FACULTY OF HEALTH SCIENCES - SCHOOL OF MEDICINE
MSc Health Statistics and Data Analytics

ARISTOTLE UNIVERSITY OF THESSALONIKI

Linear Models

Anna-Bettina Haidich
Associate Professor of Medical Statistics –Epidemiology
haidich@auth.gr

THESSALONIKI 2021-22

Materials and announcements

- elearning.auth.gr
- School of Medicine
- Postgraduate Courses
- Health Statistics & Data Analytics
- Course “Linear Models”

Elearning platform (moodle)

elearning.auth Guides Course Request Contact

Linear models

Home / My courses / Faculty of Health Sciences / School of Medicine / Postgraduate courses / Health Statistics and Data Analytics / Linear models

Activities

- Assignments
- Forums
- Quizzes
- Resources

Welcome to the course of **Linear Models**.

The course has the following tutors Anna-**Bettina** Haidich, Christos T Nakas, Eleni Verykouki, Konstantinos Bougioukas, Fani Apostolidou Kiouti, and Eirini Pagkalidou.

For this course we will meet in person for two weekends in **January** between **21-23/01/2022** and February **4-6/02/2022** and in the meantime we will be available online to answer any queries you might have. Students will also work on quizzes and exchange questions and solutions in discussion forums during this period. **The final exam will be held on February 25, 2022**. For more details, please refer to the course syllabus.

This course concentrates on advanced statistical questions:

- What is the relationship between the variables collected?
- Which procedure should be employed to explore these relationships?
- What is the interpretation of the obtained statistical results?

By the end of the course the user will be able to fit the best model to describe the relationships based on the data from their study. They will also be able to interpret the results returned from diagnostic tests and evaluate the assumptions under which their model was built.

For zoom connection the link is the following:

https://authgr.zoom.us/j/94680285974?pwd=a0FsOE93ajZlU29WNGgrR3oySFpyQT09

Meeting ID: 946 8028 5974

Passcode: 364958

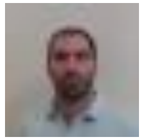
Course outline

Date	Hours	Topics	Tutor
21 January 2022	**19:30-20:00**	**Outline of the course**	**Anna-Bettina Haidich**
	20:00-21:00	**Correlation – Simple Linear Regression**	**Konstantinos Bougioukas**
22 January 2022	**09:00-11:00**	**Multiple Linear Regression**	**Konstantinos Bougioukas**
	11:30-13:00	**Model Building in Linear Regression**	**Konstantinos Bougioukas**
	13:30 -	**IQVIA presentation**	**George Nikolaidis, Andreas Karabis, Konstantina Skaltsa**
23 January 2022	**09:00-11:00**	**Logistic Regression**	**Eleni Verykouki**
	11:30-13:00	**R Practical**	**Fani Apostolidou Kiouti**
	13:00-15:00	**ROC Analysis + R Practical**	**Christos T Nakas**
4 February 2022	**17:00-18:00**	**Cox Proportional Hazards models**	**Christos T Nakas**
	18:00-19:00	**R Practical**	**Eleni Verykouki**
	19:00-21:00	**Parametric Survival Analysis + R Practical**	**Eleni Verykouki**
5 February 2022	**09:00- 11:00**	**Poisson/Negative Binomial models/Zero Inflated models**	**Eleni Verykouki**
	11:30-15:00	**R Practical**	**Fani Apostolidou Kiouti**
6 February 2022	**09:00- 11:00**	**Handling Missing Data – Imputation Methods**	**Anna-Bettina Haidich, Eirini Pagkalidou**
		Course summary + Practice	**All**

Grading system

Component	% of grade	When
Lecture participation through lectures and forum discussion	15%	
Quizzes	15%	
Final Exam	70%	February 25, 2022
Total	100%	

Communication

- Anna-**Bettina** Haidich haidich@auth.gr 
- Christos T Nakas cnakas@uth.gr 
- Eleni Verykouki everykouki@auth.gr 
- Konstantinos Bougioukas mpougioukas@auth.gr 
- Fani Apostolidou-Kiouti fania@auth.gr 
- Eirini Pagkalidou pagkalidou@auth.gr 

OPEN ACCESS Freely available online PLOS ONE

Statistical Trends in the *Journal of the American Medical Association* and Implications for Training across the Continuum of Medical Education

Lauren D. Arnold[1*¤a], Melissa Braganza[2¤b], Rondek Salih[3¤c], Graham A. Colditz[1]

Table 3. Statistical measures and methods in *JAMA* articles published in 1990, 2000, and 2010*.

