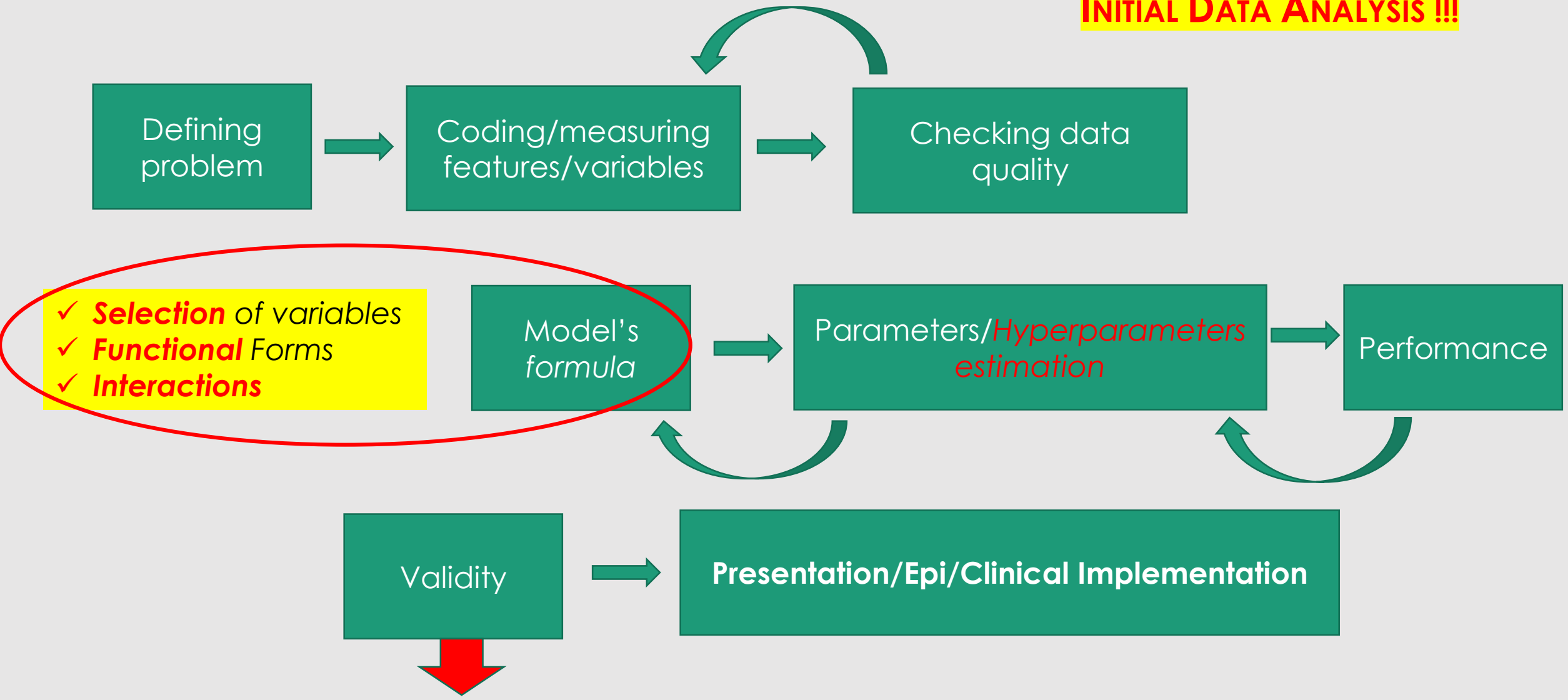
From the data design, different subjects may have **different person-times** of exposure.

Analysing risk factors while **ignoring differences** in person-times is wrong.

Note that logistic regression **completely ignores** this aspect [**difference: cohort/exp. based vs case-control**]. Observation time **is not accounted for** in the evaluation of the probability of the event.

Basic steps:

INITIAL DATA ANALYSIS !!!



