Quantitative Research Methods in Climate Change and Sustainability Science

Nguyen Bich Ngoc

VNU School of Interdisciplinary Studies

Practical Information

- Theoretical session: Sunday 23/04 & Sunday 07/05/2023 morning
- Practical session: Sunday 07/05/2023 afternoon
  - R
  - Laptop with R & RStudio installed
- Exam:
  - Exercises: 07/05/2023
  - Oral exam: 10 minutes on 20-23/05/2023 (MS Teams/Zoom)

Expected outcomes

- Focus on applied statistics
- Basic concepts/theory
- Identify problems and corresponding solutions (names)

Recommended readings

- Applied statistics with R – David Dalpiaz (https://book.stat420.org/)
- Cẩm nang nghiên cứu khoa học: từ ý tưởng đến công bố – Nguyễn Văn Tuấn (2nd edition, 2020)
- Từng bước nhập môn nghiên cứu khoa học xã hội – Phạm Hiệp & cộng sự (2022)
- Research design: qualitative, quantitative, and mixed methods approaches – John W. Creswell & J. David Creswell (5th edition, 2018)
- Fundamentals of data visualization – Claus O. Wilke (https://clauswilke.com/dataviz/index.html)
- Introduction to quantitative research methods: an investigative approach – Mark Balnaves and Peter Caputi (2001)

Content

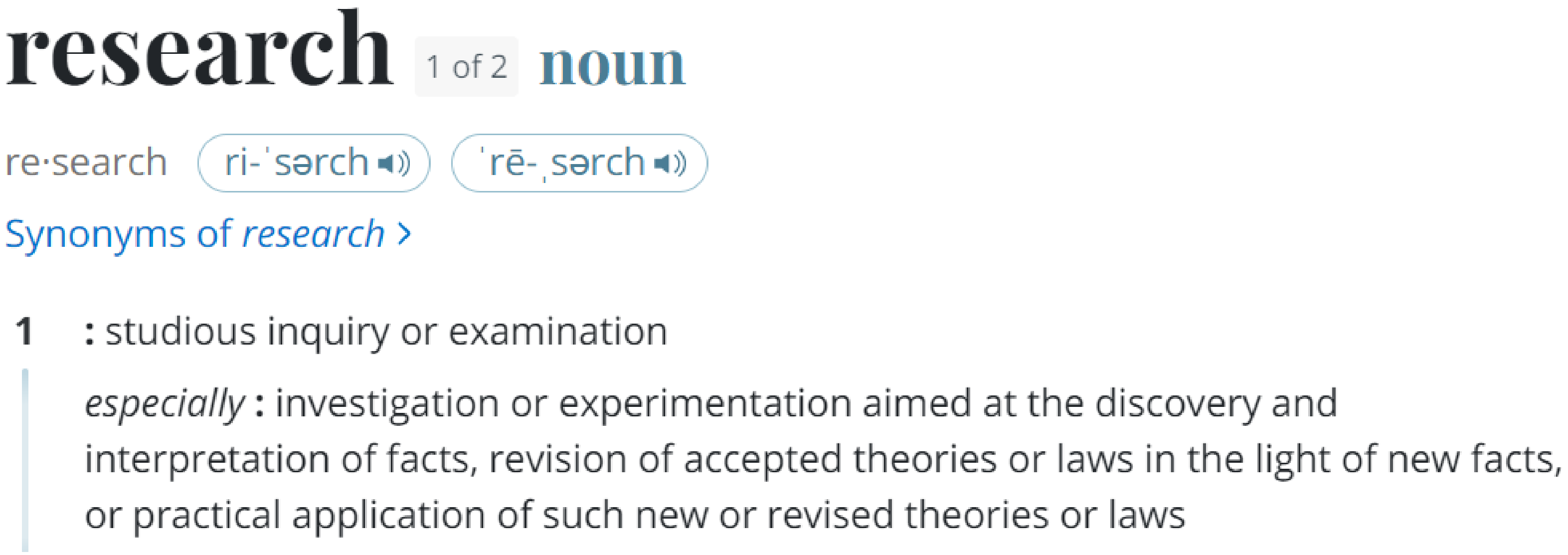
- Research: why, what, & how?
- What is data and how to collect?
- Data plotting and cleaning
- Descriptive statistics
- Inferential statistics
- Advanced topics in quantitative methods

Research: Why, what, & how?

Why?

What?

What?



research 1 of 2 **noun**

re·search ri-ˈsərch ˈrē-ˌsərch

Synonyms of *research* ›

1 : studious inquiry or examination

especially : investigation or experimentation aimed at the discovery and interpretation of facts, revision of accepted theories or laws in the light of new facts, or practical application of such new or revised theories or laws

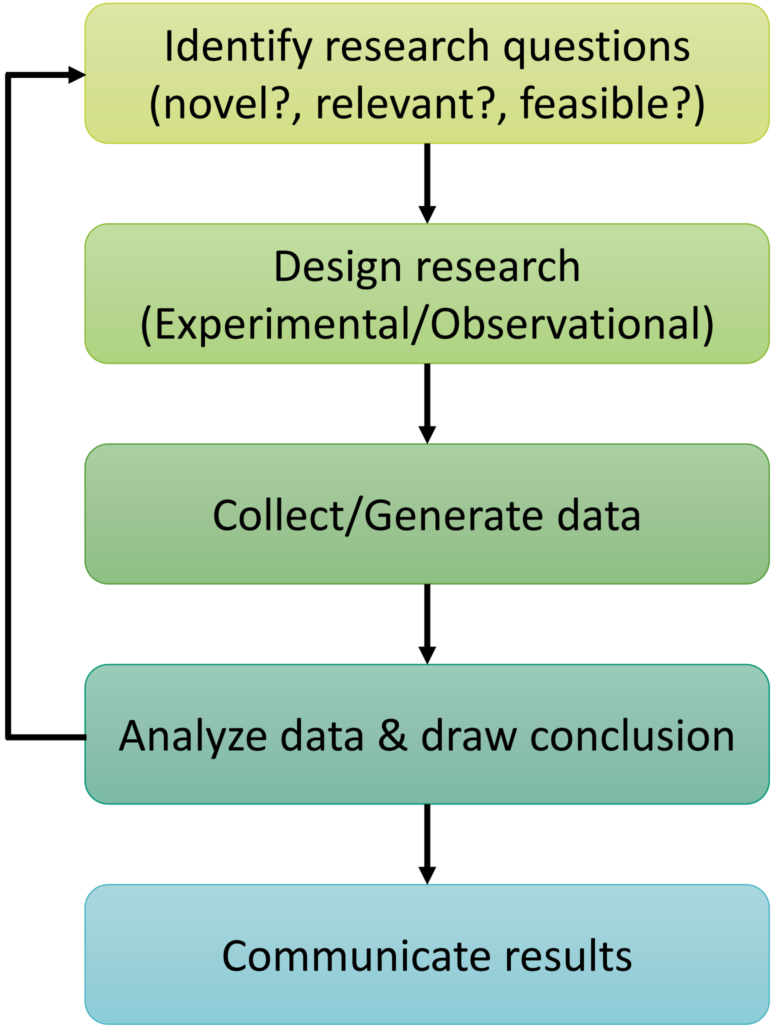
(Merriam-Webster Dictionary)

Quantitative research is an approach for testing objective theories by examining the relationship among variables. These variables, in turn, can be measured, typically on instruments, so that numbered data can be analyzed using statistical procedures.