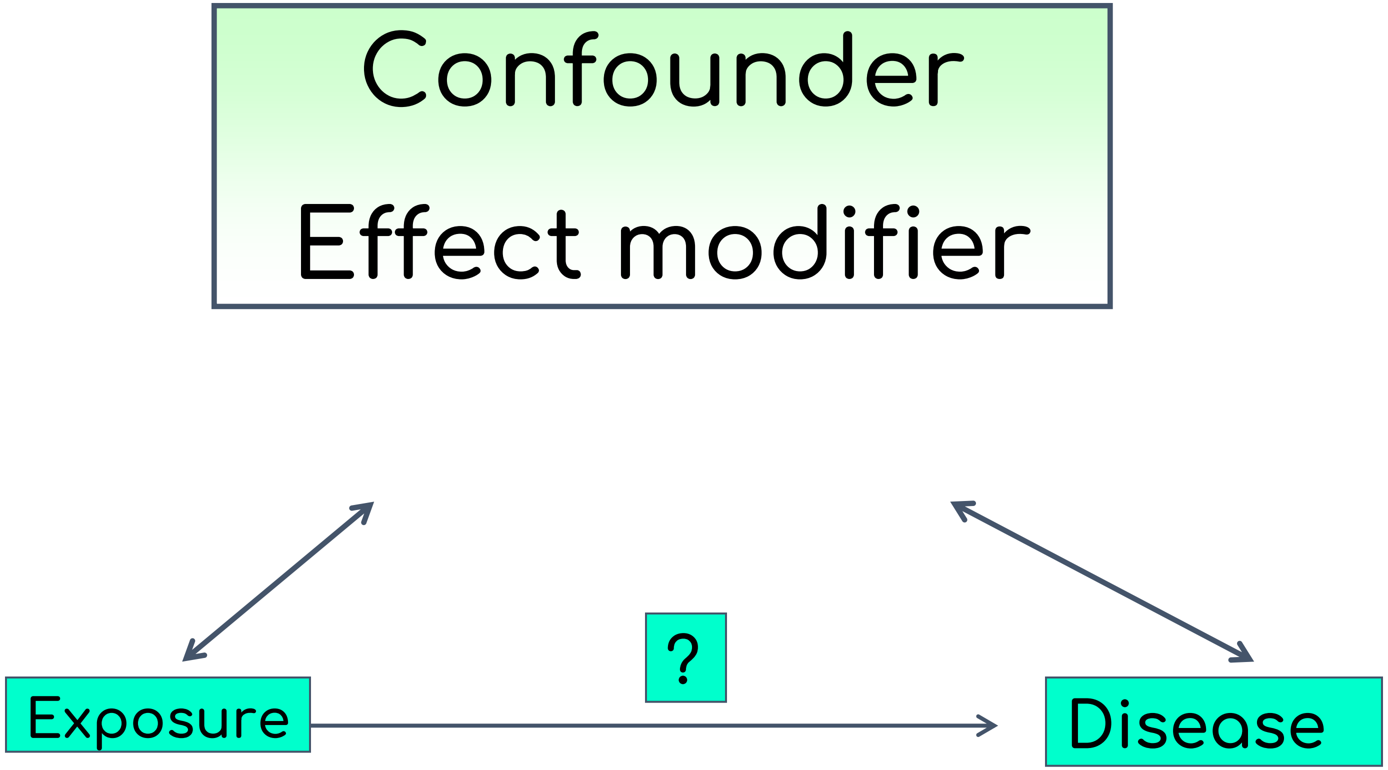
	Article Year			
Characteristics	**1990 (n = 133)**	**2000 (n = 122)**	**2010 (n = 106)**	**p-value**
Descriptive statistics	124 (93.2%)	122 (100%)	106 (100%)	-
Low-level statistical measures[†]	108 (81.2%)	116 (95.1%)	105 (99.1%)	<0.001
Morbidity & mortality	76 (57.1%)	60 (49.2%)	73 (68.9%)	0.011
ANOVA	26 (19.5%)	24 (19.7%)	18 (17.0)	0.844
Chi square	54 (40.6%)	51 (41.8%)	51 (48.1%)	0.471
Fisher exact	19 (14.3%)	18 (14.8%)	20 (18.9%)	0.583
Mantel-Haenszel	11 (8.3%)	15 (12.3%)	7 (6.6%)	0.301
Epidemiologic statistics[‡]	28 (21.1%)	34 (27.9%)	33 (31.1%)	0.190
t-test	28 (21.1%)	31 (25.4%)	28 (26.4%)	0.577
Power	7 (5.3%)	7 (5.7%)	28 (26.4%)	<0.001
p-trend	6 (4.5%)	17 (13.9%)	14 (13.2%)	0.023
Logistic regression	27 (20.3%)	42 (34.4%)	28 (26.4%)	0.039
Simple linear regression	12 (9.0%)	17 (13.9%)	13 (12.3%)	0.460
Poisson regression	0 (0.0%)	11 (9.0%)	8 (7.5%)	0.003
Multi-level modeling	3 (2.3%)	11 (9.0%)	34 (32.1%)	<0.001
Multiple comparison	7 (5.3%)	8 (6.6%)	9 (8.5%)	0.609
Multiple regression	32 (24.1%)	52 (42.6%)	51 (48.1%)	<0.001
Non-parametric test	[illegible]	[illegible]	[illegible]	[illegible]
Survival analysis	[illegible]	[illegible]	[illegible]	[illegible]
Cox models	10 (7.5%)	17 (13.9%)	34 (32.1%)	<0.001
Kaplan Meier	5 (3.8%)	13 (10.7%)	24 (22.6%)	<0.001
Transformation	9 (6.8%)	12 (9.8%)	10 (9.4%)	0.6374