*On **external** dataset !!!*

Defining the framework: initial checklist

- **Target population** who would be eligible to use the model and whatever inclusion/exclusion criteria
- **Time origin** baseline *time zero* (***if time is involved***)
- **Outcome (scale of measure)**
- ***Competing risks*** *: events after which the event of interest cannot occur or is not of interest any longer* ***[survival setting]***
- **Follow up time definition** (***if time is involved***)
- **Associated variables/exposures** list of the predictors/features [*measured at baseline*] *(how they were measured…)*

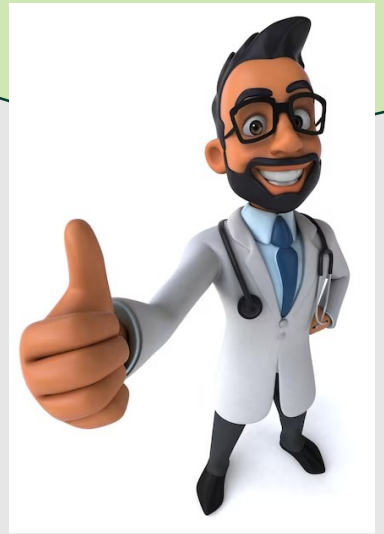
How we **select** variables in the model ?

Depend on the scientific aim !

Descriptive models

Theory-driven

Causal models



Data-driven

Predictive modelling

3 basic steps in the [*classical*] model building process:

1. ***Missing data***

2. Variables **selection (functional forms/interactions)**

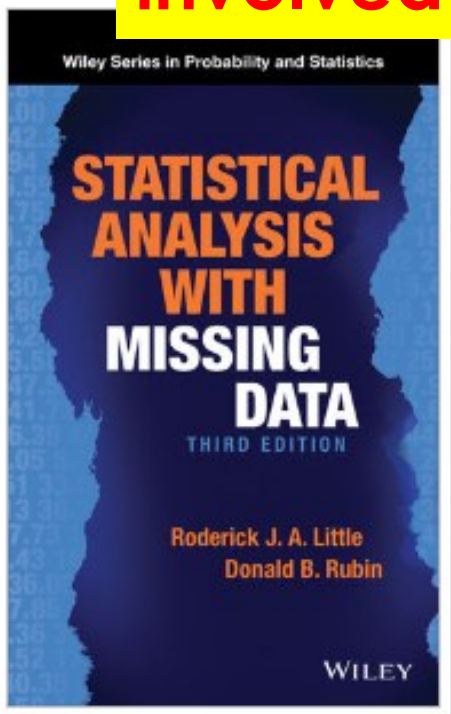
3. Evaluating performance

Missing data…pay attention …

MCAR
Missing completely at random

the fact that data are missing is ***independent*** of the observed and *unobserved* data

no **systematic** differences between participants with missing data and those with **complete** data



Particulary delicate issue in prediction, usually for descriptive aim (and *causal*) we have a *limited* number of variables involved (« low» missing rate) …

the fact that the data are missing is **systematically** related to the observed but not the *unobserved* data

Complete case analyses may or may not result in bias. Proper **accounting** for the known factors can produce unbiased results in analysis

MNAR
Missing not at random

e data are missing is **systematically** related to the unobserved data…

if the complete case analysis is biased this issue **cannot be** addressed…

Some variable selection methods…

Basic* algorithms:

- Full model
- Univariable filtering
- *Forward selection*
- *Backward elimination*

Stepwise methods

Other* ones (predictive aim):

- Information criteria (AIC/BIC)
- LASSO penalization
- ….

…

*There is no *Universal* Solution !!!

Multiverse of models …

Full Model:

1. Do not perform *any* variable selection [except for highly correlated features]

2. Select for each variable a suitable ***functional form***

3. Explore ***biologically plausible*** interactions

[**The initial list is usually *pre-selected* by expertise !!**]

If **sample size** permits…

Univariable filtering:

