Data collection 2/2

The ASTA team

Contents

1 Important take-home messages

1.1 Important take-home messages

- Population vs sample:
  - What is the population?
  - Is the entire population known – is statistics at all needed?
- Sampling
  - Sampling strategy must ensure random sampling
    - Difficult to investigate it afterwards
  - Convenience sampling often used, dangerous!
  - Be honest with yourself, describe problems: Is the sample representative for the target group/population/market segment/. . . ?
- Badly chosen big sample is much worse than a well-chosen small sample
- Watch out for biases
  - Sample/selection bias

    - Response bias
    - Non-response bias
    - (Survivorship bias)
- Data collection
    - Privacy vs necessary information (< 50 or >= 50, age in years, birth date)

2 Brief overview of terminology

2.1 Controlling (for)

- Multivariate analysis: "Controlled (for)" means that it's influence is removed
    - Size of effect often not of interest
    - Module 4: Cadmium exposure's effect on vital capacity, controlled for age
- Randomized experiments vs observational studies
- Example [A] 10.1

2.2 Confounders

- Which variables to control for?
- Effect on response variable cannot be distinguished from another (or more) of the explanatory variables
- Variables affecting the association studied, but not measured are sometimes called *lurky*
- Example: correlation between college GPA and income later in life
    - Potential lurking variables: IQ, tendency to work (hard), …
- Example:
    - Plant cucumbers in a garden, some in sun some in shade.
    - Add fertilizer to those in sun.
    - Wait…
    - More cucumbers on those in sun: due to sun light or fertilizer?
    - Effect of fertilizer confounded with effect of sun light.
- Example:
    - Ice cream sale increases with number of shark attacks
    - Weather probably (!) has an impact?
- Analyze effect of explanatory variable: not observe a confounder explaining major part of effect
    - **Omitted variable bias**

2.3 Multicolinearity

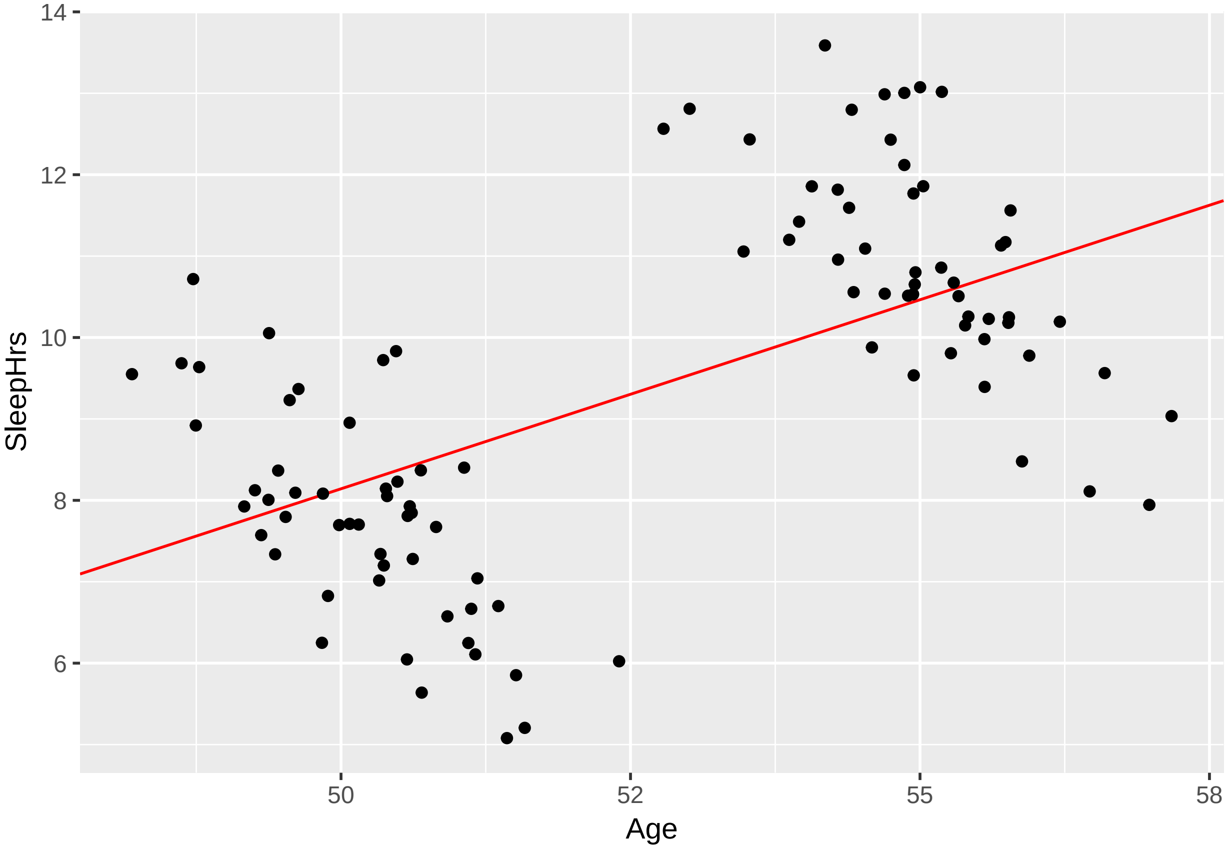
- If one or more explanatory variables are linearly dependent (or close to)