Multiple regression 32 (24.1%) 52 (42.6%) 51 (48.1%) <0.001
Cox models 10 (7.5%) 17 (13.9%) 34 (32.1%) <0.001

*Excludes statistics in which there were n<15 across all three years of review; [†]Includes standard deviations, standard errors, confidence intervals, and p-values; [‡]Includes odds ratios, relative risks, attributable risk, sensitivity, and specificity.

doi:10.1371/journal.pone.0077301.t003

Third factor



Confounding and effect modification

Confounding	Effect modification
• A distortion of the association between an exposure and an outcome that occurs when the study groups differ with respect to other factors that influence the outcome. • A type of bias that can and should be adjusted for the analysis	• It occurs when the magnitude of the effect of the primary exposure on an outcome differs depending on the level of a third variable. • A true biological phenomenon that should be further explored

Confounding – example

Exposure

Outcome

Coffee consumption → Pancreatic cancer

Confounding – example

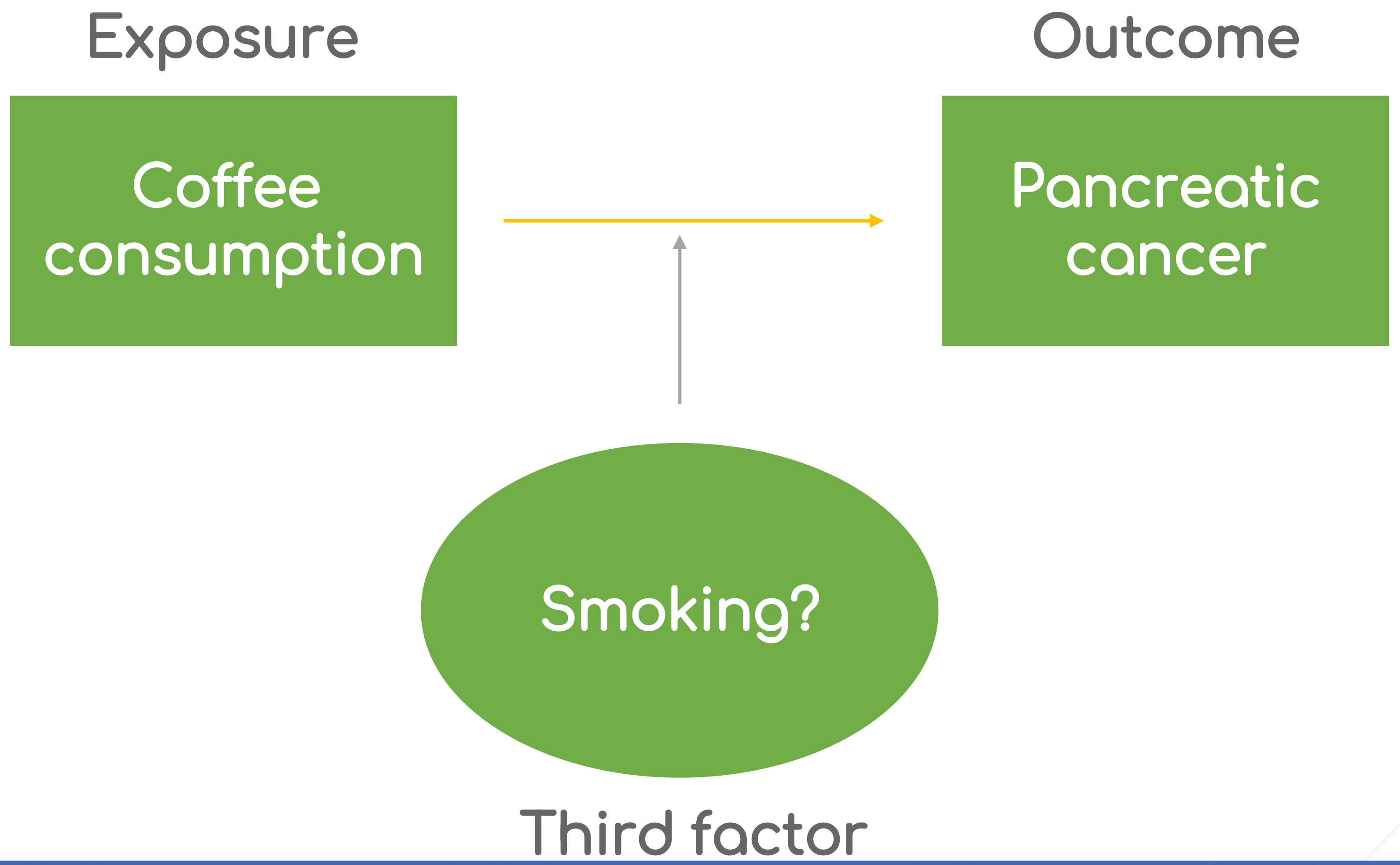
Coffee consumption	Pancreatic cancer	
	Yes	No
Yes	450	200
No	300	250
Total	**750**	**450**