Still by far the most often applied method in medical literature

1. Select a statistical significance level (e.g., α=0.20 or α=0.10)

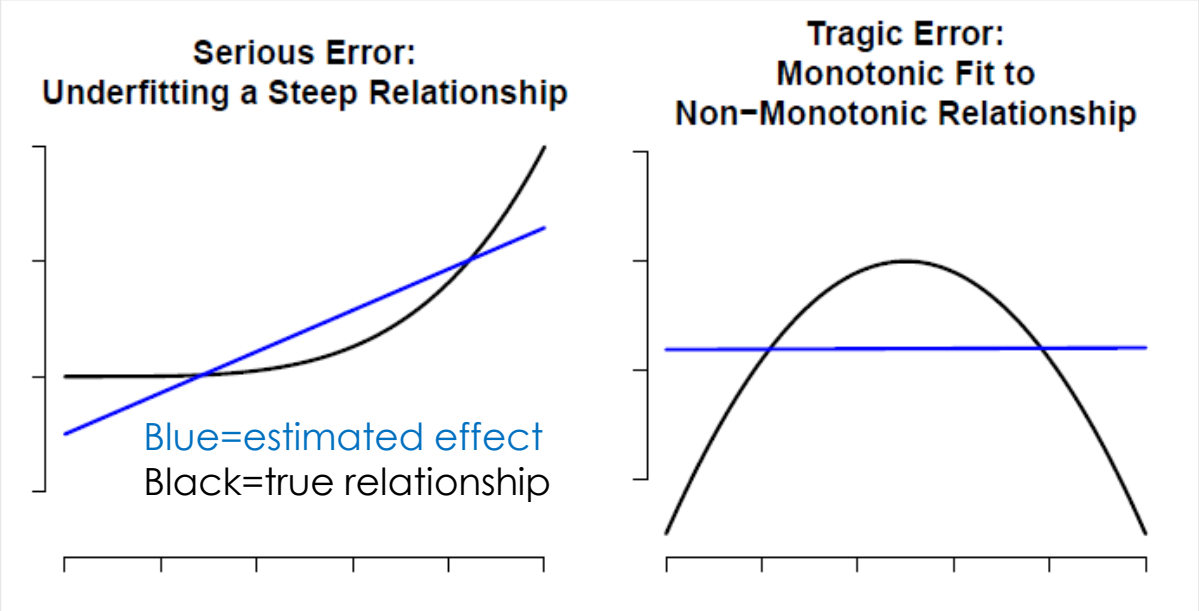
2. Estimate univariable models

3. Use all variables in multivariable model with univariable p-value < α

Univariable selection could work only with *perfectly* uncorrelated variables….

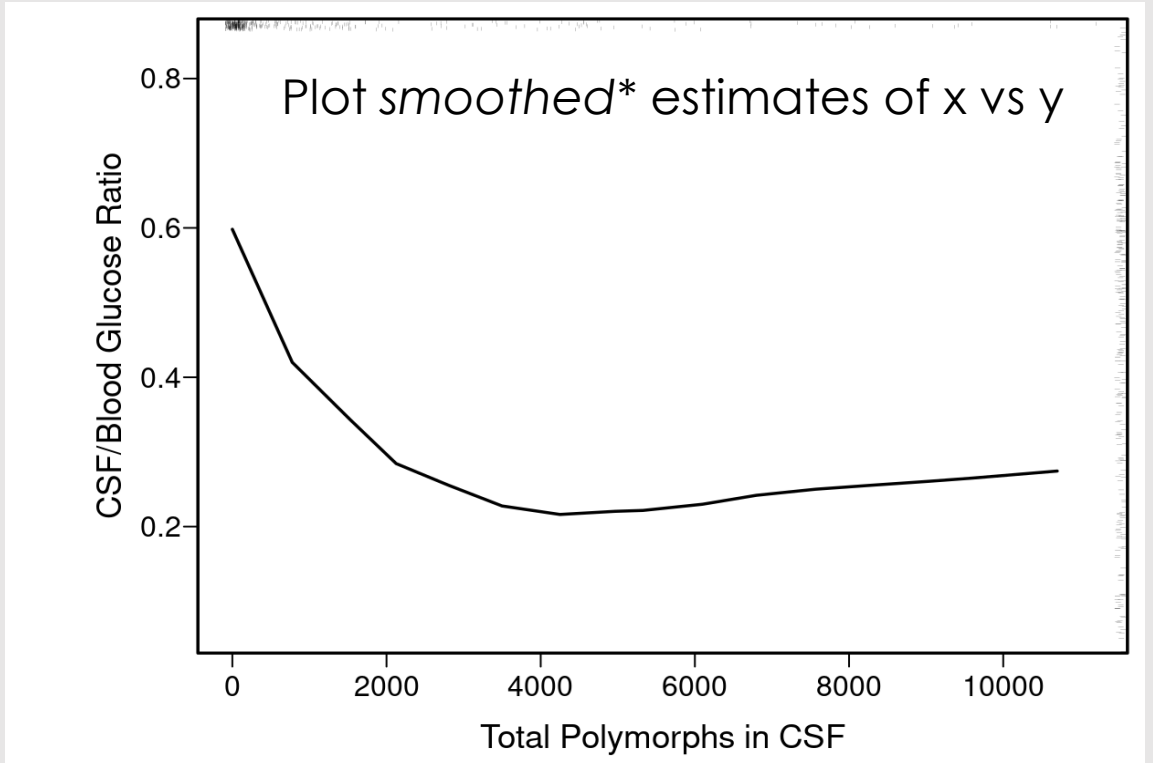
Functional forms

Numerical variables:



Rarely expect linearity.