(John W. Creswel)

How?

How?



Identify research questions (novel?, relevant?, feasible?)

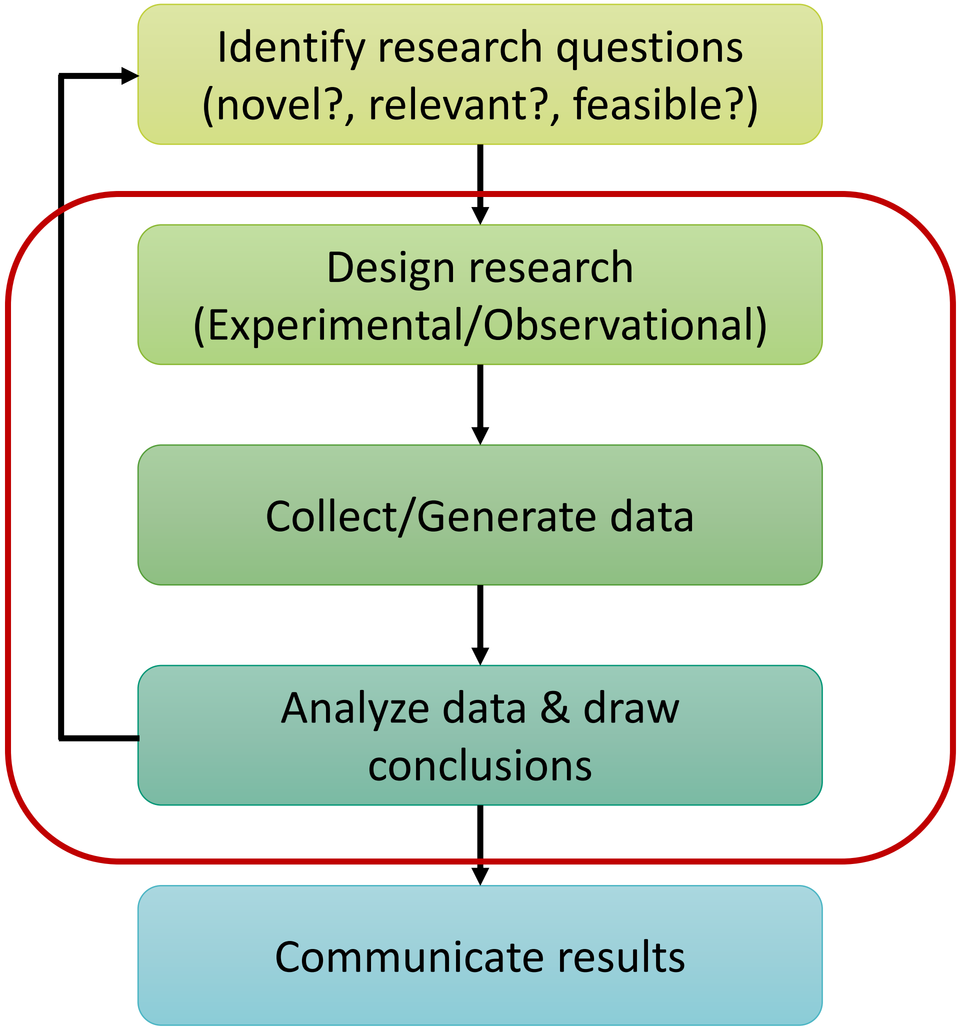
Design research (Experimental/Observational)

Collect/Generate data

Analyze data & draw conclusion

Communicate results

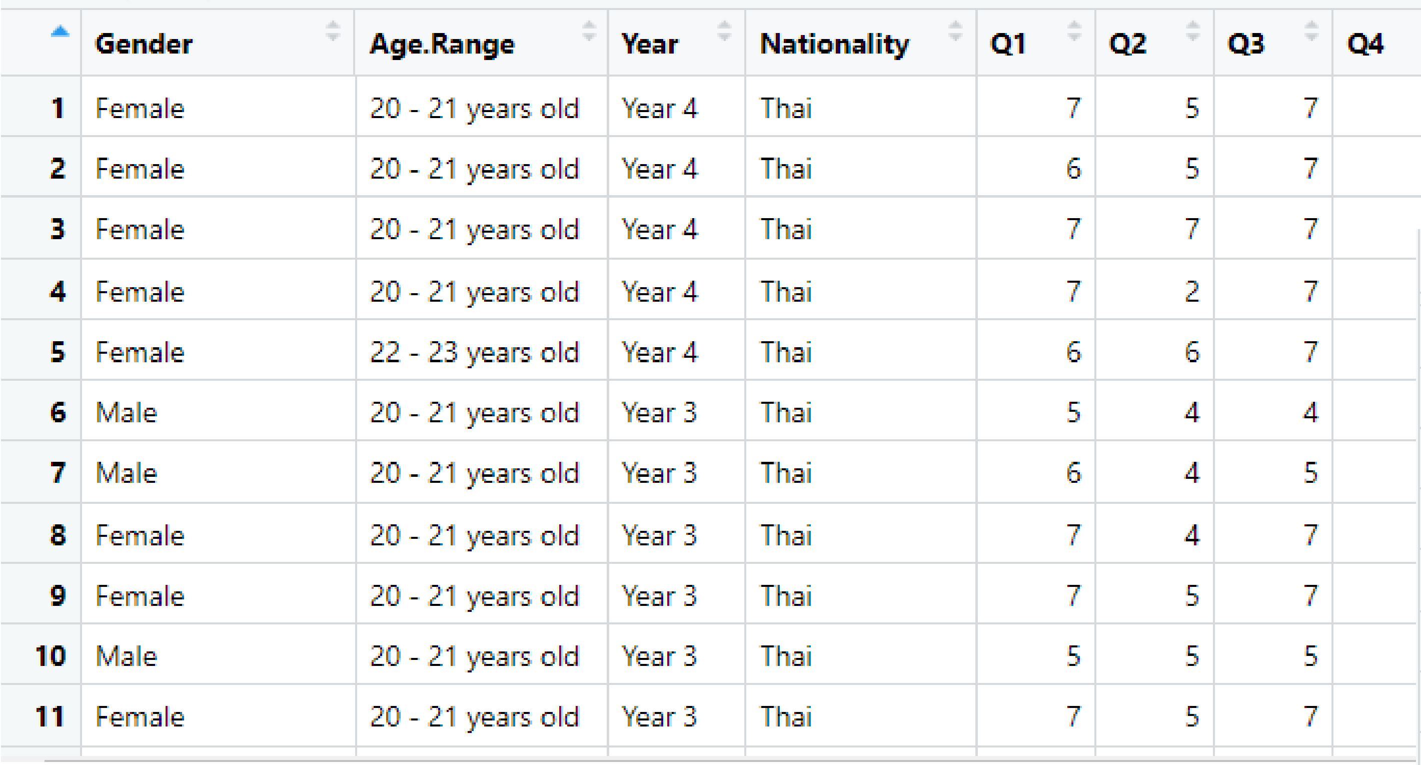
How?