OR = (450*250)/(200*300) = 1.88
95% CI: 1.48, 2.38

Crude

The odds of developing pancreatic cancer is almost 2 times higher in coffee drinkers than non coffee drinkers

Confounding – example



C… ing – example

Smoking is a risk factor for developing cancer of the pancreas

Smoking	Pancreatic cancer	
	Yes	No
Yes	600	150
No	150	300
Total	**750**	**450**

OR = (600*300)/(150*150) = 8.00   95%CI: 6.13, 10.43

The odds of developing pancreatic cancer is 8 times higher in smokers than non-smokers

Confounding – example

Smoking is associated with coffee drinking

	Coffee consumption	
	Yes	No
Yes	500	250
No	150	300
Total	**650**	**550**

OR = (500*300)/(250*150) = 4.00   95%CI: 3.12, 5.13

Coffee consumption was 4 times more likely in smokers than in non-smokers

Coffee drinkin[illegible] & Pancreatic cancer
S[illegible]ratifie[illegible]kin[illegible]

There is no association between coffee drinking and cancer of the pancreas

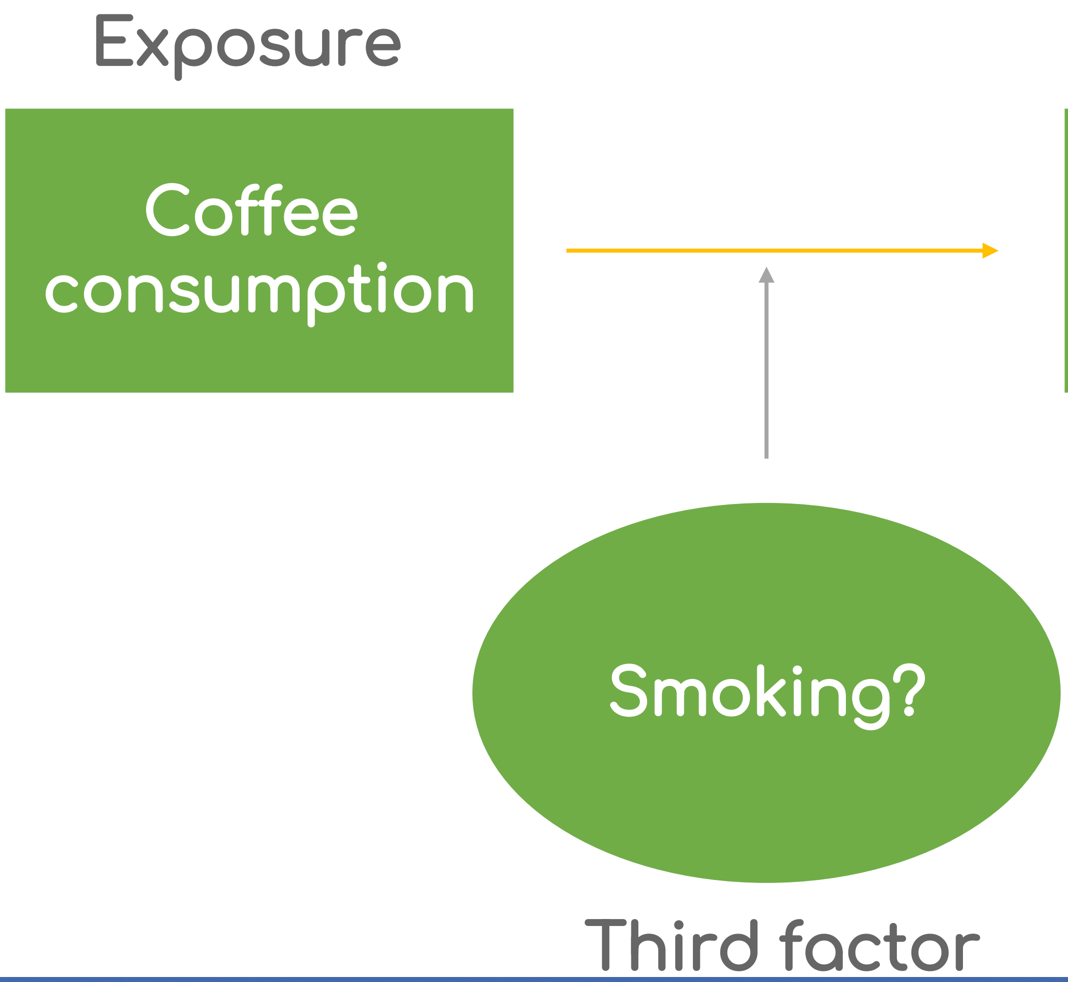
[illegible]	[illegible]ases	Controls	OR
[illegible]	400	100	1.00
[illegible]	200	50	0.68, 1.46