Visualization tools could help



*non-parametric regression tool

Nominal variables:
- choose an ***appropriate reference*** (frequent, standard group, etc.)
- collapse *rare* groups if possible

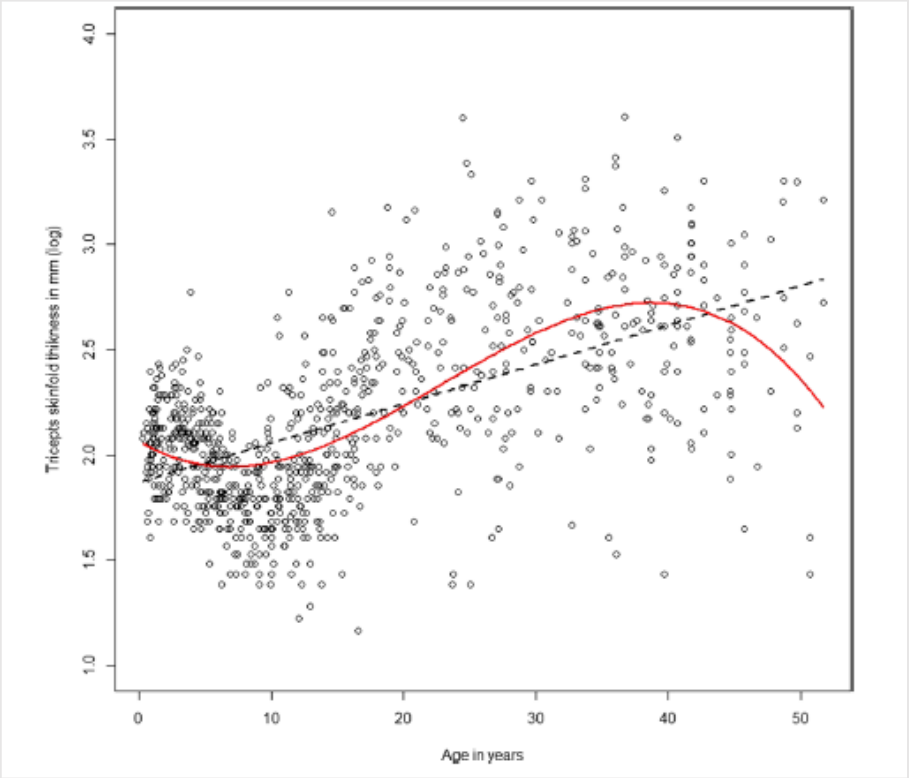
Ordinal variables:
- ordinal coding
- collapse *rare* adjacent groups if reasonable

All classical regression models **should** make assumptions about the **shape** of the relationship between predictor X and response variable Y.

Many analysts assume **linear** relationships *by default*.

Splines (piecewise polynomials) are **nonlinear** generalizations.

[Many practitioners analyze continuous data using **percentiling/classes**, but this is nearly always a bad idea].



REVIEW Open Access

A review of spline function procedures in R

Aris Perperoglou[1*], Willi Sauerbrei[2], Michal Abrahamowicz[3], Matthias Schmid[4] on behalf of TG2 of the STRATOS initiative

