Epidemiology Basics

Max Salvatore

2022-05-16

Contents

1 **Introduction** **5**

2 **Causality** **7**

2.1 Three criteria for causation 7

2.2 Probabilistic approach 7

2.3 Four causal types 8

2.4 Four causal inference frameworks 9

2.5 1. Sufficient-component cause model (Rothman's Causal Pie Model) 9

2.6 2. Bradford-Hill's Viewpoints 11

3 **Study design** **13**

3.1 Cross-sectional studies 13

3.2 Case-control studies 14

3.3 Cohort studies 16

3.4 Randomized controlled trials (RCTs) 20

3.5 Study design flowchart 21

4 **Confounding bias** **23**

4.1 Identifying confounding 23

5 **Selection bias** **29**

5.1 Selection bias definitions 29

5.2 Key points 33

6 Regression **35**

6.1 Linear regression 35

7 Literature **39**

Chapter 1

Introduction

Hello and welcome! This book will contain notes on a variety of epidemiologic topics including study design and biases.

At present, this is a **not a functional book** and the content is **a work in progress** .

The current version of the book has not been reviewed for correctness and references.

Chapter 2

Causality

Material excerpted and adapted from Lindsay Kobayashi, PhD for EPID601 at the University of Michigan School of Public Health

Epidemiology is concerned with identifying causes of health outcomes in order to inform effective and equitable interventions to improve them.

What is a cause?

- "An antecedent event, condition or characteristic that was necessary for the occurrence of the disease at the moment it occurred, given that other characteristics are fixed."

2.1 Three criteria for causation

1. A causal relationship between two variables must have a temporal order, in which the cause must precede the effect in time (i.e., if A is a cause and B is an effect, then A must occur before B).
2. The two variables should be empirically correlated with one another.
3. The observed empirical correlation between two variables cannot be explained away as the result of a third variable that causes both A and B. In other words, the relationship is not spurious and occurs regularly.

2.2 Probabilistic approach

- In epidemiology, the deterministic concept of causation is supplemented or replaced with probabilistic methods so that instead of demonstrating causality in individuals, we make causal inferences about a hypothesized relation between a given exposure and a disease in a particular population.

  - Probabilistic – X leads to a distribution of possible outcomes in a population (or a % probability)
  - Deterministic – X leads to an outcome

2.3 Four causal types

1. Type 1: Doomed

- Disease occurs with or without exposure

2. Type 2: Effect causative

- Disease occurs if and only if person exposed

3. Type 3: Effect preventive

- Disease occurs if and only if person unexposed

4. Type 4: Immune

- Disease does not occur with or without exposure

Populations consist of individuals of all four causal types

Causal type	Effect	Disease outcome if exposed	Disease outcome if unexposed
Type 1	No effect (doomed)	Case	Case
Type 2	Effect causative	Case	Noncase
Type 3	Effect preventive	Noncase	Case
Type 4	No effect (immune)	Noncase	Noncase

- The exposure has an effect on the disease in Types 2 & 3, but not on Types 1 & 4.
- Cannot observe these types because we cannot observe both exposure states simultaneously in the same individuals, i.e., one exposure condition is counterfactual.
  - If we observe an exposed person who becomes a case, we cannot determine whether he/she is a Type 1 or 2; If we observe an unexposed person who becomes a case, we cannot whether he/she is a Type 1 or 3.
  - Similarly, if we observe a certain proportion of exposed persons at risk developing a disease in a population, we cannot say how many (if any) were caused by the exposure.

2.4 Four causal inference frameworks

1. Sufficient-Component Cause Model
2. Bradford-Hill's Viewpoints/Criteria
3. Counterfactual (Potential Outcomes) Framework
4. Directed Acyclic Graphs (DAGs)

We use several models of causation in epidemiology because each model has its own strengths and helps us to examine different epidemiologic concepts

2.5 1. Sufficient-component cause model (Rothman's Causal Pie Model)

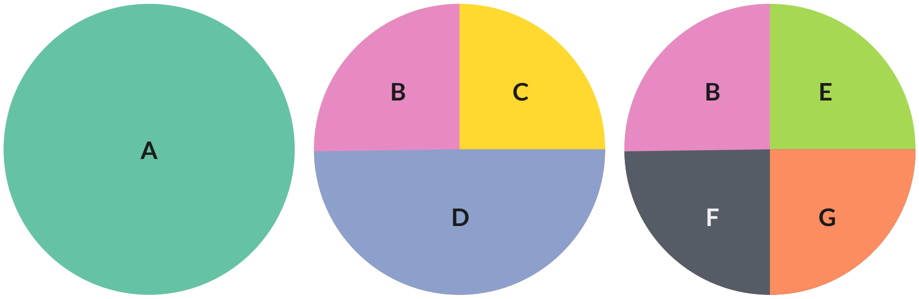

Figure 2.1: Causal pies

- Theoretical mechanistic model
    - Each pie represents a sufficient causal mechanism
    - A **sufficient cause** model represents a minimal set of conditions or events that are sufficient for the outcome to occur
    - Acknowledges the multifactorial etiology of health outcomes

2.5.1 Example

The above figure represents hypothetical sufficient cause models for tuberculosis (TB) infection. U represents unknown causal components. The presence of each causal component implies that their absence (e.g., T=0) would block occurrence of the outcome.

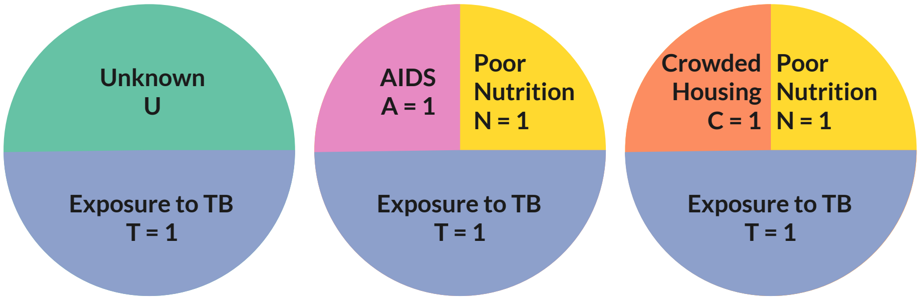
Unknown
U
Exposure to TB
T = 1
AIDS
A = 1
Poor
Nutrition
N = 1
Exposure to TB
T = 1
Crowded
Housing
C = 1
Poor
Nutrition
N = 1
Exposure to TB
T = 1