Non-smokers	Cases	Controls	OR
Coffee	50	100	1.00
No coffee	100	200	0.66, 1.52

Stratified OR with the Mantel-Haenszel method: 1.00

Crude OR : 1.88

Confounding



Conditions satisfied:

- Smoking is associated with coffee drinking
- Smoking is a risk factor of cancer of the pancreas

Confounding factor

- Confounding is the situation where an association between an exposure and an outcome is entirely or partially due to another exposure (called the confounder).

  - Positive confounding → stronger association (Crude OR > adjusted OR)
  - Negative confounding → weaker association (Crude OR < adjusted OR)

Three conditions must be satisfied:

- it must be associated with the exposure of interest
- it must be a risk factor for the outcome of interest
- It must not be on the causal pathway

Controlling for it

- In the design phase of the study
  - Randomization
  - Minimization
  - Matching

- In the analysis phase of the study
  - Stratification
  - Multivariable models

Effect modification

- Effect modification occurs when the effect of an exposure is different among different subgroups (e.g. gender, race)

- In statistics, it is synonym with interaction

- The crude estimate of the association (e.g. odds ratio) is expected to lay between the estimates of the odds ratio for the stratum-specific estimates

Why do we care about it?

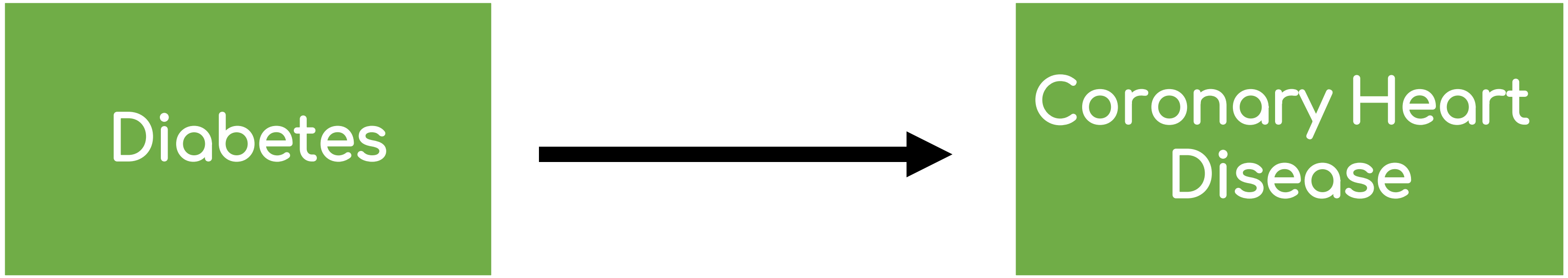
- To define high-risk subgroups for preventive actions
- To increase precision of effect estimation by taking into account groups that may be affected differently
- To increase the ability to compare across studies that have different proportions of effect-modifying groups
- To aid in developing a causal hypotheses for the disease

Handling effect modification

- Designing the study
  - Collect information on potential effect modifiers.
  - Power the study to test potential effect modifiers a priori
  - Don't match on a potentially important effect modifier

- Analyzing the study
  - Stratification and report stratum specific estimates
  - Multivariable models test for interaction

