CANCER RESEARCH UK | Cambridge Institute

Together we are beating cancer

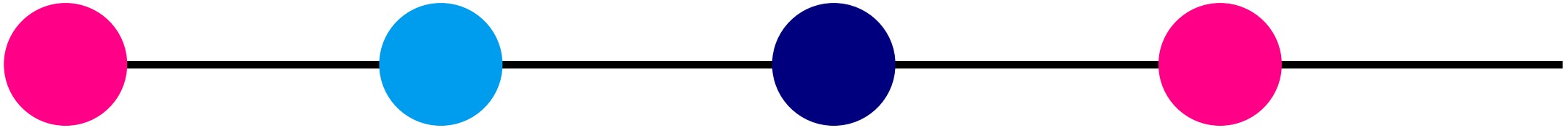
Luca Porcu & Chandra Chilamakuri (Bioinformatics core)

21st February 2025

# Linear regression models

Fixed-effects models

# Process flow



**Stage 1**

Effects and regression models

**Stage 2**

Analysis of variance (ANOVA)

**Stage 3**

Simple regression model

**Stage 4**

Multivariable regression model

CANCER RESEARCH UK | Cambridge Institute

# Effects

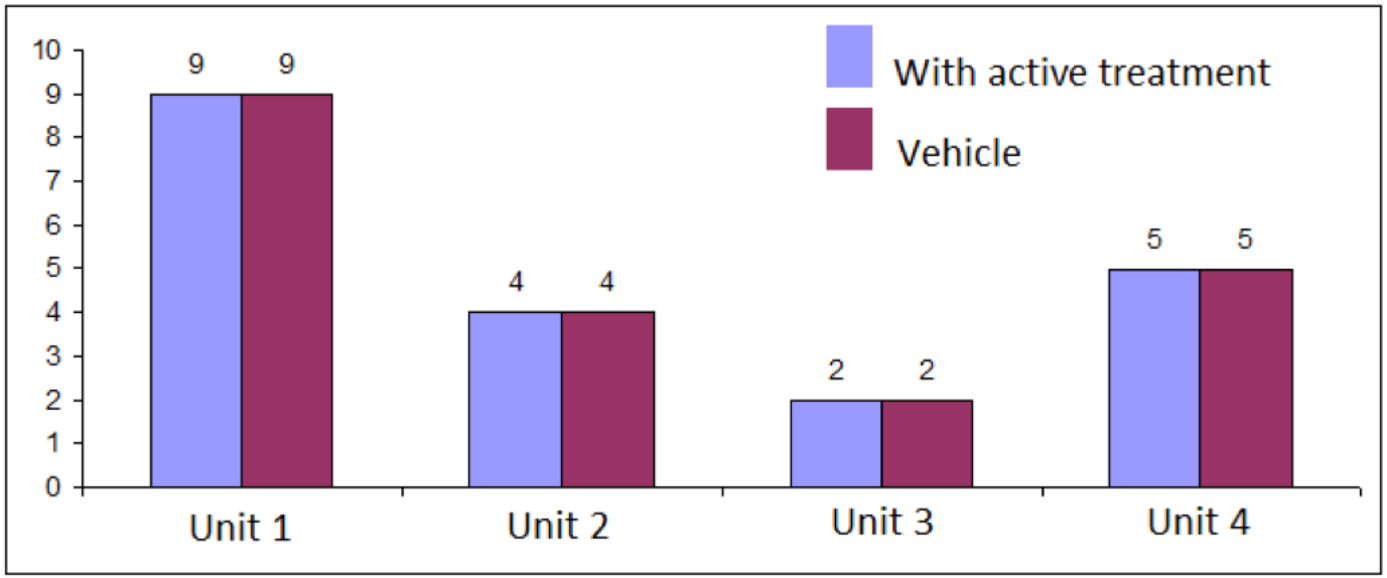