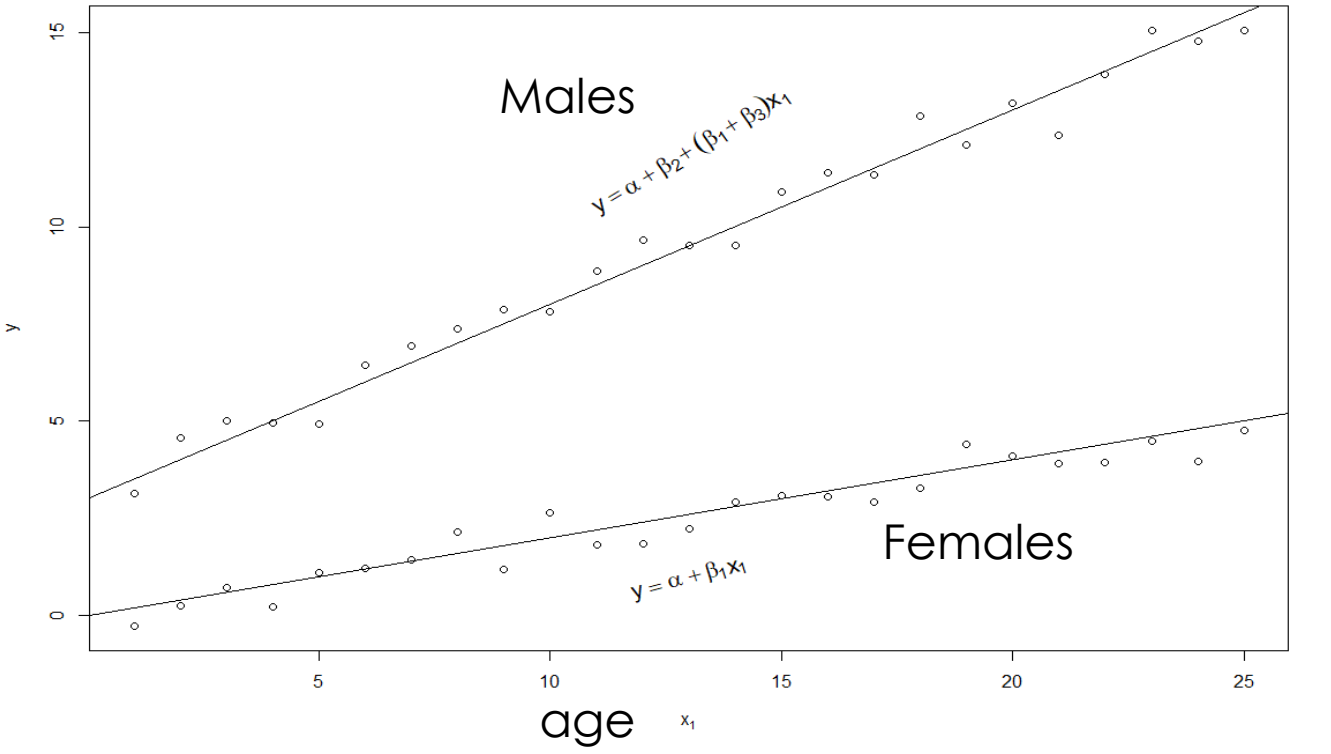
Check for updates

Interactions (basic example)

This is how one allows the slope (*effect*) of a predictor ***to vary*** by categories of another variable.

Example: separate slope for males and females:

$$E(y|x) = \alpha + \beta_1 * age + \beta_2 * [sex = m] + \beta_3 * age * [sex = m]$$



$$E(y|age, sex = m) = \alpha + \beta_1 * age + \beta_2 + \beta_3 * age$$
$$= (\alpha + \beta_2) + (\beta_1 + \beta_3) * age$$