2.4 Simpsons "paradox"

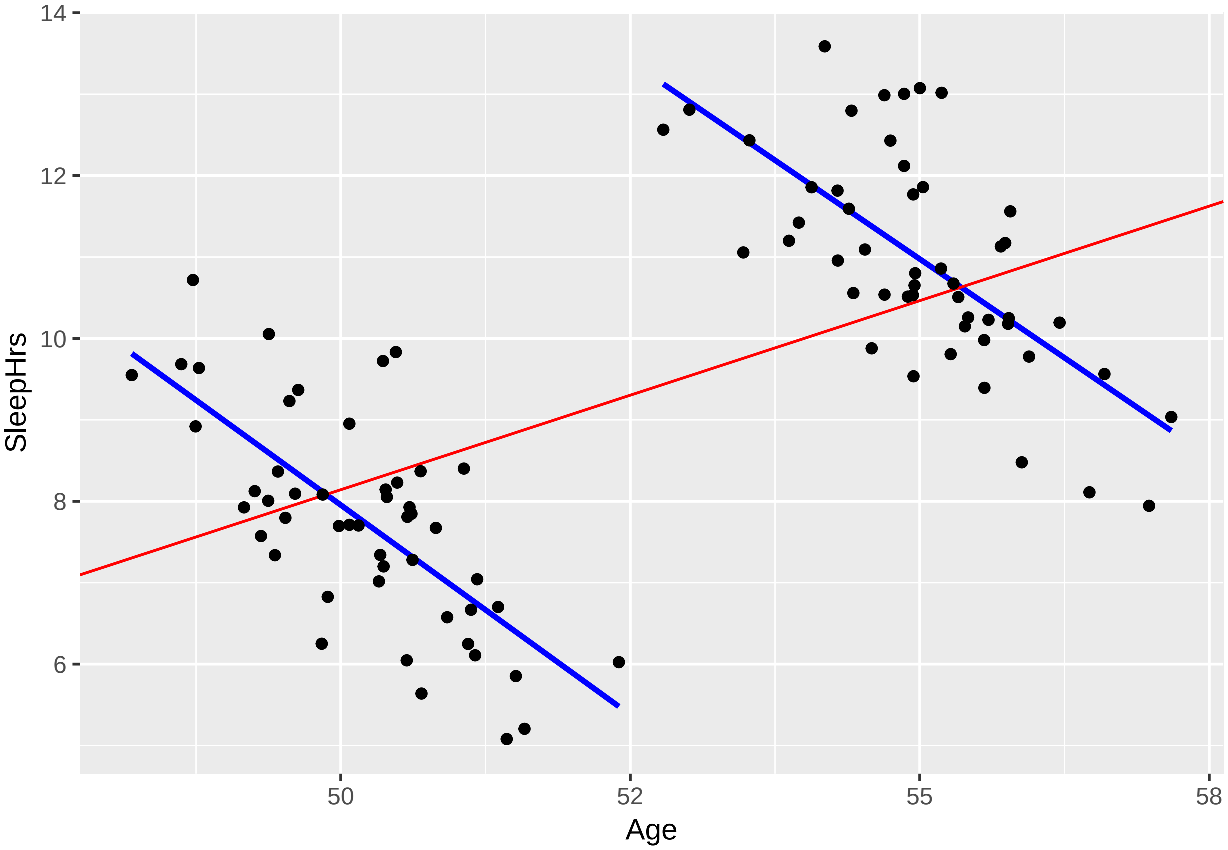
```
mylm <- lm(SleepHrs ~ Age, data = DF)
summary(mylm)
```

```
##
## Call:
## lm(formula = SleepHrs ~ Age, data = DF)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.728 -0.917 -0.102  1.338  3.505
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -15.0791     3.4825   -4.33  3.6e-05 ***
## Age           0.4644     0.0661    7.02  2.9e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.7 on 98 degrees of freedom
## Multiple R-squared:  0.335,  Adjusted R-squared:  0.328
## F-statistic: 49.3 on 1 and 98 DF,  p-value: 2.86e-10
```

2.5 Simpsons “paradox”



2.6 Simpsons “paradox”



2.7 Summary

- Some terms introduced, a lot more to it – but gives some ideas of potential problems

3 Data wrangling

3.1 Data wrangling

Read data:

- `rio`: A Swiss-Army Knife for Data I/O
  - `rio`: A Swiss-Army Knife for Data I/O
  - Excel: `readxl` (part of `rio`)
- R for Data Science

4 Case-study

4.1 Case: Questionnaire about biking habits in Region Sjælland

- Questionnaire:

    - Shared in approx 30 different Facebook groups

- Questions:

    - Representative for the entire region?

        - Each municipality represented in sample proportional to its population size?
        - Disabled people?
        - People biking (municipalities' age distribution may vary)

- Important take-home messages:

    - Sampling strategy must ensure random sampling

        - Difficult to investigate it afterwards

    - Convenience sampling often used, dangerous!

4.2 Analysis

Demo