Example



Diabetes & CHD

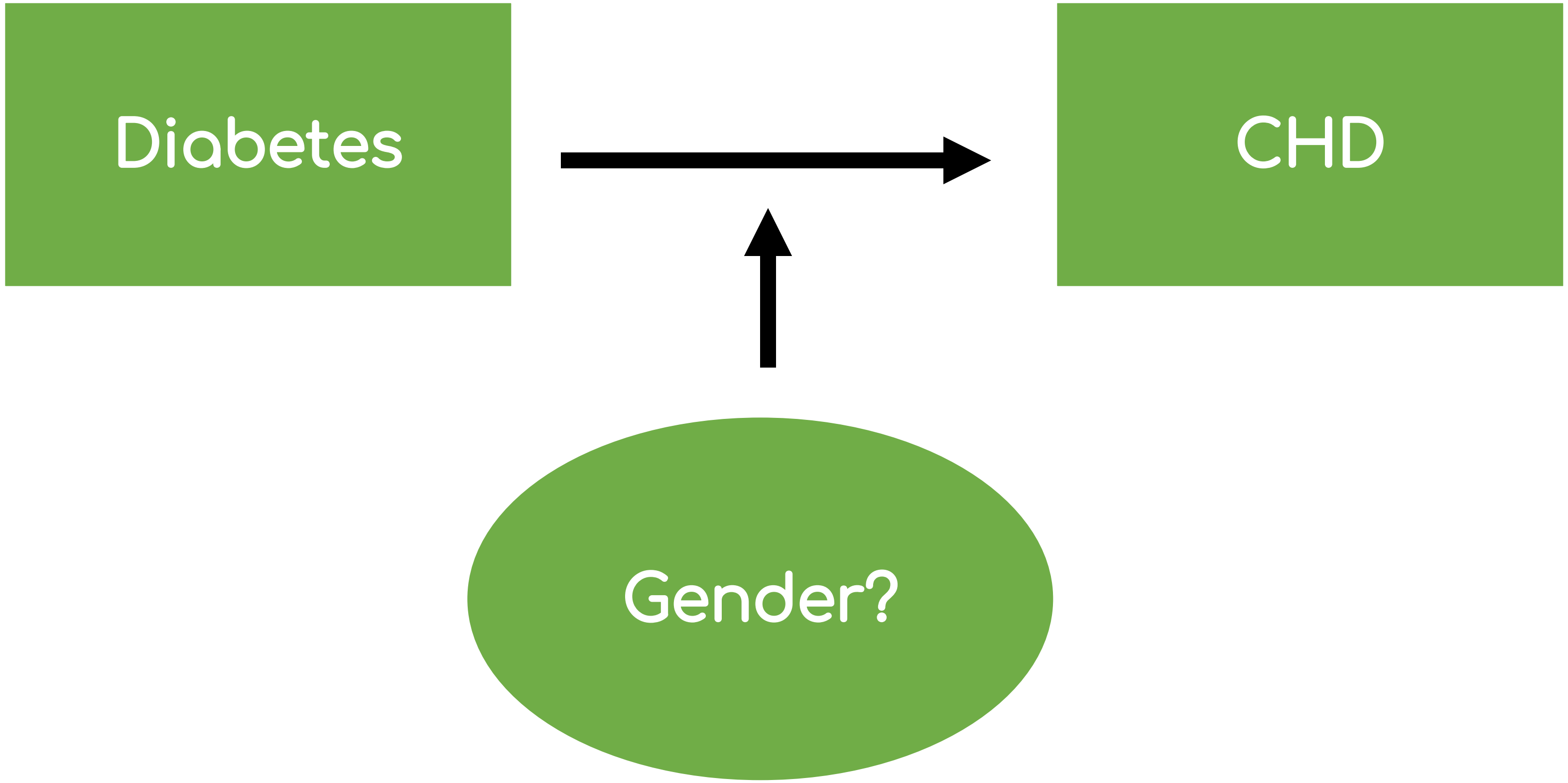
	CHD	No CHD	OR
Diabetes	25	170	3.4
No diabetes	95	2194	

The odds of CHD is 3.4 times higher in patients with diabetes than no diabetes

Crude OR: 3.4

95%CI: 2.13, 5.42

Effect modification



Diabetes & CHD

Females	CHD	No CHD	OR
Diabetes	13	93	6.66
No diabetes	25	1191	3.30, 13.45

Males	CHD	No CHD	OR
Diabetes	12	77	2.23
No diabetes	70	1003	1.16, 4.30

Crude OR: 3.4

Confounding or effect modification?