α : mean y for 0-year-old female

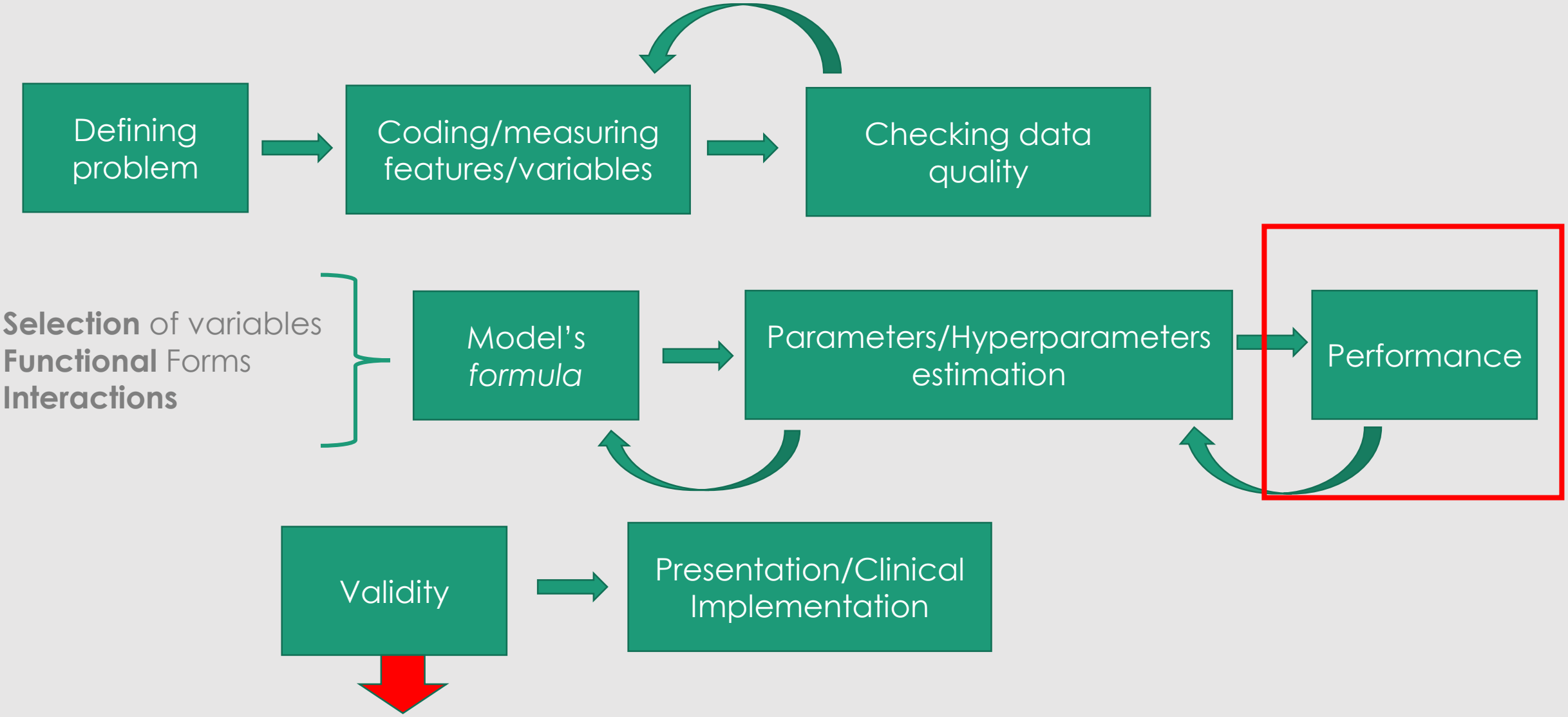
β_1: slope of age for females

β_2 : mean y for males - mean y for females, (0-year-olds)

β_3: increment in slope in going from females to males

$$E(y|age, sex = f) = \alpha + \beta_1 * age$$

INITIAL DATA ANALYSIS !!!



Possibly on **external** dataset !!!

Classical measure of overall performance: R squared

R^2 Values	Interpretation	Graph

$$y = f(x) + \varepsilon$$

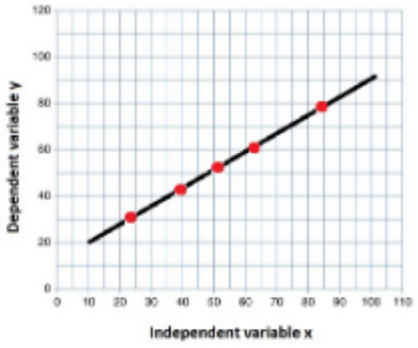
$R^2 = 1$ All the variation in the y values is accounted for by the x values

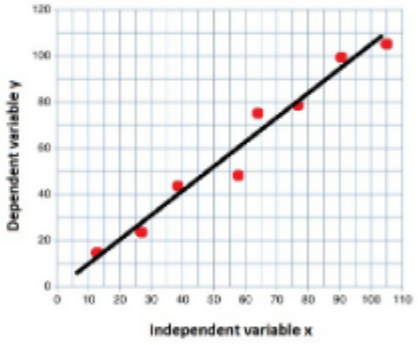
$$\bar{y} = \frac{1}{n}\sum_i y_i \qquad e_i = y_i - f_i$$