Figure 2.2: Causal pies example

2.5.1.1 Definitions

- **Component cause**: an event or condition that plays a necessary role in the occurrence of some cases of a given disease
- **Necessary cause**: an event or condition that plays a necessary role in the occurrence of all cases of a given disease; must be present for the disease to occur.
- **Causal complement**: the other factors which are necessary and sufficient condition for a factor to produce disease

2.5.1.2 Benefits of sufficient-cause model

Helps us to understand

- Multifactorial nature of disease causation
- That there is often an unknown (U) contribution to disease causation
    - The composition of U may vary between people/populations
- The fraction of a given disease attributable to its component causes often sums to >100%
    - This is due to the multifactorial nature of disease etiology

2.5.1.3 Criticisms of sufficient-cause model

- More useful on a conceptual basis than in application
- Does not depict:
    - Sequential mechanisms (timing of causal components)
    - Direct vs. indirect effects of component causes

2.6 2. Bradford-Hill's Viewpoints

1. Strength of Association

- The stronger the association the more likely that the relationship is causal (not representing bias or error)
- Observation of a weak observation (RR<~2.0) does not negate the possibility of causality

2. Consistency

- Repeated observation of an association in different populations under different circumstances
- Consistency of results across epidemiologic studies gets at the heart of inductive reasoning and is used to infer causality in observational studies
- Should be viewed cautiously, as it may just reflect consistency of confounding or bias across studies
- Also, may result from publication bias whereby "positive" results are more likely to be published

3. Specificity

- Cause leads to a single effect, not multiple effects
- This criteria is INVALID

4. Temporality

- Exposure must precede the onset of disease
- Essential for causation
- Can be difficult to establish temporality

5. Biological gradient (dose-response relationship)

- Relationship between magnitude (dose) of exposure and risk of outcome is regarded as strong evidence

6. Plausibility

- A scientifically plausible mechanism between exposure and outcome is helpful
- Should be some biological rationale
- Hill noted that knowledge of the mechanism is limited by the current state of evidence

- Plausibility can change with time as knowledge grows

7. Coherence

- Implies a cause-and-effect interpretation for an association does not conflict with what is known of the natural history and biology of disease

8. Experimental evidence

- Laboratory experiments in animals or humans
- Often not ethical or possible to randomly assign exposures – we can't always get experimental data

9. Analogy

- The effect of similar factors may be considered
- Not a useful criterion

2.6.1 3. Neyman-Rubin Counterfactual Framework

- Also called Rubin Causal Model, Potential Outcomes Model, or simply Counterfactual Framework
- Counterfactual - contrary to the observed fact
- Asks the question, would the outcome have been different if the exposure was not present?

Chapter 3

Study design

Material adapted from slides by Jack Jennings, PhD, MPH

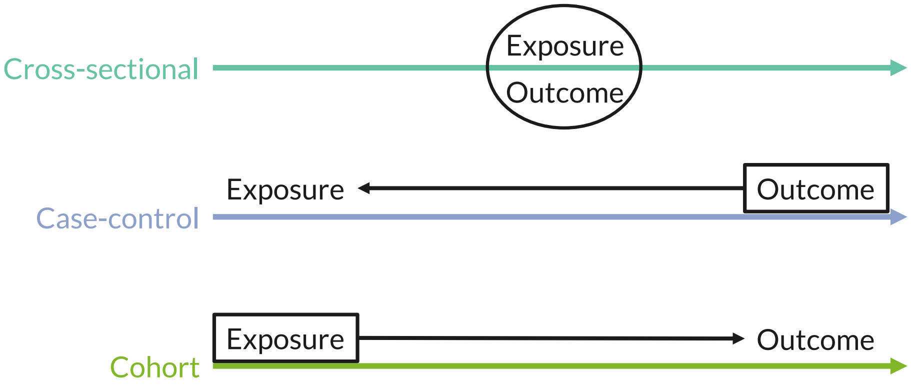
Cross-sectional
Exposure
Outcome
Exposure
Outcome
Case-control
Exposure
Outcome
Cohort

Figure 3.1: Comparison of study designs

3.1 Cross-sectional studies

- Measure existing disease and current exposure levels at one point in time
- Sample without knowledge of exposure or disease
- Ex. prevalence studies

3.1.1 Advantages

- Often early study design in a line of investigation

- Good for hypothesis generation
- Relatively easy, quick and inexpensive … *depends on question*
- Examin multiple exposures or outcomes
- Estimate prevalence of disease and exposures

3.1.2 Disadvantages

- Cannot infer causality
- Prevalent vs. incident disease
- May miss latent disease
- May be subject to recall bias

3.1.3 Example

Research question

- Determine whether there are differences in rates of stroke and myocardial infarction by gender and race among patients

Hypothesis

- There will be differences in rates of stroke by gender and race
- There will be differences in rates of myocardial infarction by gender and race

3.2 Case-control studies

- Identify individuals with existing disease/s and *retrospectively* measure exposure

3.2.1 Advantages

- Good design for rare, chronic and long latency diseases
- Relatively inexpensive (population size and time)
- Allows for the examination of multiple exposures
- Estimate odds ratios
- Hospital-based studies and outbreaks

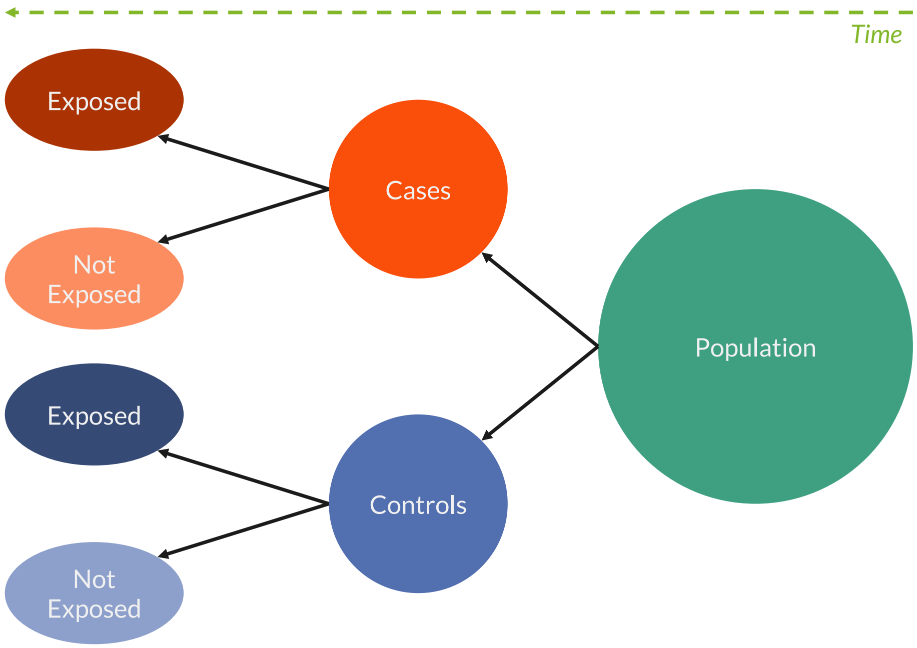
Time
Exposed
Not Exposed
Cases
Population
Exposed
Not Exposed
Controls