Definition and classification

9.00 -09.15 am

Together we are beating cancer

# Definition





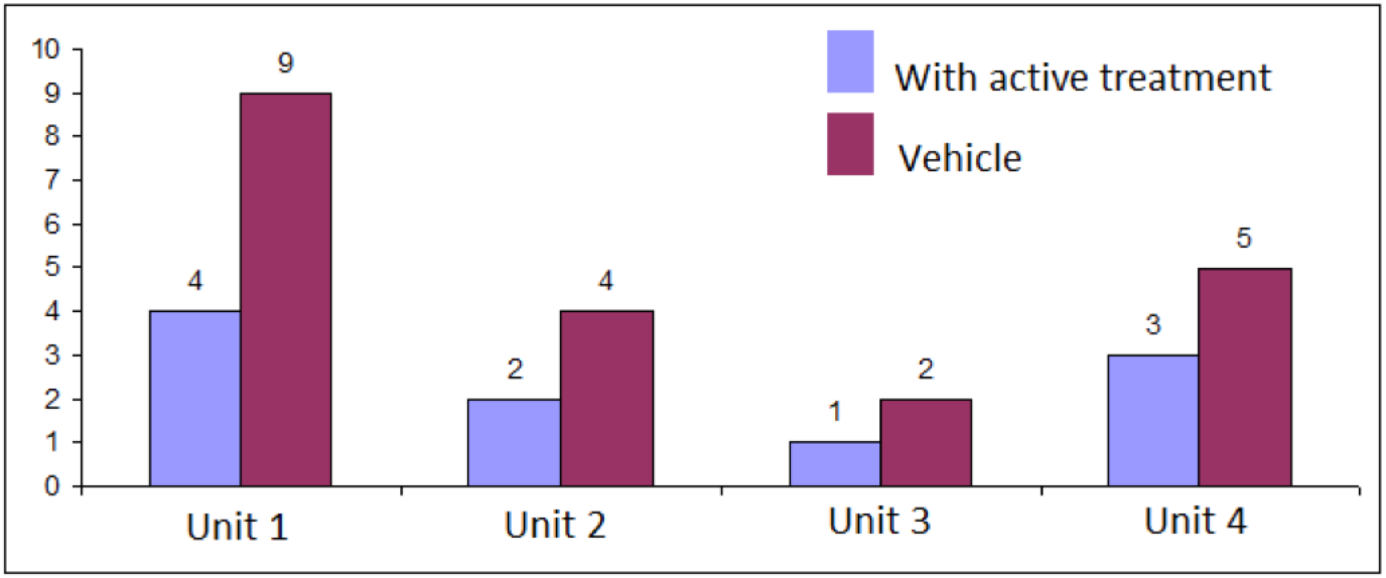

**Number of metastases**

Each unit exhibits a **response** that is observed some time after **exposure**.

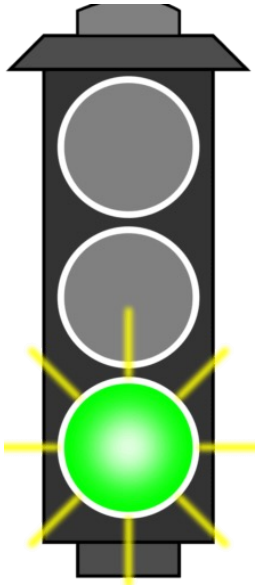
**Definition of "no effect":** each unit would exhibit the same value of the response whether assigned to exposure or not. If changing the exposure assigned to a unit changed the unit's response then the exposure has at least some effect.