Identify research questions
(novel?, relevant?, feasible?)
Design research
(Experimental/Observational)
Collect/Generate data
Analyze data & draw conclusions
Communicate results

What is data and how to collect?

- Data, Variables, Distributions
- Sample and Sampling methods
- Survey design
- Experimental design

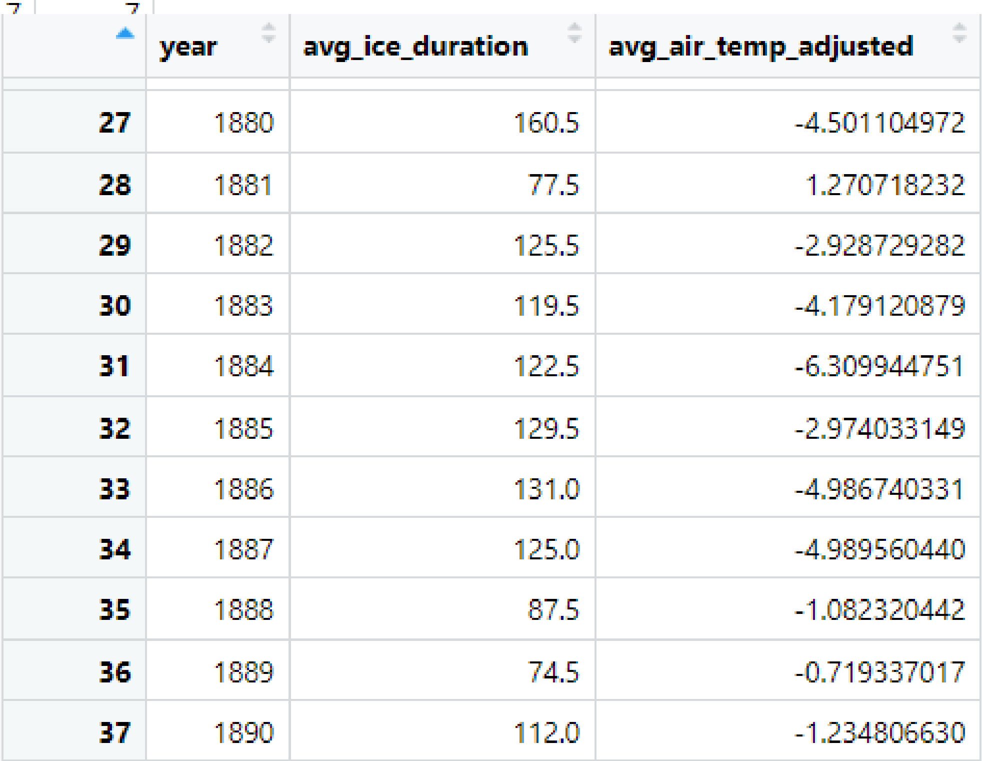
Data



	Gender	Age.Range	Year	Nationality	Q1	Q2	Q3	Q4	Q5
1	Female	20 - 21 years old	Year 4	Thai	7	5	7	7	7
2	Female	20 - 21 years old	Year 4	Thai	6	5	7	5	6
3	Female	20 - 21 years old	Year 4	Thai	7	7	7	7	7
4	Female	20 - 21 years old	Year 4	Thai	7	2	7		
5	Female	22 - 23 years old	Year 4	Thai	6	6	7		
6	Male	20 - 21 years old	Year 3	Thai	5	4	4		
7	Male	20 - 21 years old	Year 3	Thai	6	4	5		
8	Female	20 - 21 years old	Year 3	Thai	7	4	7		
9	Female	20 - 21 years old	Year 3	Thai	7	5	7		
10	Male	20 - 21 years old	Year 3	Thai	5	5	5		
11	Female	20 - 21 years old	Year 3	Thai	7	5	7		

Variables

Observations/Units



	year	avg_ice_duration	avg_air_temp_adjusted
27	1880	160.5	-4.501104972
28	1881	77.5	1.270718232
29	1882	125.5	-2.928729282
30	1883	119.5	-4.179120879
31	1884	122.5	-6.309944751
32	1885	129.5	-2.974033149
33	1886	131.0	-4.986740331
34	1887	125.0	-4.989560440
35	1888	87.5	-1.082320442
36	1889	74.5	-0.719337017
37	1890	112.0	-1.234806630