Figure 3.2: Case-control schematic

3.2.2 Disadvtanges

- Multiple outcomes cannot be studies
- Recall bias
- Sampling bias
- Cannot calculate prevalence, incidence, population relative risk or attributable risk
- Beware of reverse causation

3.2.3 Challenges

- Selection of controls
  - Sample size
  - Matching (group or individual)
- Selection of cases
  - Incident or prevalent cases

3.2.4 Example

Hypothesis

- Buprenorphine-exposed neonates will exhibit less NAS than methadone-exposed neonates

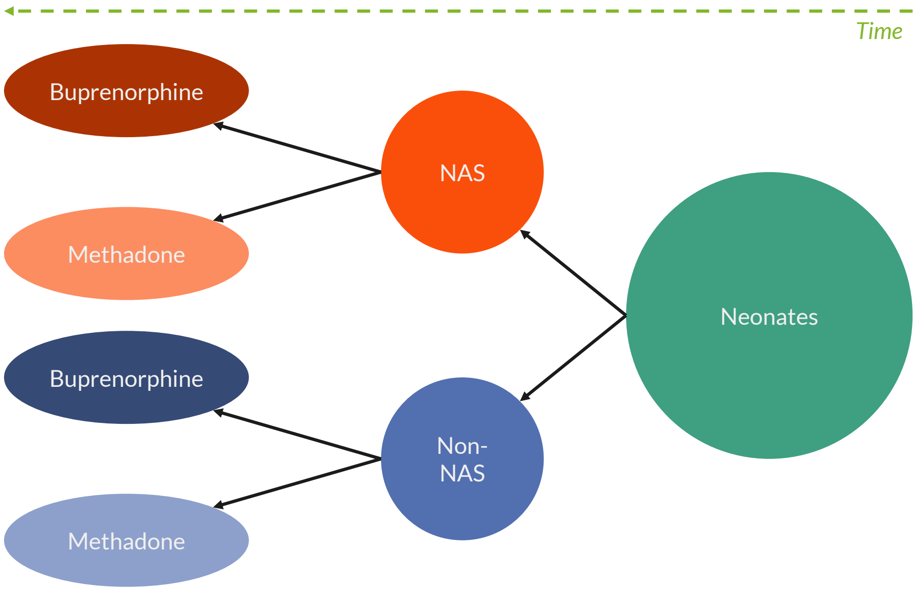
Time
Buprenorphine
Methadone
NAS
Neonates
Buprenorphine
Non-NAS
Methadone

Figure 3.3: Case-control example

3.3 Cohort studies

- Identify exposed and unexposed individuals and follow them over time measuring outcomes/s (prospective)

3.3.1 Prospective vs retrospective cohort studies

- Prospective
  - Study begins before or after exposure but always before collection of the outcome measure
- Retrospective
  - Study begins after collection of the outcome measure

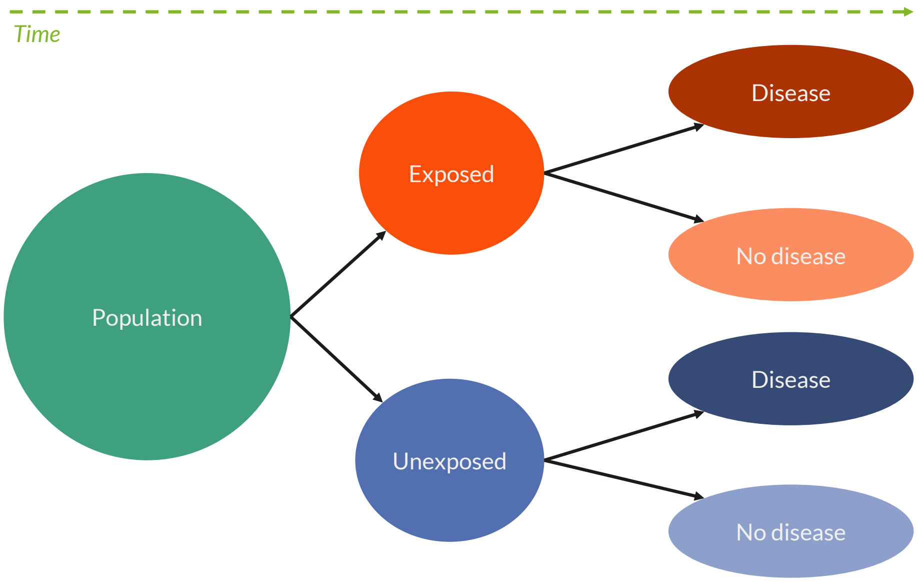
Time
Population
Exposed
Unexposed
Disease
No disease
Disease
No disease

Figure 3.4: Cohort study schematic

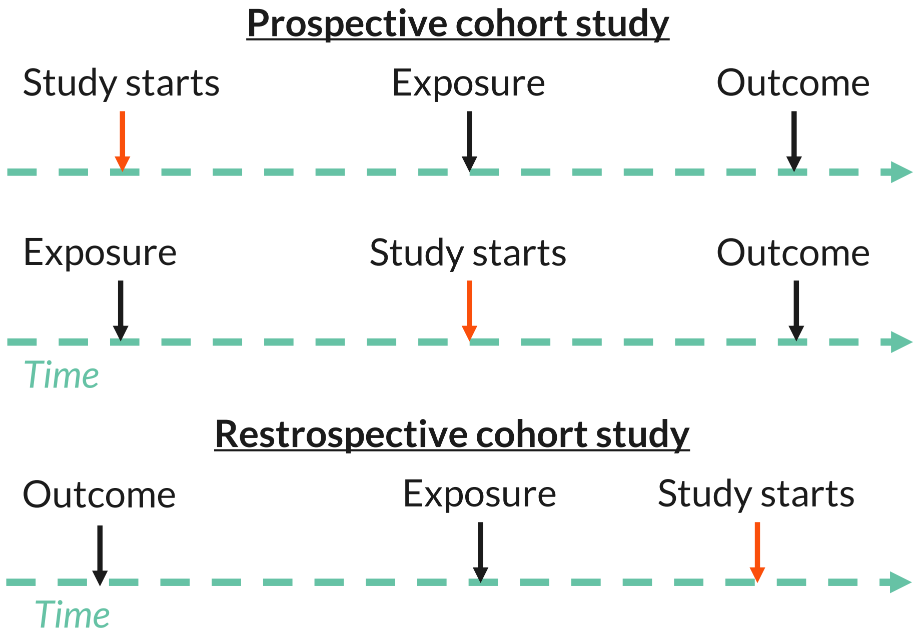
Prospective cohort study
Study starts
Exposure
Outcome
Exposure
Study starts
Outcome
Time
Restrospective cohort study
Outcome
Exposure
Study starts
Time