- Confounding
  - The crude estimator (e.g. RR, OR) is outside the range of the two stratum-specific estimators
  - A confounder changes the estimate of the risk by 10% or more in the adjusted analysis compared to the unadjusted

  - **Mantel-Haenszel stratified analysis or multivariable analysis**

- Effect modification
  - The crude estimator (e.g. RR, OR) is closer to a weighted average of the stratum-specific estimators
  - The two stratum-specific estimators differ from each other

  - **Separate analysis by subgroup or multivariable analysis with the interaction term**

Multivariable models

$$y = \overbrace{\underbrace{b_0 + b_1x_1 + ... + b_px_p}_{\text{Independent variables}}}^{\text{Exposure, risk factors, confounders and effect modifiers}} + \varepsilon$$

outcome

Dependent variable

Multivariable models

- Dependent variable
  - continuous → linear regression
  - binary → logistic regression
  - Time to event → Cox proportional hazards regression
  - Counts → Poisson regression
- Independent variable
  - continuous/categorical

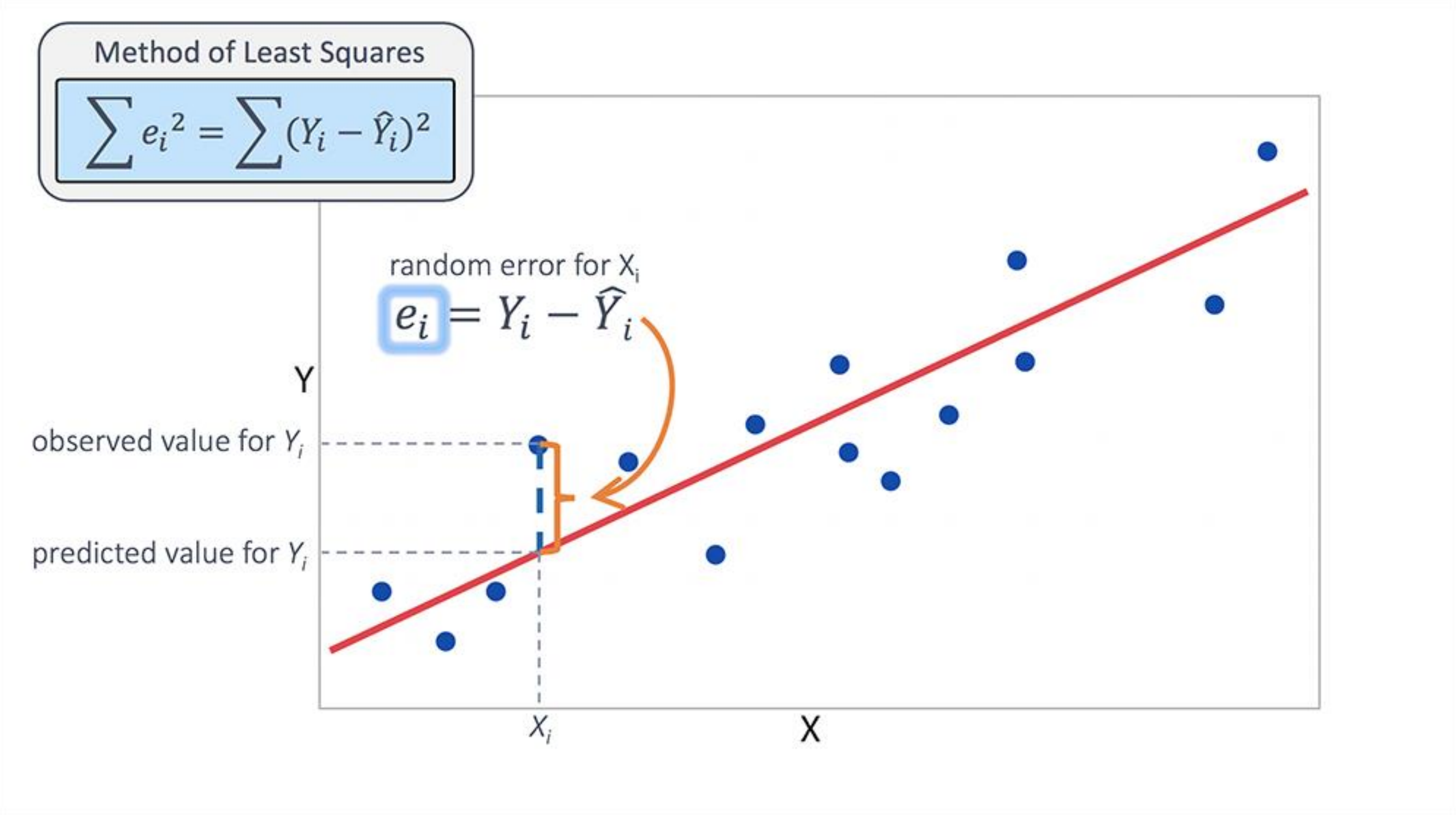
Multivariable models

- Data (y)= Fitted model ($b_0+b_1x_1+....+ b_px_p=\hat{y}$) + residuals ($\varepsilon$)