Variables

- Levels of measurement

  Nominal, Ordinal, Interval, Ratio

- Possible received values

  Categorical, Discrete, Continuous

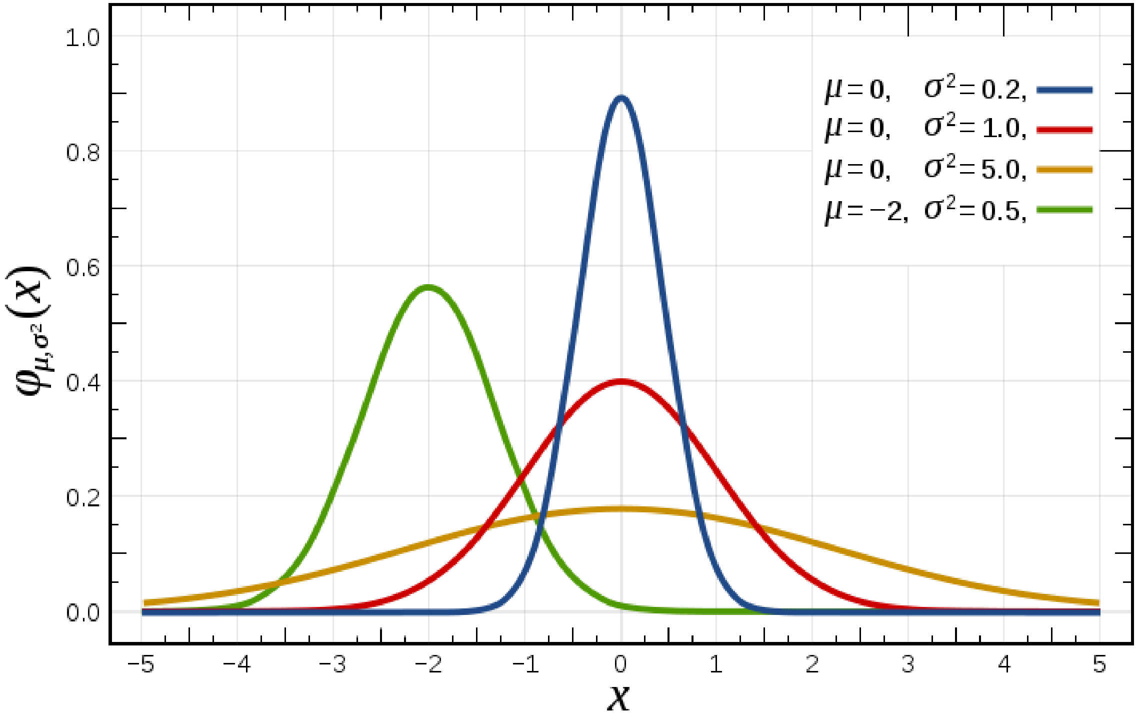
- Relationships

  Independent/Explanatory, Dependent/Outcome, Controlled

Common distributions

- Normal distribution $N(\mu, \sigma^2)$

  Standard normal distribution $N(0,1\ )$

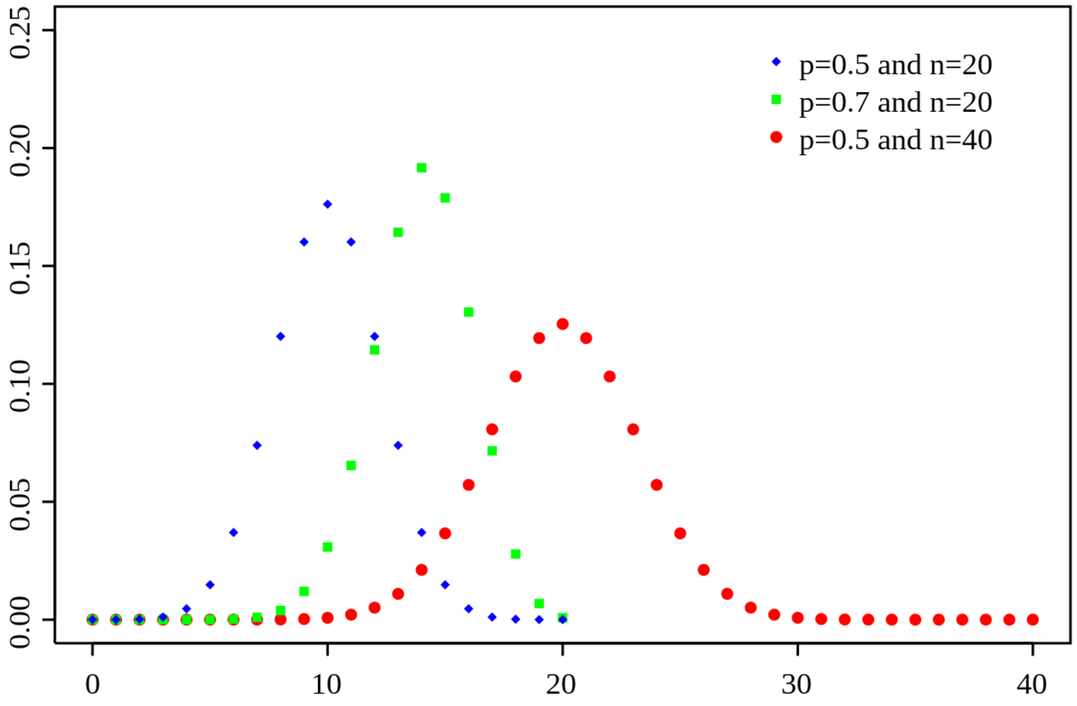


Common distributions

- Normal distribution $N(\mu, \sigma^2)$

  Standard normal distribution $N(0,1\ )$

- Binomial distribution $B(n, p)$

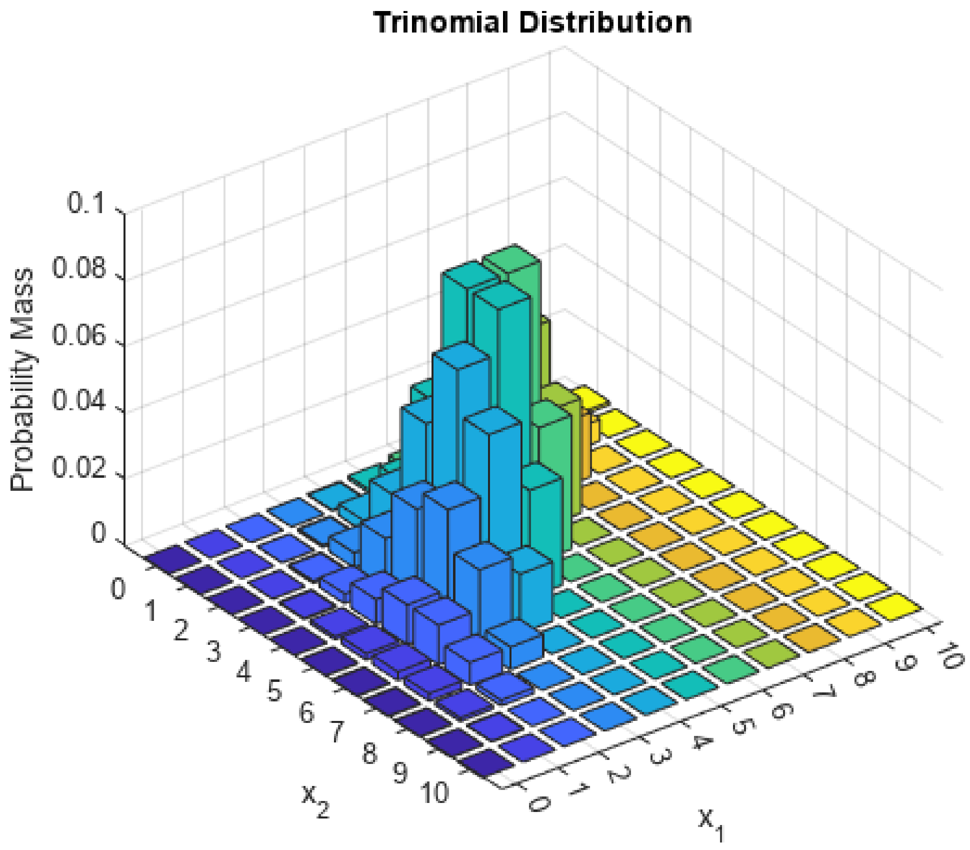


Common distributions

- Normal distribution $N(\mu, \sigma^2)$

  Standard normal distribution $N(0,1\ )$

- Binomial distribution $B(n, p)$
- Multinomial distribution

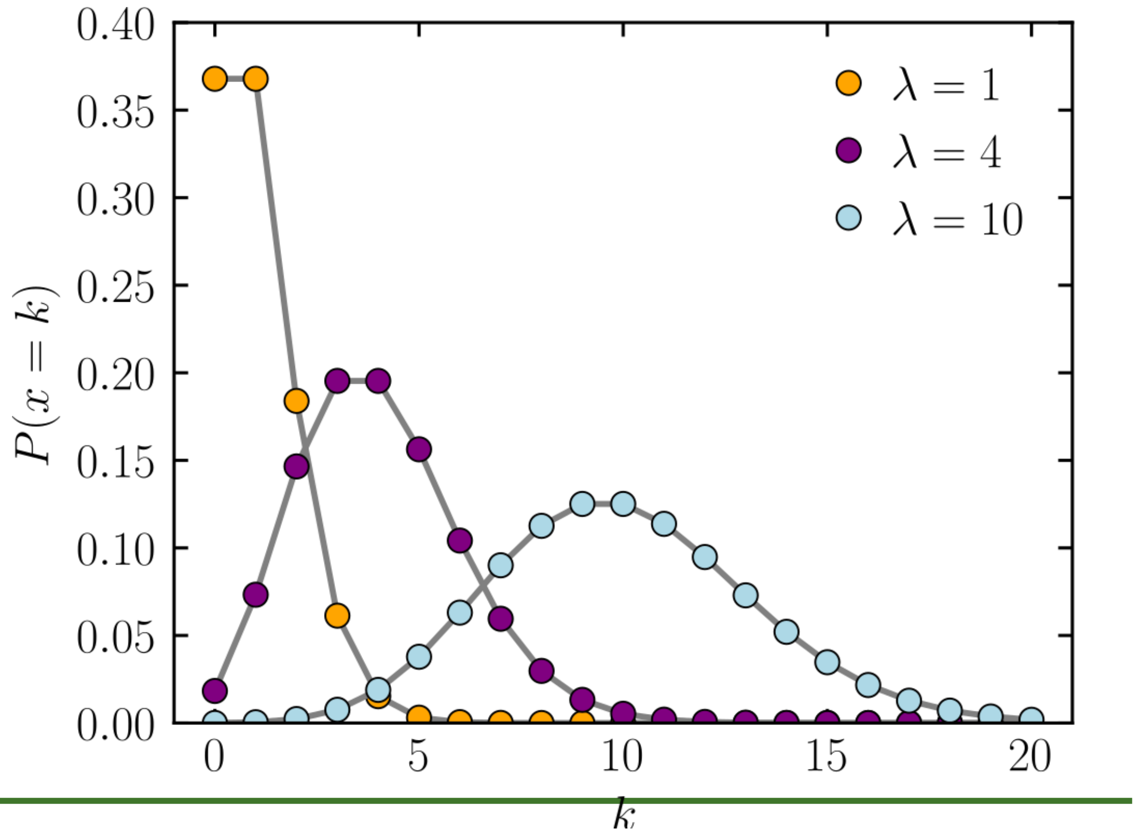


Common distributions

- Normal distribution $N(\mu, \sigma^2)$

  Standard normal distribution $N(0{,}1\ )$

- Binomial distribution $B(n, p)$