# Examples of effects






**Number of metastases**



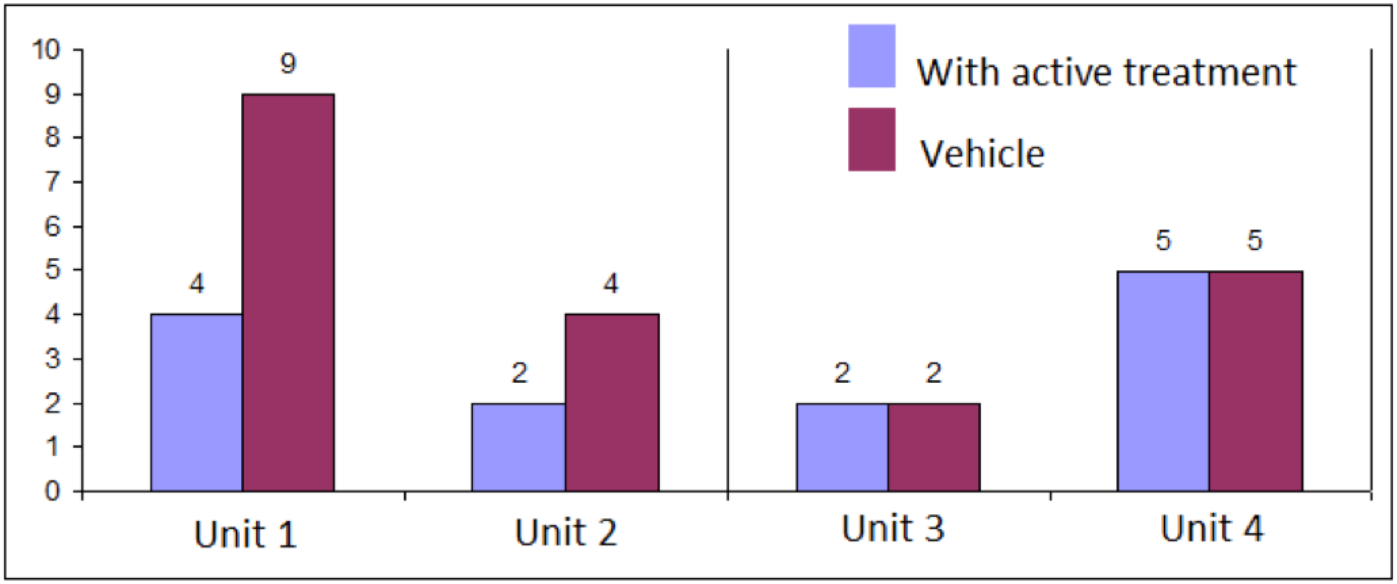
N° of metastases in the treated group: 10

N° of metastases in the vehicle: 20

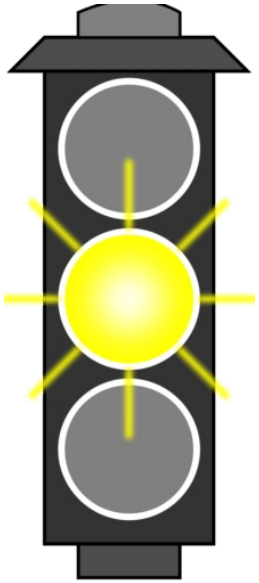
**This effect is understandable, easily detectable and systematic**