- $\varepsilon= y - \hat{y}$ (error) <- residuals

- Approximation

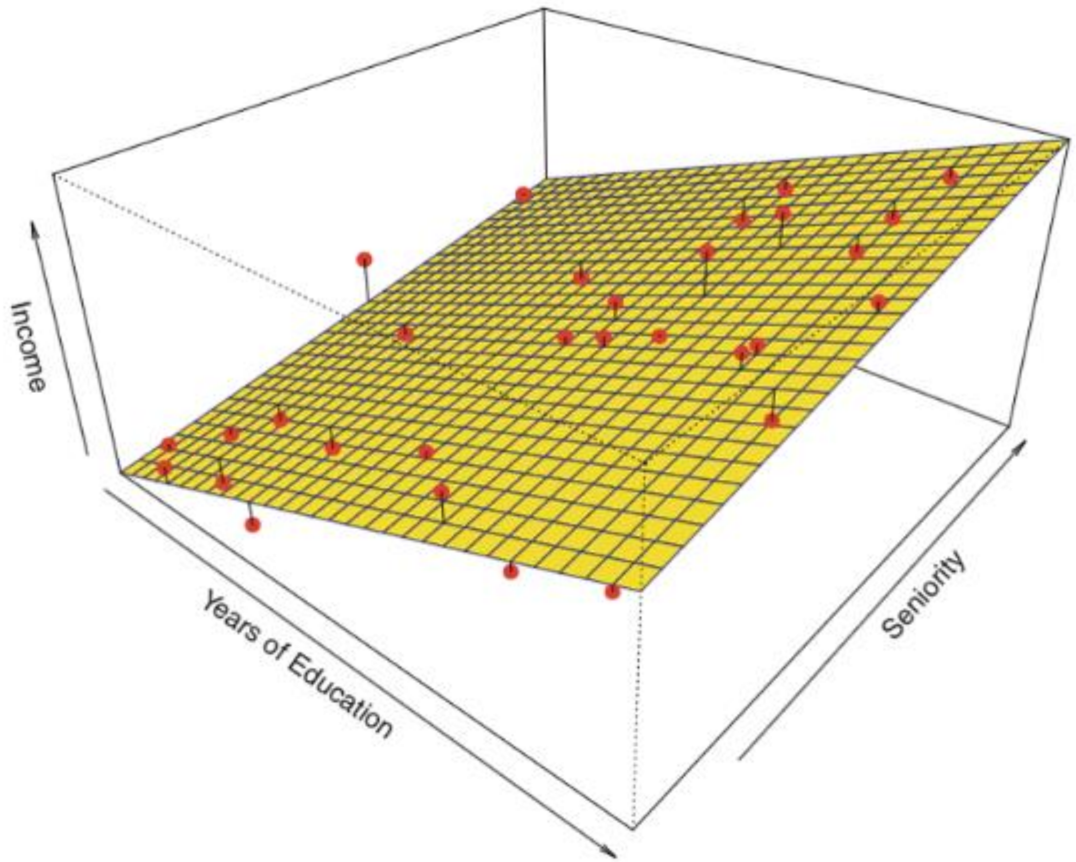
Residuals



Method of Least Squares
$\sum e_i^2 = \sum (Y_i - \hat{Y}_i)^2$
random error for X_i
$e_i = Y_i - \hat{Y}_i$
Y
observed value for Y_i
predicted value for Y_i
X_i
X

Multiple regression

$$Income = b_0 + b_1 * years\ education + b_2 * seniority$$



Steps for model building

1. **Define and Design**
    - Construct research questions
    - Define the study design
        - ❖ Experimental or observational
        - ❖ Simple or stratified randomization
        - ❖ Potential confounders and control variables
        - ❖ Longitudinal or repeated measurements on a study unit
    - Variable selection and measurement defined
        - ❖ Continuous or categorical variables
    - Write an analysis plan
    - Calculate sample size estimations

Steps for model building

2. **Prepare and explore**
   - Collect, code, enter and clean data
   - Create new variables
   - Run univariate and bivariate Statistics
   - Run an initial model
3. **Refine the model**
   - Refine predictors and check model fit
     - ❖ Test interactions
     - ❖ Drop non significant control variables
   - Test assumptions
   - Check for and resolve data issues
   - Interpret results

Goal of the Model?

Prediction

Minimize prediction error rather than causal interpretation

Evaluating a predictor of primary interest

Have to account for confounding or effect modifiers

Identifying the important independent predictors of an outcome

Most difficult
false-positive associations, potential complexity and not a single best model


All models are wrong
but some are useful
George E.P. Box

Have a nice start!