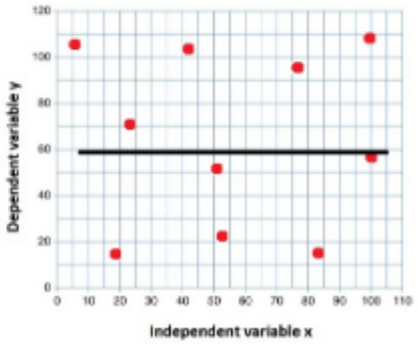
$R^2 = 0.83$ 83% of the variation in the y values is accounted for by the x values

$$SS_{tot} = \sum_i (y_i - \bar{y})^2 \qquad SS_{res} = \sum_i (y_i - f_i)^2 = \sum_i e_i^2$$

$R^2 = 0$ None of the variation in the y values is accounted for by the x values







R^2 (coefficient of determination) is the proportion of the variance for a dependent variable that's explained by an independent variable in a regression model.

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

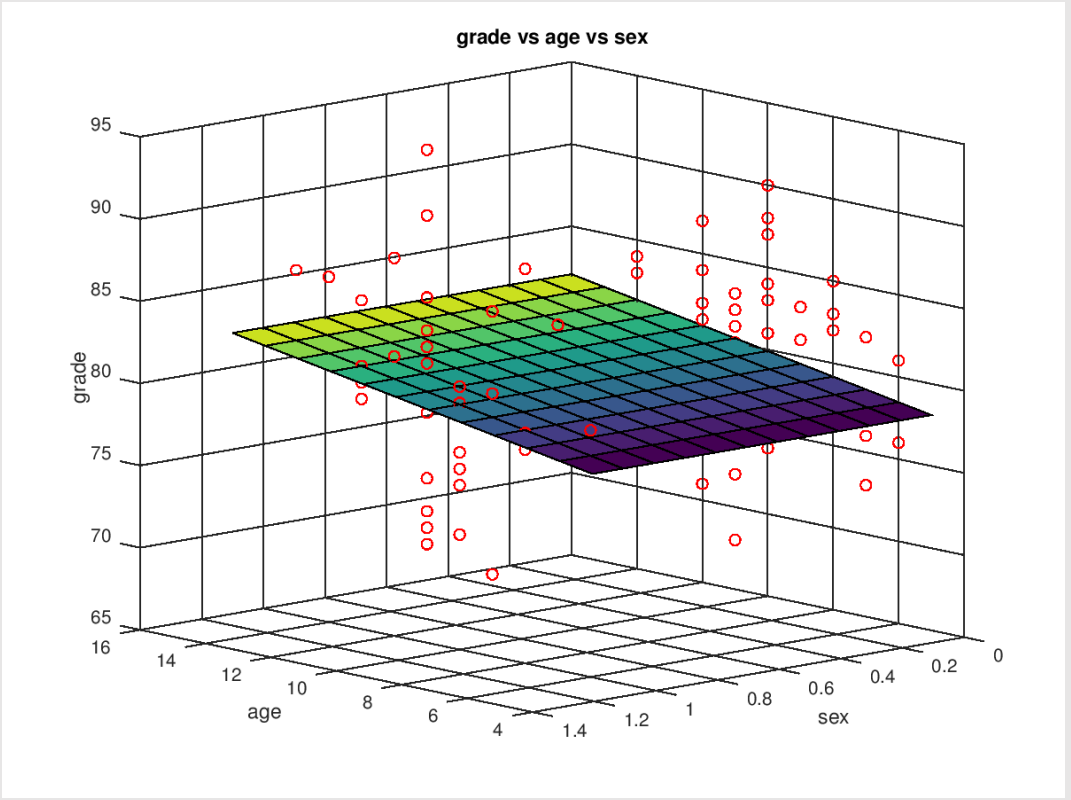
fraction of variance **unexplained**

$$SS_{reg} = \sum_i (f_i - \bar{y})^2$$

R squared for multivariable (*generalized*) models

R^2 : % of variation in Y explained by the model
[**adjusted for *p*=#covariates, *n*=sample size**]

$$R^2_{adj} = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$



Binary/[time-to-event] models:

- Cox and Snell R^2
- Nagelkerke's R^2

$$R^2_{CS} = 1 - \exp\left[\frac{2}{n}\left(ln(Lik_{Null}) - ln(Lik_{Model})\right)\right]$$

likelihood of the null model with only the intercept *vs* a given set of parameters