# Examples of effects






**Number of metastases**

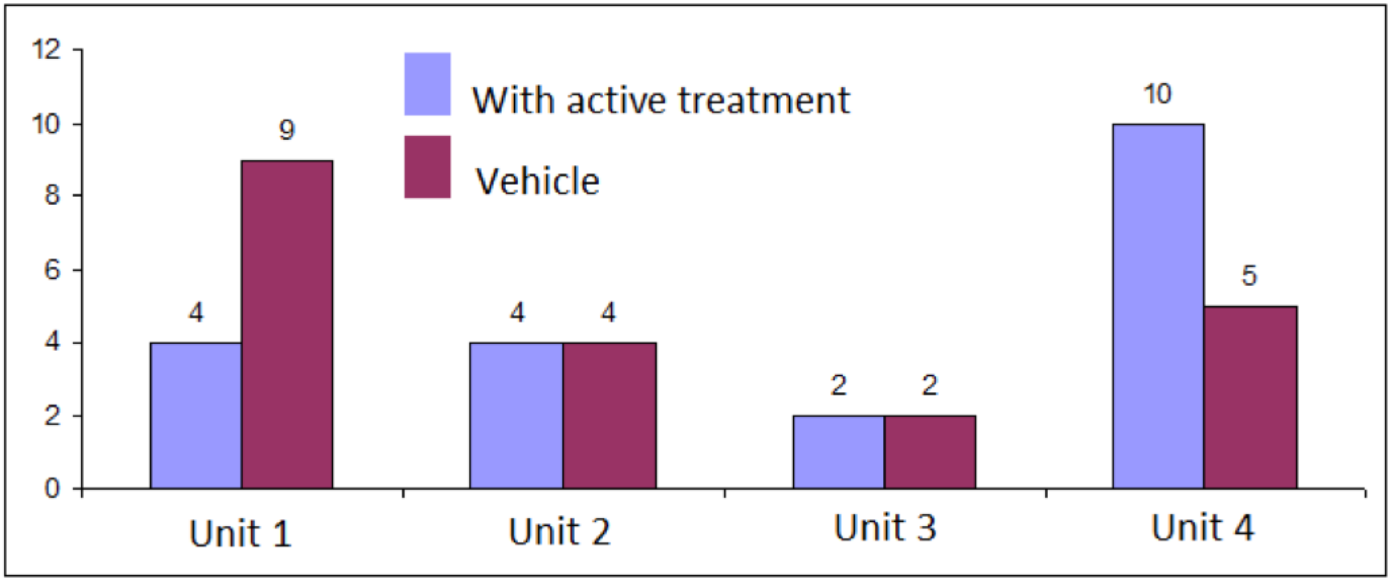


| Experimental group | Male mice | Female mice |
|---|---|---|
| Active | 6 | 7 |
| Vehicle | 13 | 7 |

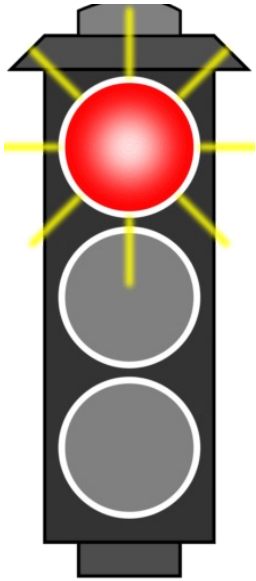
**This effect is understandable, detectable with sufficient units and systematic**