- Multinomial distribution
- Poisson distribution $Pois(\lambda)$



Data sources

Primary

Experiment

Survey/questionnaire

Observation/Measurement

Secondary

Open-source database

Government publications

Internal reports

Sample and sampling methods

- Sub-population
- Representative?
- Sample size
    - Too small?
    - Too large?

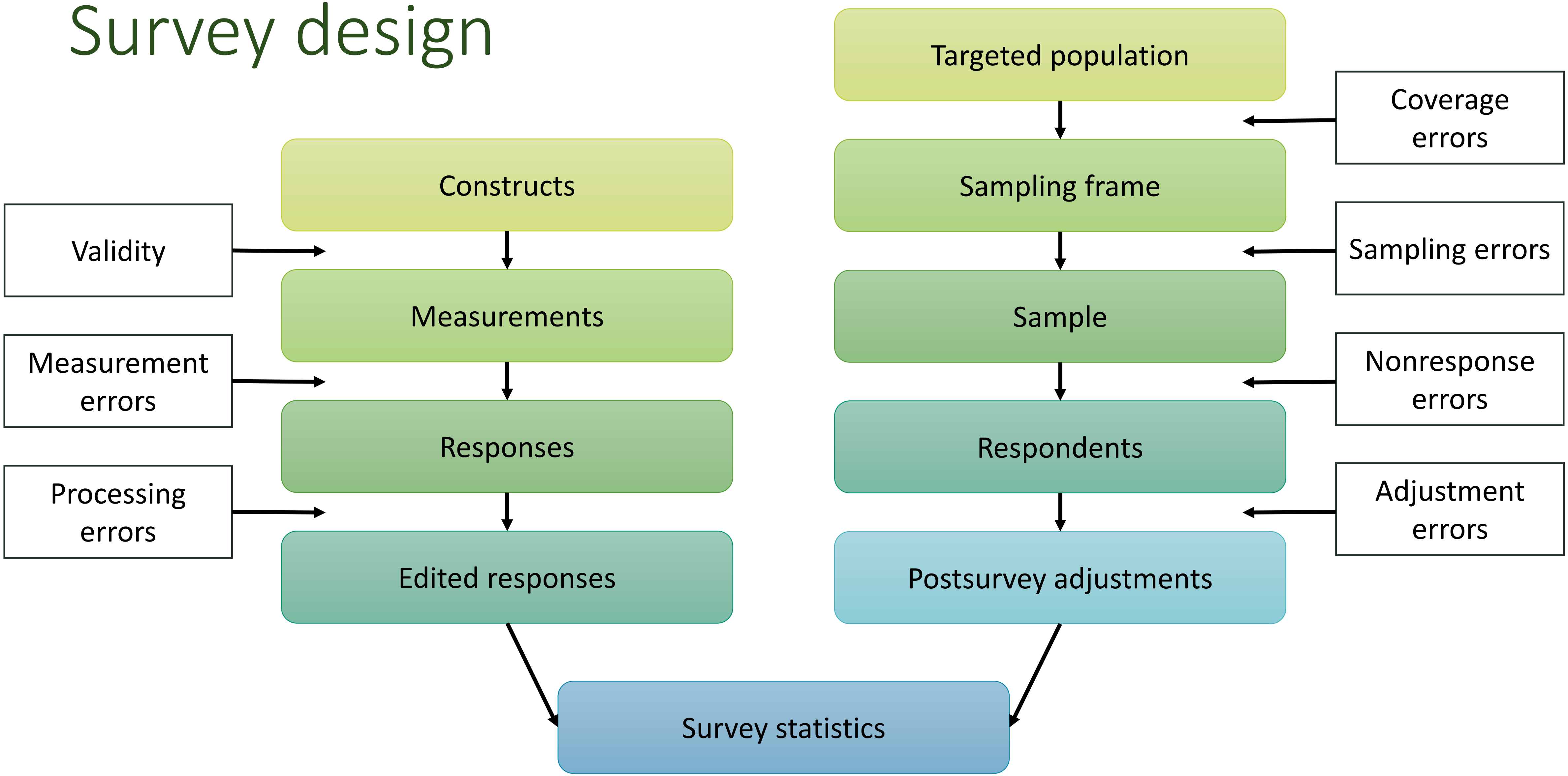
Sample and sampling methods

- Sampling methods
  - Probability sampling
    - Simple random sampling
    - Systematic sampling
    - Stratified sampling
    - Cluster sampling
  - Non-probability sampling
    - Convenience sampling
    - Quota sampling
    - Judgement (or purposive) sampling
    - Snowball sampling

Survey design

- Constructs of interest
  - What?
  - **How to measure?**
- Questionnaires design
  - Wording
  - Use of single question
  - Cognitive processes in answering questions
  - **PRETEST** survey questions