Figure 3.5: Prospective vs retrospective cohort study

3.3.2 Advantages

- Measure population-based incidence
- Relative risk and risk ratio estimations
- Rare exposures
- Temporality
- Less likely to be subject to biases (recall and selection as compared to case-control)
- Possible to assess multiple exposures and/or outcomes

3.3.3 Disadvantages

- Impractical for rare diseases and diseases with a long latency
- Expensive
  - Often large study populations
  - Time of follow-up
- Biases
  - Design: Sampling, ascertainment, and observer
  - Study population: Non-response, migration, and loss-to-follow-up

3.3.4 Example

Research question

- Determine whether circulating biomarkers (i.e., C-reactive protein; exhaled breath condensate - pH, hydrogen peroxide, 8-isoprostene, nitrie, nitrate levels; sputum - TNF-α, IL-6, UK-8, IL-1β, neutrophil elastase; and fractional exhaled nitric oxide) predict individuals who will benefit from initiation of antibiotic therapy for the treatment of a mild decrease in FEV_1

Hypothesis

- Biomarkers at the time of presentation with a mild increase in pulmonary symptoms or small decline in FEV_1 can be used to identify which patients require antobiotics to recover.

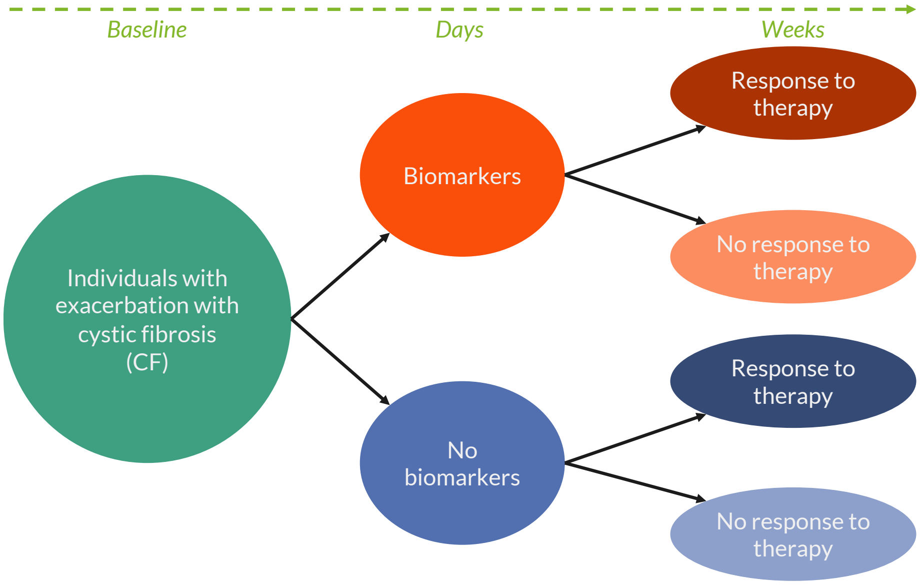
Baseline
Days
Weeks
Individuals with exacerbation with cystic fibrosis (CF)
Biomarkers
No biomarkers
Response to therapy
No response to therapy
Response to therapy
No response to therapy

Figure 3.6: Cohort study example

3.3.5 Important features

- How much selection bias was present?
    - Were only people at risk of the outcome included?
    - Was the exposure clear, specific, and measurable?
    - Were the exposed and unexposed similar in all important respects except for the exposure?
- Were steps taken to minimize information bias?
    - Was the outcome clear, specific, and measurable?
    - Was the outcome identified in the same way for both groups?
    - Was the determination of the outcome made by an observer blinded to treatment?
- How complete were the follow-up of both groups?
    - What efforts were made to limit loss-to-follow-up?
    - Was loss-to-follow-up similar in both groups?
- Were potential confounding factors sought and controlled for in the study design or analysis?
    - Did the investigators anticipate and gather information on potential confounding factors?
    - What methods were used to assess and control for confounding?

3.4 Randomized controlled trials (RCTs)

- Experimental: exposure is assigned
- Randomization assignment
  - Random allocation of exposure or treatment
  - Results (or should result) in two equivalent groups on all measured and unmeasured confounders
  - Gold standard for causal inference

3.4.1 Advantages

- Least subject to biases of all study designs (**IF** designed and implemented well…!)

3.4.2 Disadvantages

- Intent-to-treat
- Loss-to-follow-up
- Randomization issues
- Not all exposures can be "treatments," i.e., are assignable
- Some exposures/treatments cannot be ethically randomized

3.4.3 Example

Research question

- To determine whether resident's attitude and skills in diabetes management and counseling change after a curricular intervention
- To determine whether patient outcomes related to diabetes (e.g., weight, smoking status) change after a curricular intervention among residents

Hypothesis

- Attiudes and skills related to diabetes management and counseling will improve among residents after a curricular intervention
- Fewer patients with diabetes will smoke over time after curricular intervention among residents

3.4.4 Randomization strategies

- Randomly assigned
- Quasi-randomization
- Block randomization
    - Method of randomization that ensures that at any point in the trial, roughly equal numbers of participants have been allocated to the comparison groups