# Examples of effects






**Number of metastases**



N° of metastases in the treated group: 20

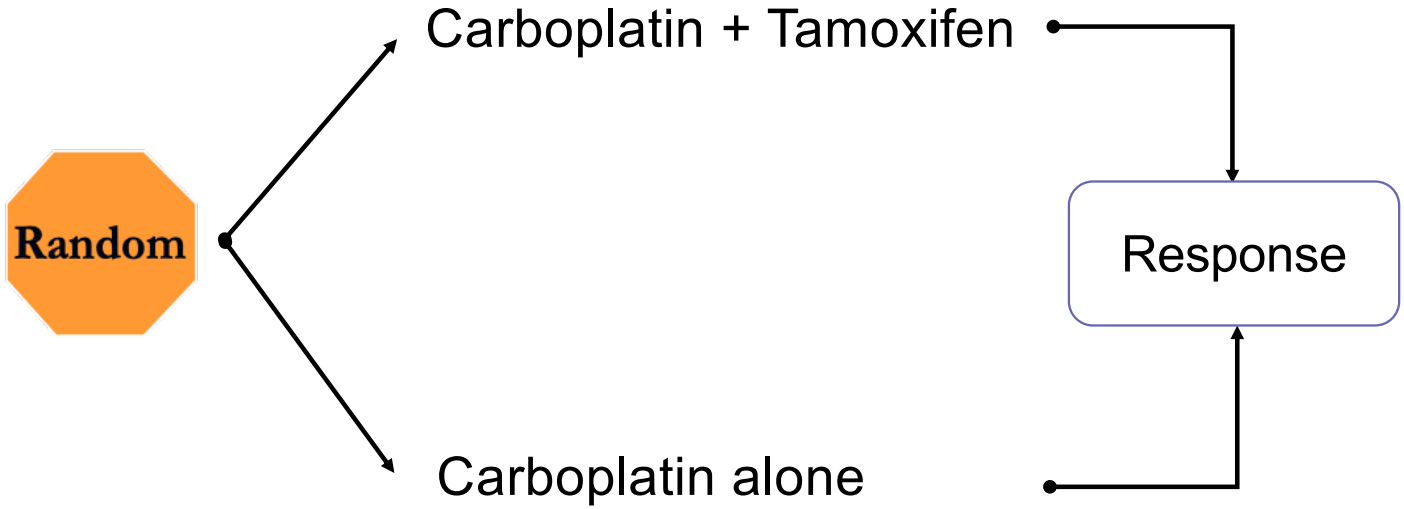
N° of metastases in the vehicle: 20

**This effect can someday be understood, not easily detectable and unsystematic**

# Different sources of effects



## Treatment



# Different sources of effects



## Environment and features of experimental units

Temperature, humidity, season, barometric pressure, lunar cycle, noise, air movement, light, smells, room characteristics, cage size and design, bedding material, nest box design, nest materials, number of animals in group, water quality, diet type, diet availability, diet quality, frequency and duration of handling

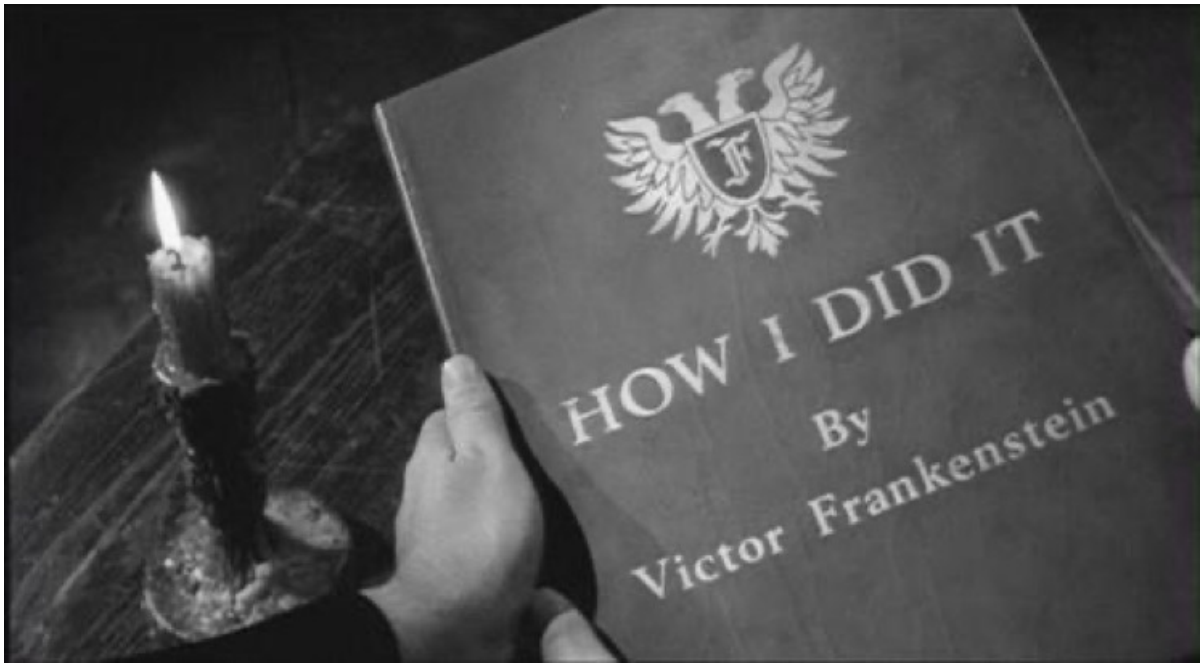
Species, sex, strain, genotype, health status, batch, supplier, age, body weight, litter size, oestrus stage of females, level of inter-animal aggression

# Different sources of effects



## The operator

- Calibration of instruments
- Measurement errors
- Recording errors
- Preparation of test materials
- Operative procedures



# Classification of effects



## Fixed effects

Effects attributable to a finite set of levels of a source (i.e. *predictor* in statistical terms) that occur in the data and which are there because we are interested in them. Fixed effects are **parameters** associated with an entire **population**.

**Treatment** **Sex** **Genotype** **Time**

# Classification of effects



## Random effects

Effects attributable to a (usually) infinite set of levels of a source (i.e. *predictor*), of which only a **random sample** are deemed to occur in the data. Random effects are associated with **individual observational units** drawn at random from a population.