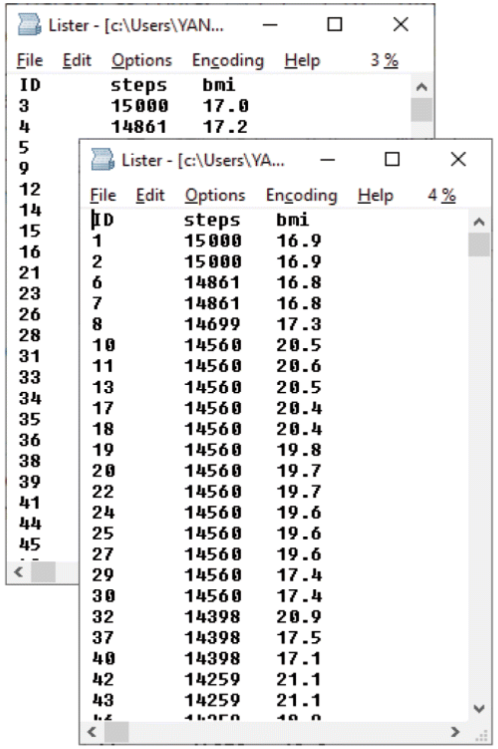
Survey design



Data plotting and cleaning

- Common plots
- Plotting as tool for data cleaning

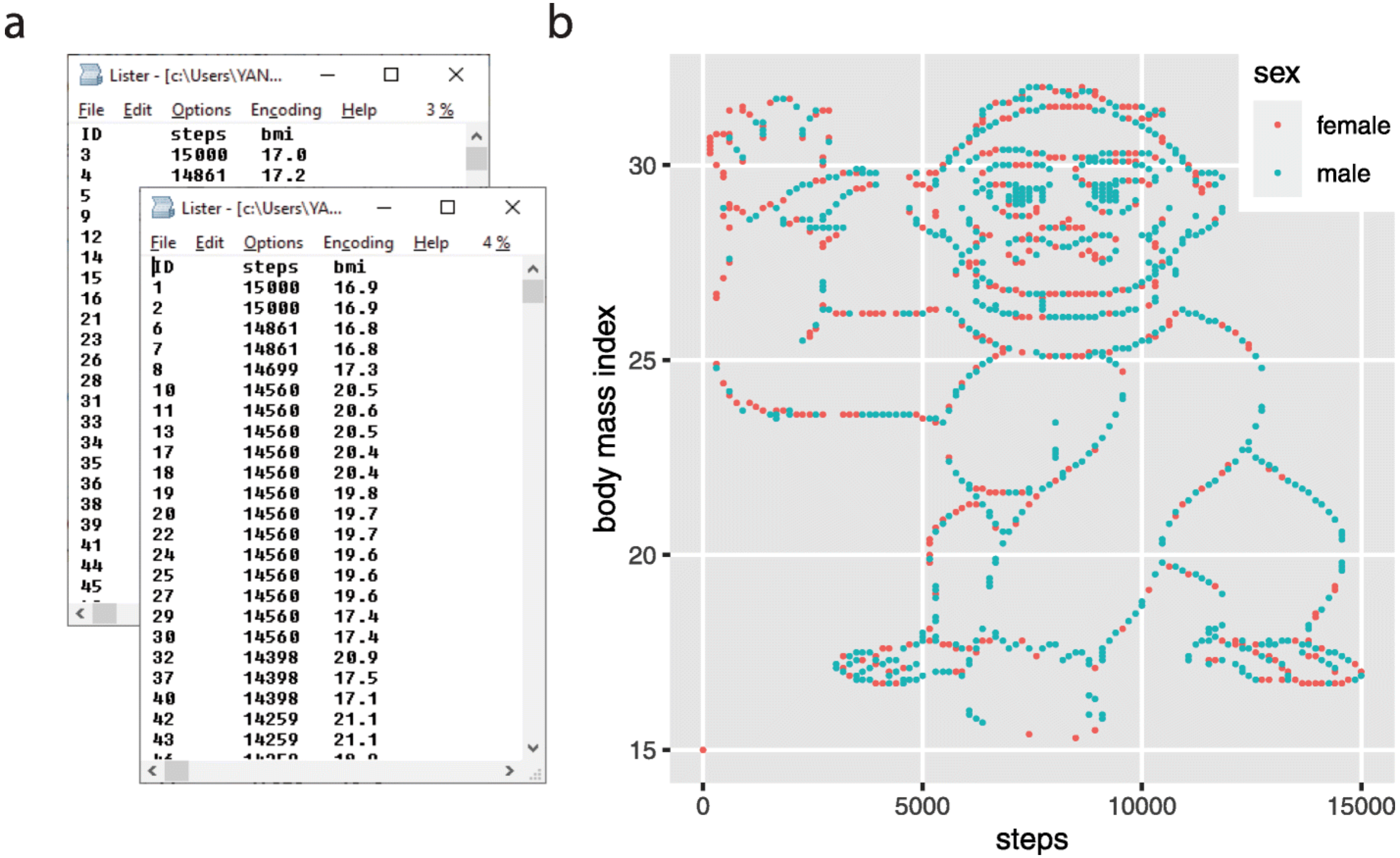
a


Lister - [c:\Users\YAN...
File Edit Options Encoding Help 3 %
Lister - [c:\Users\YA...
File Edit Options Encoding Help 4 %

```
ID        steps    bmi
3         15000    17.0
4         14861    17.2
5
9
12
14
15
16
21
23
26
28
31
33
34
35
36
38
39
41
44
45
```

```
ID        steps    bmi
1         15000    16.9
2         15000    16.9
6         14861    16.8
7         14861    16.8
8         14699    17.3
10        14560    20.5
11        14560    20.6
13        14560    20.5
17        14560    20.4
18        14560    20.4
19        14560    19.8
20        14560    19.7
22        14560    19.7
24        14560    19.6
25        14560    19.6
27        14560    19.6
29        14560    17.4
30        14560    17.4
32        14398    20.9
37        14398    17.5
40        14398    17.1
42        14259    21.1
43        14259    21.1
```

Yanai, I., Lercher, M. A hypothesis is a liability. *Genome Biol* **21**, 231 (2020). https://doi.org/10.1186/s13059-020-02133-w