3.5 Study design flowchart

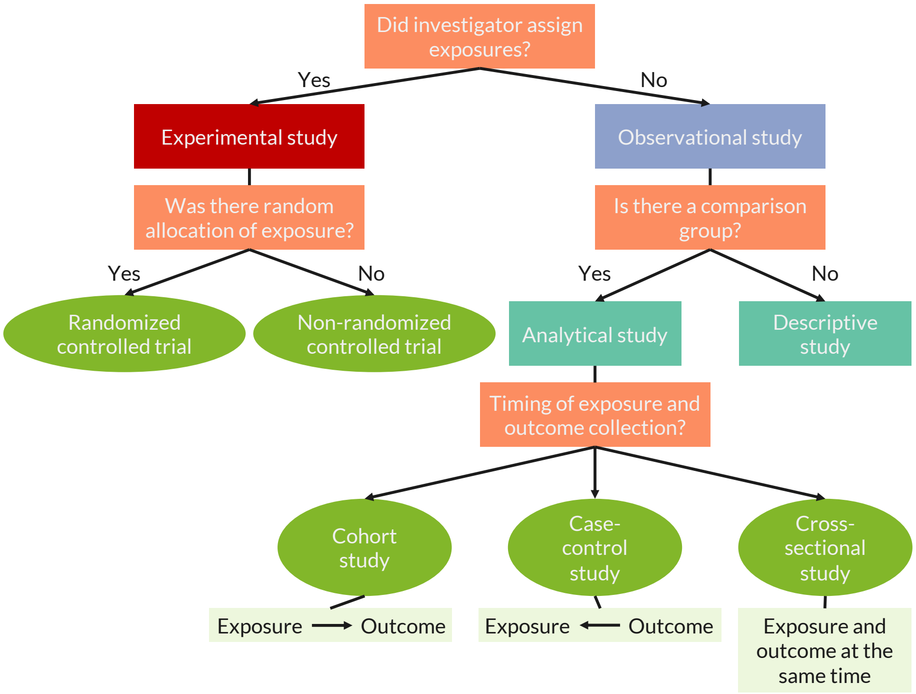
Did investigator assign exposures?
Yes
No
Experimental study
Observational study
Was there random allocation of exposure?
Is there a comparison group?
Yes
No
Yes
No
Randomized controlled trial
Non-randomized controlled trial
Analytical study
Descriptive study
Timing of exposure and outcome collection?
Cohort study
Case-control study
Cross-sectional study
Exposure → Outcome
Exposure ← Outcome
Exposure and outcome at the same time

Figure 3.7: Study design flowchart

Chapter 4

Confounding bias

4.1 Identifying confounding

4.1.1 Data-driven approaches

- There are many data-driven approaches that are used to identify confounding that are based on statistical approaches to examine associations in study data
  - Stepwise regression
    - Throw all suspected confounders into a model and remove those not associated with the outcome (e.g., $p > 0.05$) in a stepwise fashion
  - Change-in-estimate approach (10% rule)
    - Throw all suspected confounders into a model, retain those whose removal changes the exposure -> outcome effect estimate by > 10%

But confounding is about **causal** relationships, thus it is best to identifying confounding by using causal relationships

- The observed data structure and *a prior* theory or knowledge about the suspected data structure are used to identify confounding
- This is better than stepwise regression or the change-in-estimate approach, which use arbitrary rules based on statistical significance

4.1.2 Structural approach

4.1.2.1 Three criteria

A confounding factor must:

1. Be a cause of the outcome under study
2. Be associated with the exposure under study in the source population
3. Must not be caused by the exposure or disease

4.1.2.2 Criteria 1: Confounding factor must cause the outcome (either directly or indirectly)

- A confounding factor must be a cause of the outcome
  - May be an actual cause of the disease
  - May be a surrogate/proxy or indirect cause of the disease
    - Household income as a surrogate for a milieu of social factors correlated with income
    - Education as a proxy for literacy
  - Prior theory or knowledge (not the data itself) is used to determine the relation of the suspected confounding factor to the outcome

4.1.2.3 Criteria 2: Confounding factor must be associated with the exposure in the source population

- We can generally identify this directly from our data, however varies a bit by design
- Cohort Study
  - Cohort is source population. Therefore, this relationship can be determined from the observed study data.
- Case-control Study
  - In a C-C study, the controls are selected from the source population, however, the control group needs to be very large and have no selection bias or measurement error in order to accurately reflect the association in the source population.
  - External information can be used when available, or prior knowledge.

4.1.2.4 Criteria 3: The factor cannot be caused by the exposure or the disease

- A confounding factor must not be affected by the exposure or the disease.

- Two scenarios can occur here:
    1. The factor that we think is a confounder is actually an intermediate on the causal pathway between exposure and outcome (a **mediator**)
    2. The factor that we think is a confounder is a common outcome of the exposure and outcome of interest (a **collider**)

4.1.2.5 Intermediate on the causal pathway (Mediator)

- Here, low birth weight is on the causal pathway from maternal smoking and perinatal mortality

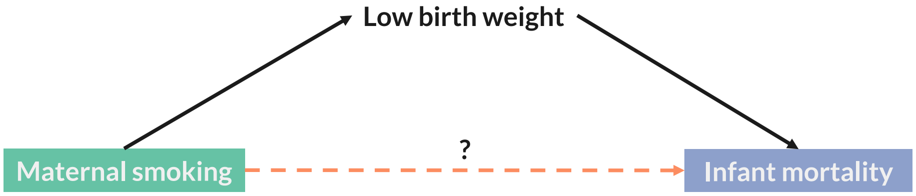

Figure 4.1: Example of mediator

- Low birth weight is a **mediator**. Adjusting for a mediator is referred to as "over-adjustment"

Another example: Fluoridation, Diet sugar, and Tooth decay

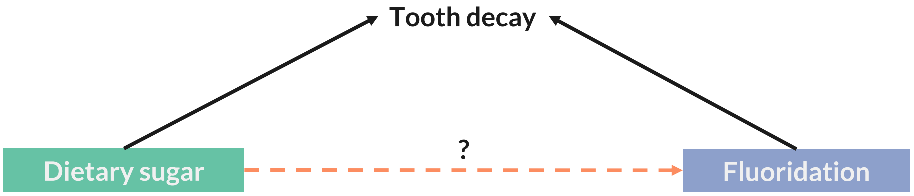

Figure 4.2: Example of a collider

4.1.2.6 Confounding is structural

- Confounding arises because of how, structurally, the variables are related to each other
    - Either how they are naturally related to each other, or, how they are related to each other after an investigator has changed the relationships by adjustmnet, matching, conditioning, etc.

How do we know what the proper structural relationships are?