**Mouse** **Litter** **Batch** **Place of residence**

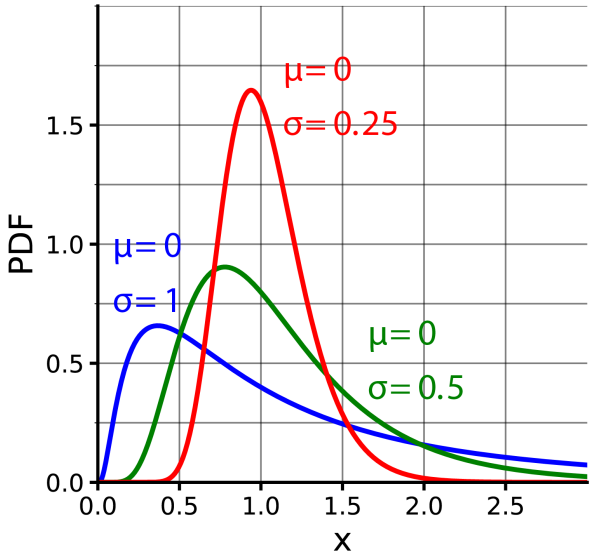
# Classification of effects

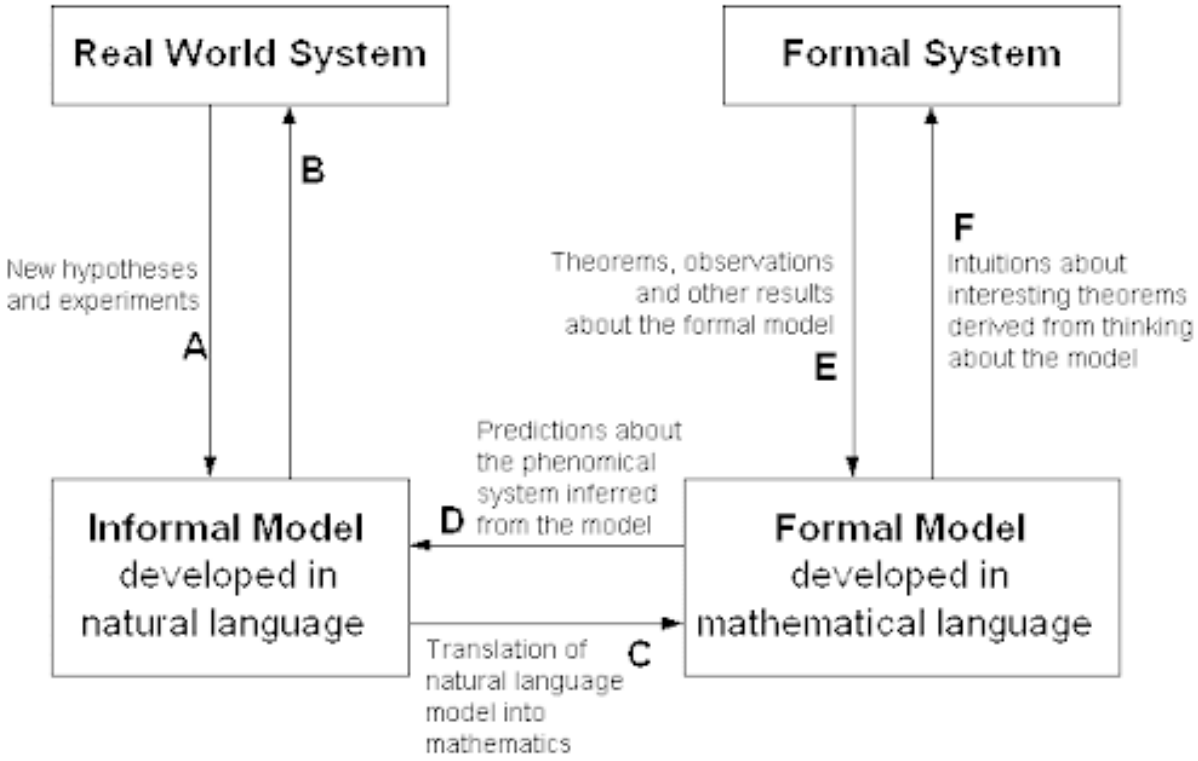


## Unpredictable effect (i.e. *error*)

**Unpredictable** effect attributable to “hidden” sources whose consequence is the deviation of the observed value from the “true value” (i.e. parameter) of the population or the individual observational unit.