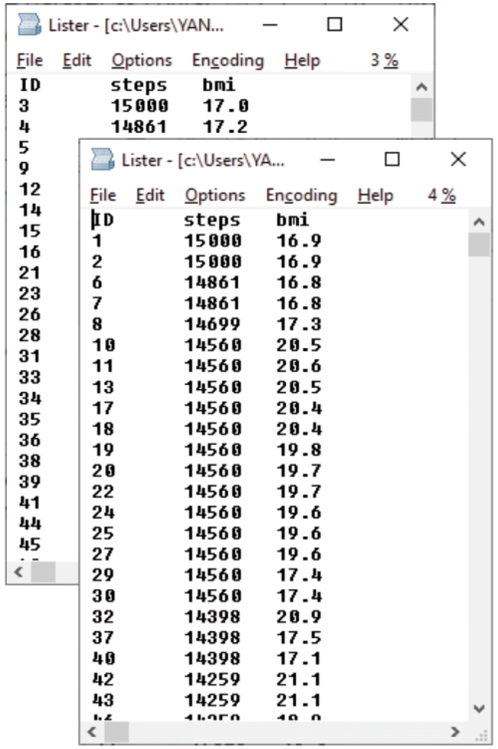


```
ID        steps   bmi
3         15000   17.0
4         14861   17.2
5
9
12
14
15
16
21
23
26
28
31
33
34
35
36
38
39
41
44
45
```

```
ID        steps   bmi
1         15000   16.9
2         15000   16.9
6         14861   16.8
7         14861   16.8
8         14699   17.3
10        14560   20.5
11        14560   20.6
13        14560   20.5
17        14560   20.4
18        14560   20.4
19        14560   19.8
20        14560   19.7
22        14560   19.7
24        14560   19.6
25        14560   19.6
27        14560   19.6
29        14560   17.4
30        14560   17.4
32        14398   20.9
37        14398   17.5
40        14398   17.1
42        14259   21.1
43        14259   21.1
```

Yanai, I., Lercher, M. A hypothesis is a liability. *Genome Biol* **21**, 231 (2020). https://doi.org/10.1186/s13059-020-02133-w

a

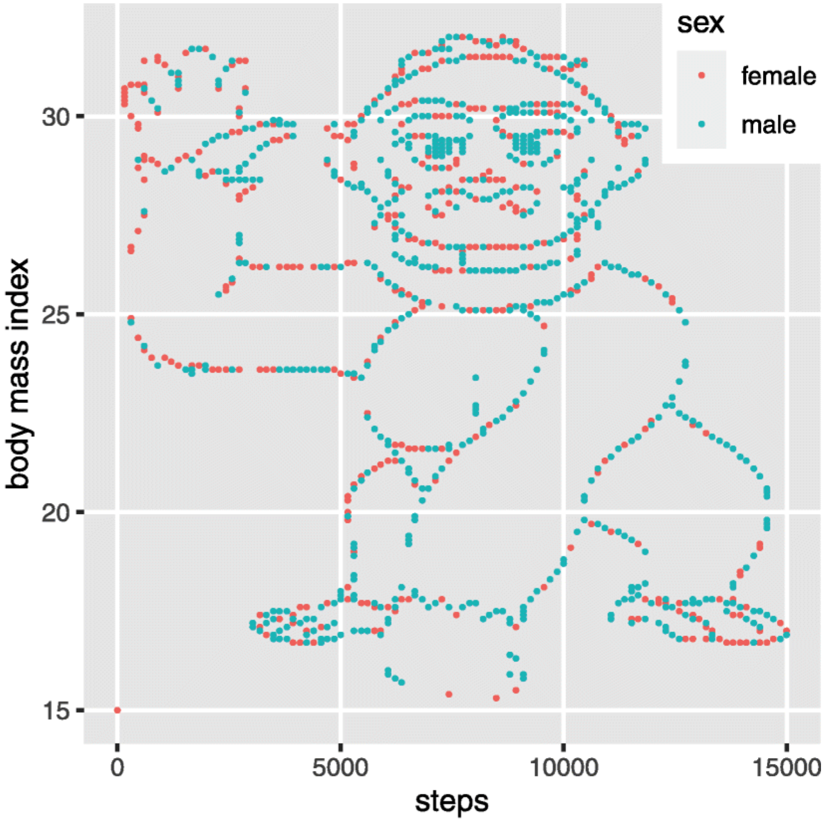

Lister - [c:\Users\YAN...
File Edit Options Encoding Help 3 %

```
ID        steps   bmi
3         15000   17.0
4         14861   17.2
5
9
12
14
15
16
21
23
26
28
31
33
34
35
36
38
39
41
44
45
```

Lister - [c:\Users\YA...
File Edit Options Encoding Help 4 %

```
ID        steps   bmi
1         15000   16.9
2         15000   16.9
6         14861   16.8
7         14861   16.8
8         14699   17.3
10        14560   20.5
11        14560   20.6
13        14560   20.5
17        14560   20.4
18        14560   20.4
19        14560   19.8
20        14560   19.7
22        14560   19.7
24        14560   19.6
25        14560   19.6
27        14560   19.6
29        14560   17.4
30        14560   17.4
32        14398   20.9
37        14398   17.5
40        14398   17.1
42        14259   21.1
43        14259   21.1
```