- Directed Acyclic Graphs (DAGs)
  - NOT a method of data analysis
  - They are used to IDENTIFY confounders based on the assumptions we are willing to make
- DAGs help us to depict the assumed temporal structure of the relationships between our factors
  - Both in the state of nature
  - And after investigators have intervened on the natural structure by conditioning on a set of factors

So

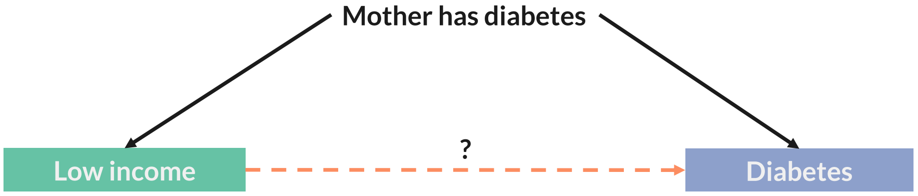

Figure 4.3: Example of confounding

- We want exchangeability of those with low vs. non-low income, condition on having a mother with diabetes (conditional exchangeability)
- The goal is to block all backdoor paths from the exposure to the outcome on the DAG

4.1.2.7 DAG rules for identifying confounding

1. To get from A to Y through a backdoor path, you can move along any path, regardless of the arrow's directionality

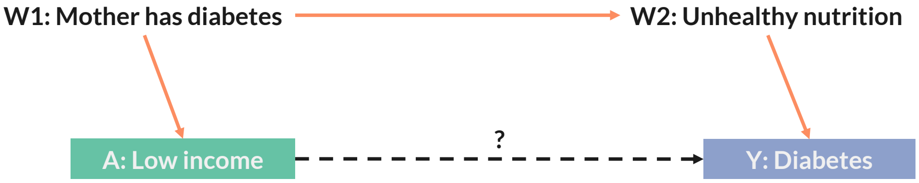

Figure 4.4: An open backdoor path

In the above example (Fig. 2.4), a backdoor path is open through the orange arrows. The effect of the exposure on the outcome is **not** identified.

2. Conditioning on a *common cause of the exposure and outcome* (**confounder**) closes the backdoor path

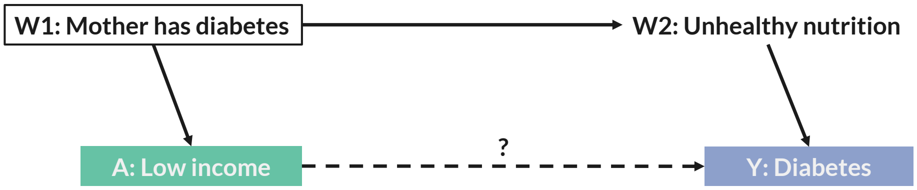
W1: Mother has diabetes
W2: Unhealthy nutrition
A: Low income
?
Y: Diabetes

Figure 4.5: Conditioning on confounder W1 closes backdoor path

In the above example (Fig. 2.5), a backdoor path starting from A is blocked at W1. The effect of A (the exposure) on Y (the outcome) **is** identified.

3. **Unmeasured** factors (U) may still lead to confounding, even if you closed the backdoor path through measured factors

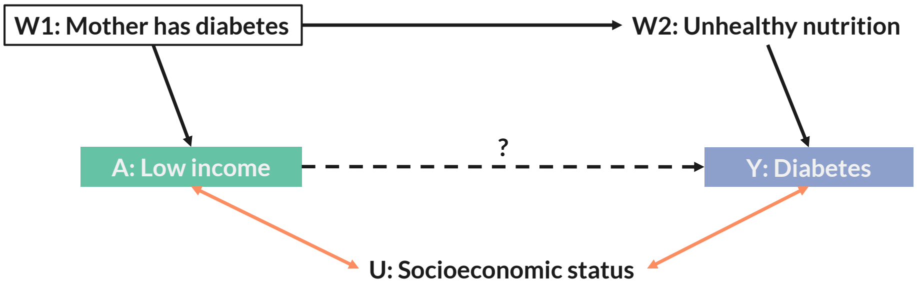
W1: Mother has diabetes
W2: Unhealthy nutrition
A: Low income
?
Y: Diabetes
U: Socioeconomic status

Figure 4.6: Unmeasured factors (U) may still lead to confounding

In the above example (Fig 2.6), a backdoor path is blocked at W1, but it is open through U. The effect of A on Y is **not** identified.

4. The existence of a collider will block the backdoor path

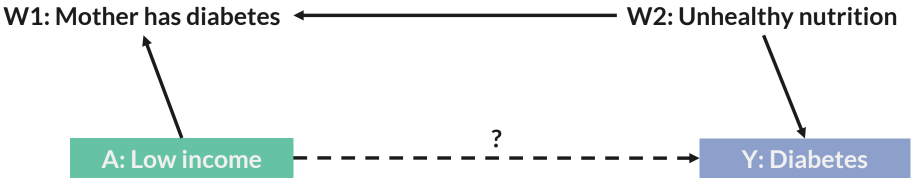
W1: Mother has diabetes
W2: Unhealthy nutrition
A: Low income
?
Y: Diabetes

Figure 4.7: Colliders (W1) block backdoor paths

In the above example (Fig 2.7), a backdoor path starting from A is blocked at W1. The effect of A on Y **is** identified.

5. Condition on a collider will open the backdoor path

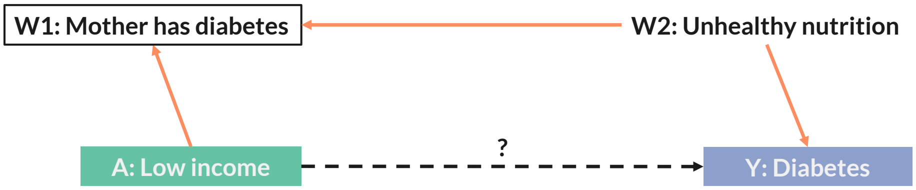
W1: Mother has diabetes
W2: Unhealthy nutrition
A: Low income
?
Y: Diabetes

Figure 4.8: Colliders (W1) block backdoor paths

In the above example (Fig 2.8), a backdoor path starting from A is opened by condition on W1. The effect of A on Y is **not** identified.

6. Conditioning on a descendant (outcome) of a collider will open the backdoor path

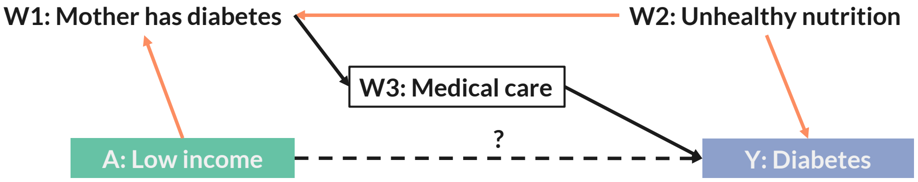
W1: Mother has diabetes
W2: Unhealthy nutrition
W3: Medical care
A: Low income
?
Y: Diabetes

Figure 4.9: Conditioning on descendant of collider opens backdoor path

In the above example (Fig 2.9), a backdoor path starting from A is opened by conditioning on W3. The effect of A on Y is **not** identified.

4.1.2.8 Quantifying confounding

- Non-collapsibility of strata
  - When the association of exposure and outcome is different across the strata of a third variable identified as a confounder and the crude (non-stratified) association, then the data are not collapsible and confounding is present
  - To be continued....