At any given combination of age, gender, height, smoking status and place of residence, many different values of FEV (lung capacity) could be recorded, and so produce a distribution of recorded FEV values. At this given combination of age, gender, height, smoking status and place of residence, the distribution of FEV values has a unique mean FEV (“the true value”).






# Regression models

Definition and classification

9.15-9.30 am

Together we are
beating cancer

# Definition



A regression model describes mathematically the population distribution of the **response** (e.g. FEV distribution in UK) as a function of **fixed effects** (e.g. age, gender, height, smoking status), **random effects** (e.g. place of residence) and ***error***.

$$Y_i = f(x_{1,i},..,x_{n,i}; x_{n+1,i},\ldots, x_{n+k,i}; \varepsilon_i)$$

where i is the observational unit of the population.

Hi, my name is Mark Brown. I am 35 years old, male, 1.75 meters tall and ex-smoker. I live in Cambridge, UK. ***This morning*** my FEV is 4.73 liters in 1 second.

# Definition



Every statistical regression model consists of two components.

| Component | Meaning | Example | |
|---|---|---|---|
| Systematic component | It describes the relationship between the predictors and the population parameter of interest (e.g. mean) | $\mu = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n$, where $x_1,..,x_n$ could be **fixed** and **random** predictors | |
| Random components | It describes the properties of **random effects** and ***error*** | Random effects are normally distributed | $N(0, \sigma_{random})$ |
| | | Error is normally distributed | $N(0, \sigma_{error})$ |
| | | Error variance is constant | $\sigma_{error}^2$ = constant |

# Examples of regression models



| Model | Component | Structure |
|---|---|---|
| 1 | Systematic | $\boldsymbol{\mu}$ (Y) = $\boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 x_1 + \ldots + \boldsymbol{\beta}_n x_n$ |
| | Random | Random effects and error are normally distributed. Error variance is constant. |
| 2 | Systematic | log [$\boldsymbol{\mu}$ (Y)] = $\boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 x_1 + \ldots + \boldsymbol{\beta}_n x_n$ |
| | Random | Random effects and error are log-normal distributed |
| 3 | Systematic | Median (Y) = $\boldsymbol{\beta}_0 + \exp(\boldsymbol{\beta}_1 x_1) + \ldots + \exp(\boldsymbol{\beta}_n x_n)$ |
| | Random | Random effects are normally distributed |
| | | Error is exponentially distributed |

**Legend:** Y = response; $x_1, \ldots x_n$ = fixed and random effects; exp = exponential function; $\boldsymbol{\beta}_0$ ,$\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_n$ = coefficients (i.e. parameters) of the regression model to estimate

# Assumptions of linear regression models

| Component | Assumption | Meaning | |
|---|---|---|---|
| Systematic component | Population mean | We are interested to describe the population mean | $\boldsymbol{\mu}$ |
| | Linearity | The coefficients are assumed to combine the effects of the predictors linearly | $\boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 x_i + \ldots + \boldsymbol{\beta}_n x_i$ |
| Random components | Random effects | They are normally distributed | N (0, $\boldsymbol{\sigma}_{random}$ ) |
| | Error | Error is normally distributed | N (0, $\boldsymbol{\sigma}_{error}$ ) |
| | | Error variance is constant | $\boldsymbol{\sigma}^2$ = constant |

# Linear fixed-effects and mixed-effects models

A linear model that incorporate only fixed effects is called
linear fixed-effects model (or merely **linear model**)

A linear model that incorporate both fixed and random effects is called
**linear mixed-effects model**

http://bioinformatics-core-shared-training.github.io/IntroductionToStats/practical.html