Chapter 5

Selection bias

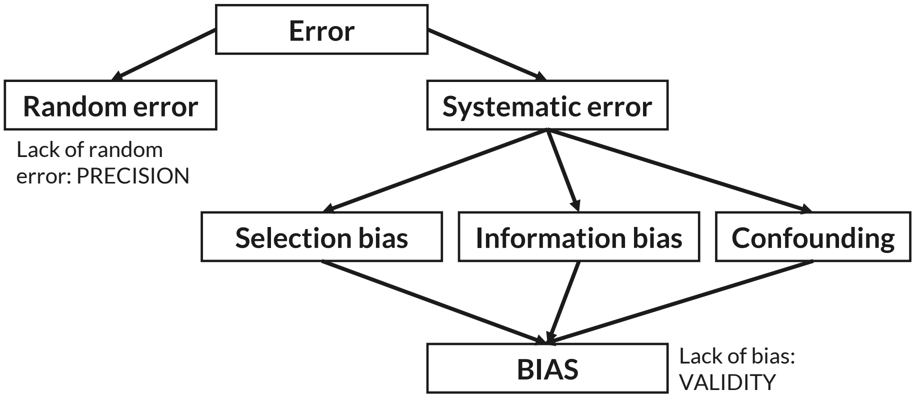
Error
Random error
Lack of random error: PRECISION
Systematic error
Selection bias
Information bias
Confounding
BIAS
Lack of bias: VALIDITY

Figure 5.1: Types of error

There are three broad types of systematic error in epidemiologic studies: selection bias, information bias, and confounding.

Random error differs from systematic error in that its error gets smaller as the sample size (`n`) gets larger. Systematic error does not get better with larger `n`. Selection bias is a type of systematic error.

5.1 Selection bias definitions

Selection bias is:

- Distinct from confounding and information bias because of its mechanisms

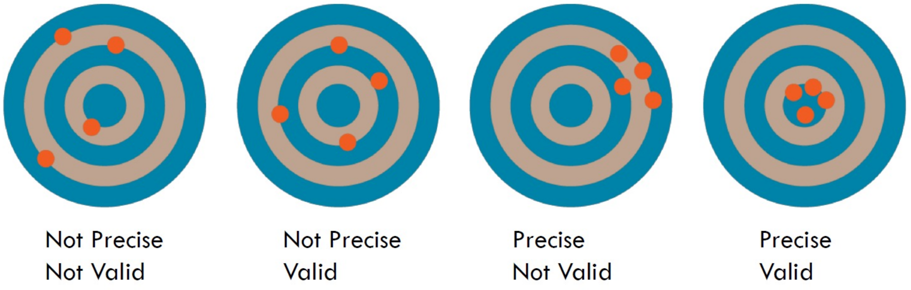

Figure 5.2: Visualization of precision and validity

- Similar to confounding because **exchangeability** is violated
  - $Pr[A = a]$ is not independent of Y^a

5.1.1 Definitions

- Traditional: "Selection bias is present when individuals have different probabilities of being included (or retained) in a study sample according to relevant characteristics, namely the exposure and outcome of interest." (Szklo & Nieto, 2019)
- Structural: "Bias resulting from conditioning on a common effect (a collider) of two variables, one of which is the exposure or associated with the exposure and the other is either the outcome or associated with the outcome." (adapted from Hernán, Hernández-Díaz, & Robbins 2004)

5.1.2 Selection bias in DAGs

5.1.2.1 Brief review of paths

- $A \rightarrow Y$ (A causes Y)
- $A \leftarrow C \rightarrow Y$ (A and Y share a common cause; aka confounding)
- $A \rightarrow \boxed{\text{S}} \leftarrow Y$ (A and Y share a common effect; S is a collider)
  - We must condition on the collider S (adjustment or restriction) or a on a descendant of a collider for us to detect a statistical association between A and Y in this scenario

5.1.2.2 Selection bias DAG examples

There are some more complex ways that selection bias can be captured in a DAG.

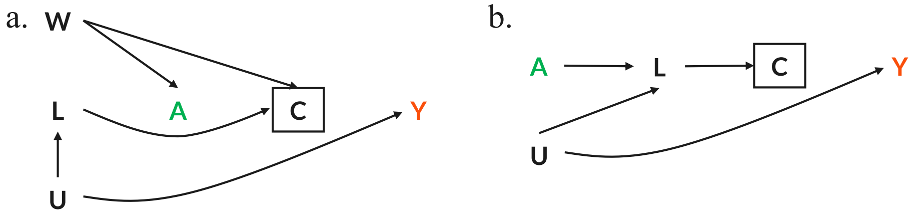

Figure 5.3: Examples of selection bias in DAGs

In Fig. 3.3.a., conditioning on the collider, C, opens a backdoor pathway between the exposure, A, and the outcome, Y. This path is $A \leftarrow W \rightarrow \boxed{\text{C}} \leftarrow L \leftarrow U \rightarrow Y$.

In Fig. 3.3.b., conditioning on C, the descendant of the collider, L, opens a backdoor path.

Here is a motivational example that displays the intuition behind a collider.

Rain → $\boxed{\text{Wet sidewalk}}$ ← Neighbor's sprinkler

Take, for example, the question: did it rain last night? Suppose we only observe when the sidewalk is wet. The sidewalk could be wet for two reasons: (1) it rained or (2) your neighbor ran the sprinkler. If we know the sidewalk is wet and that the neighbor's sprinkler is broken, then it probably rained. Knowing information about one of the causes of a collider gives us information about the other.

5.1.2.3 Traditional vs. structural definition

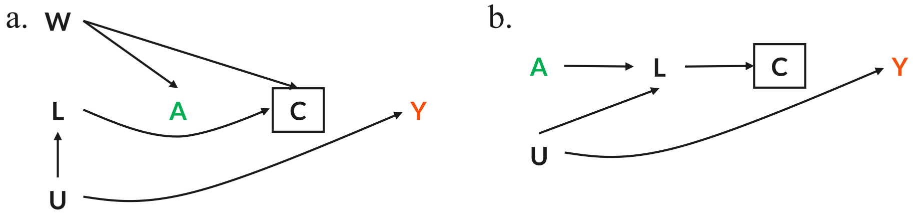

Figure 5.4: Examples of selection bias in DAGs

The above figure graphically depicts the traditional (Fig 3.4.a) and structural (Fig 3.4.b) definitions of selection bias.

5.1.3 Examples of selection bias

- There are many ways a study can be subject to selection bias

- Learning about well accepted types of selection bias can help with finding it in our own studies
- Differs somewhat by study design: case-control, cohort, and RCT

5.1.3.1 Selection bias in case-control studies

- "Berkson's bias"
    - Particularly relevant for hospital-based case-control studies
    - Occurs when **controls are not selected independent of exposure**
    - Case-control studies are thought to be particularly susceptible to selection bias because at the very least $Y \rightarrow \boxed{\text{S}}$

Take the DAG in Fig. 3.4.b to represent a hospital-based case-control study of the malnutrition (A) on depression (Y). Because of the study design, those with depression (Y) are more likely to be included in the study (C). Malnutrition by itself is likely to cause somebody to be admitted to the hospital (C). This is called "Berkson's Bias."

5.1.3.2 Selection bias in cohort studies

- Selection bias in cohort studies typically arises because of **loss to follow-up** or mortality related to both the outcome and the exposure.

Take the DAG in Fig. 3.3.a to represent a cohort study of occupational exposure (A) on risk of stroke (Y). The terrible work conditions of the job (W) is a common cause of exposure (A) and likelihood that a person will quit (C). Underlying health status (U) is a cause of stroke (Y) and causes a person to quit (C) through deteriorating physical health (L). This is also called the "healthy worker effect."

5.1.3.3 Selection bias in cross-sectional studies

- Cross-sectional studies are also susceptible to selection bias.
- Sometimes this is referred to as "incidence-prevalence bias."
- Prevalent cases with better prognosis or underlying health are more likely to show up in your study.

Take the DAG in Fig. 3.4.b to represent a study of the effect of folic acid (A) at conception on the prevalence of birth defects (Y). Only those babies born were included in the study (C).

5.1.3.3.1 Selection bias in randomized control trials (RCT)

- RCTs are often thought to be the “gold standard” of causal inference
- While they are less likely to be subject to confounding, they are equally likely to be subject to selection bias due to loss to follow-up.

Take the DAG in Fig. 3.3.b to represent a study of AZT (A) on development on AIDS (Y). Treatment A and illness severity (U) both cause side effects (L) which leads to dropout (C).

5.2 Key points

- Selection bias is a type of systematic error related to recruitment or retention of participants.
- Recruitment or retention must be related to exposure and outcome to cause bias.
- We can visualize selection bias on DAGs.
- All types of studies are subject to selection bias but it might look different depending on the study.

Chapter 6

Regression

Material excerpted and adapted from Bender 2009 and epiRhandbook

One of the most important analysis related methods to deal with confounding is multiple regression analysis.

The most important methods are:

- Linear regression for continuous outcomes
- Logistic regression for binary outcomes
- Cox regression for time-to-event data, and
- Poisson regression for frequencies and rates

Other methods are, for example, stratification, the use of instrumental variables, or the application of propensity scores. In this chapter, an overview of the most important multiple regression models is given with a focus on applications in modern epidemiology.

6.1 Linear regression

The use of simple linear regression is possible if the effect of one continuous explanatory variable X, for example, Body mass index measured in kilograms per square meter on one continuous response variable Y, for example, systolic blood pressure measured in millimeters of Hg is to be investigated. The fundamental model equation is given by

$$Y = \beta_0 + \beta_1 x + \epsilon$$

where β_0 is the intercept, β_1 is the regression coefficient for X, x is the observed value for X, and ϵ is the random error describing individual deviations from the mean of Y given $X = x$, also called the residual term.

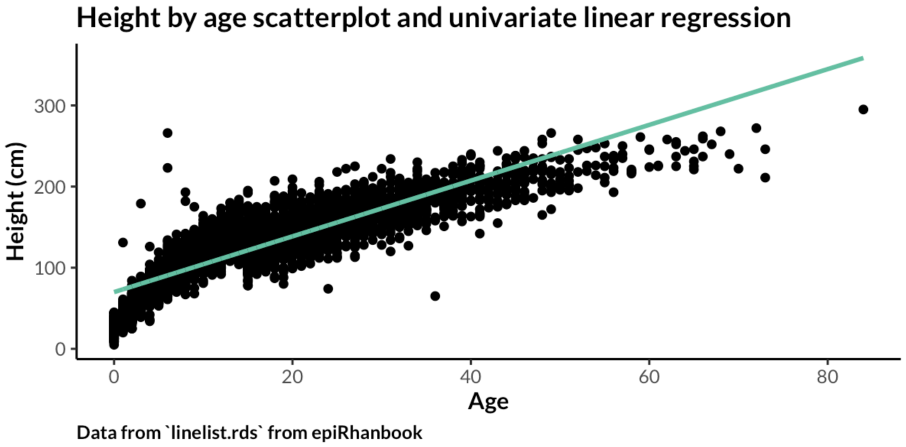

Figure 6.1: An example of a univariate regression analysis

The ordinary least squares estimates of the parameters β_1 and β_0 are given by

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

and

$$\beta_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

where $\bar{x}$ and $\bar{y}$ are the arithmetic means of x_i and y_i, $i = 1, ..., n$, respectively.

In summary, the goal of regression models are threefold:

1. To determine whether the variables Y and X are systematically related (test of $H_0 : \beta_1 = 0$).
2. To estimate the effect size of X on Y by means of $\hat{\beta}_1$ (complemented by a 95% confidence interval)
3. To predict the expected value of Y for given values of X (with 95% confidence interval)

6.1.1 Multiple linear regression

Simple linear regression plays a negligible role in epidemiology for two reasons:

1. Binary (logistic) and time-to-event (Cox) outcomes are much more common than continuous response variables
2. The effects of **confounders** have to be taken into account, so that multiple linear regression should be applied even if we are interested mainly in the effect of one primary explanatory variable. The multiple linear regression model is an extension of the univariate linear regression model where instead of only one explanatory variable X several explanatory variables $X_1, ..., X_k$ with observed values $x_1, ..., x_k$ are considered. The fundamental model equation is given by

$$Y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + \epsilon$$

As mentioned above, we can test whether th explanatory variable $X_j (j = 1, ..., k)$ has a significant effect on the response Y, we can describe the estimated effect size of X_j by means of $\hat{\beta}_j$ for $j = 1, ..., k$ (with CIs), and we can predict the expected value of Y for given values of $X_1, ..., X_k$ (with CIs).

This model can describe the effects of several explanatory variables simultaneously and the regression coefficient β_j represents the effect for X_j adjusted for all other explanatory variables. In general, it is misleading to assess the associations of the response Y and several explanatory variables by means of several simple linear regression models.

6.1.2 Assumptions

There are four assumptions for conducting linear regression:

1. Linear relationship
2. Independence
3. Homoscedasticity
4. Normality ….

Chapter 7

Literature

Here is a review of existing methods.