b

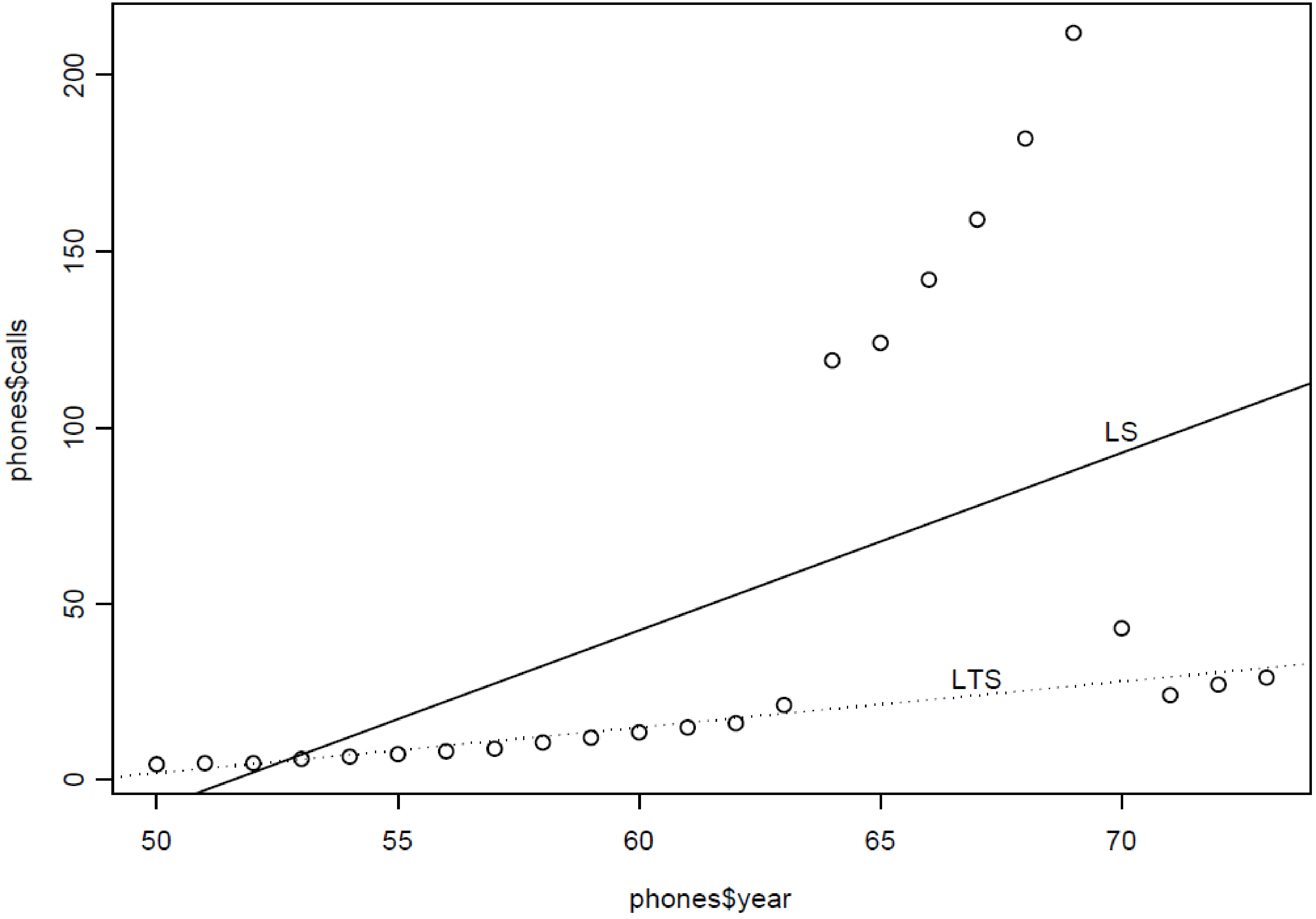


c

	Gorilla not discovered	Gorilla discovered
Hypothesis-focused	14	5
Hypothesis-free	5	9

Yanai, I., Lercher, M. A hypothesis is a liability. *Genome Biol* **21**, 231 (2020). https://doi.org/10.1186/s13059-020-02133-w

- Telephone data
- Calls (in millions) from Belgium in the years 1950-1973.

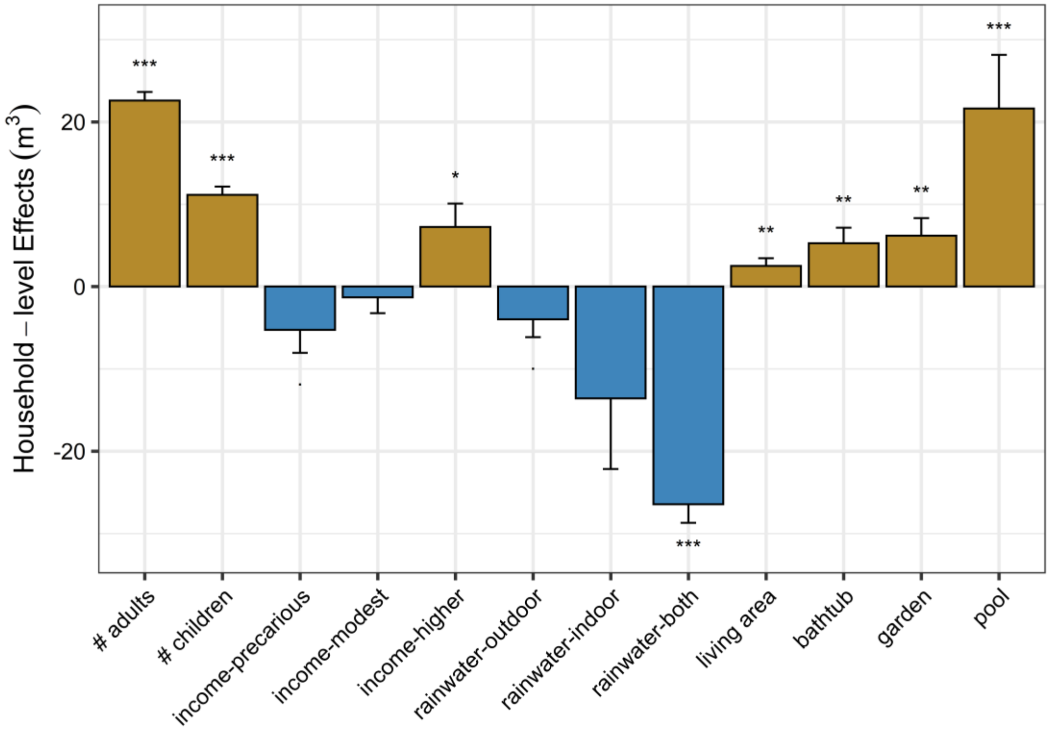


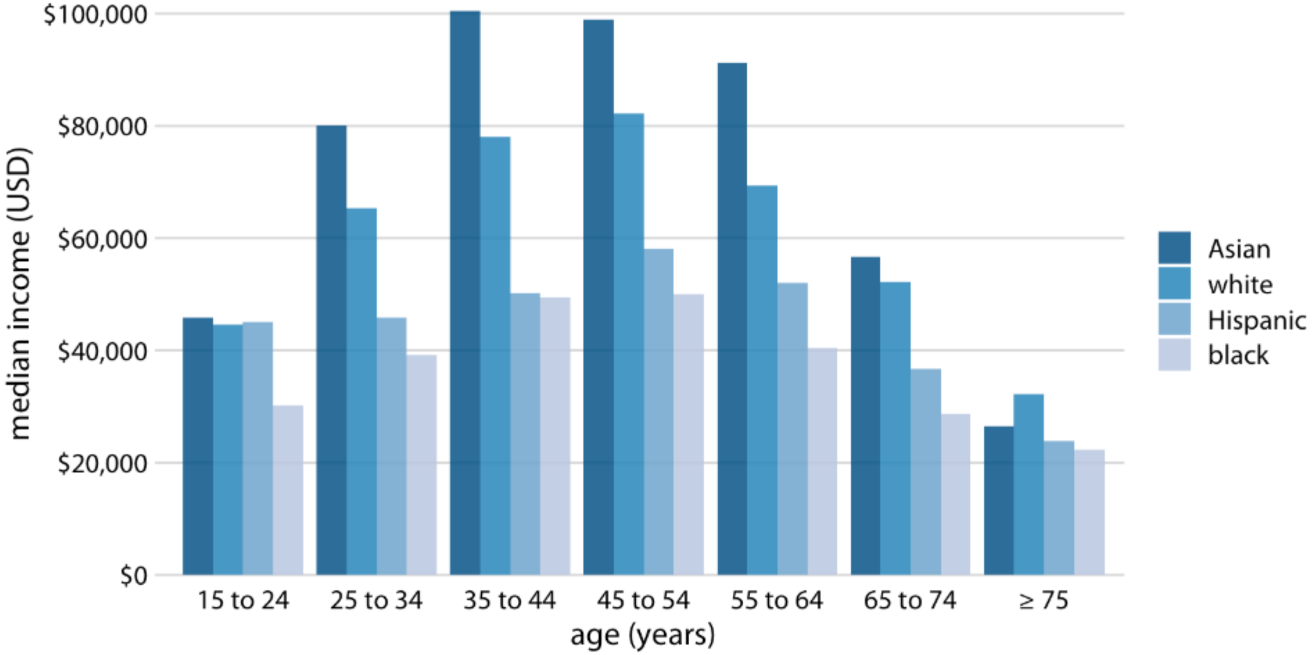
Data plotting/visualization

- Clarity
- Precision
- Efficiency
- Maximize ideas, minimize ink

https://www.ted.com/talks/hans_rosling_the_best_stats_you_ve_ever_seen

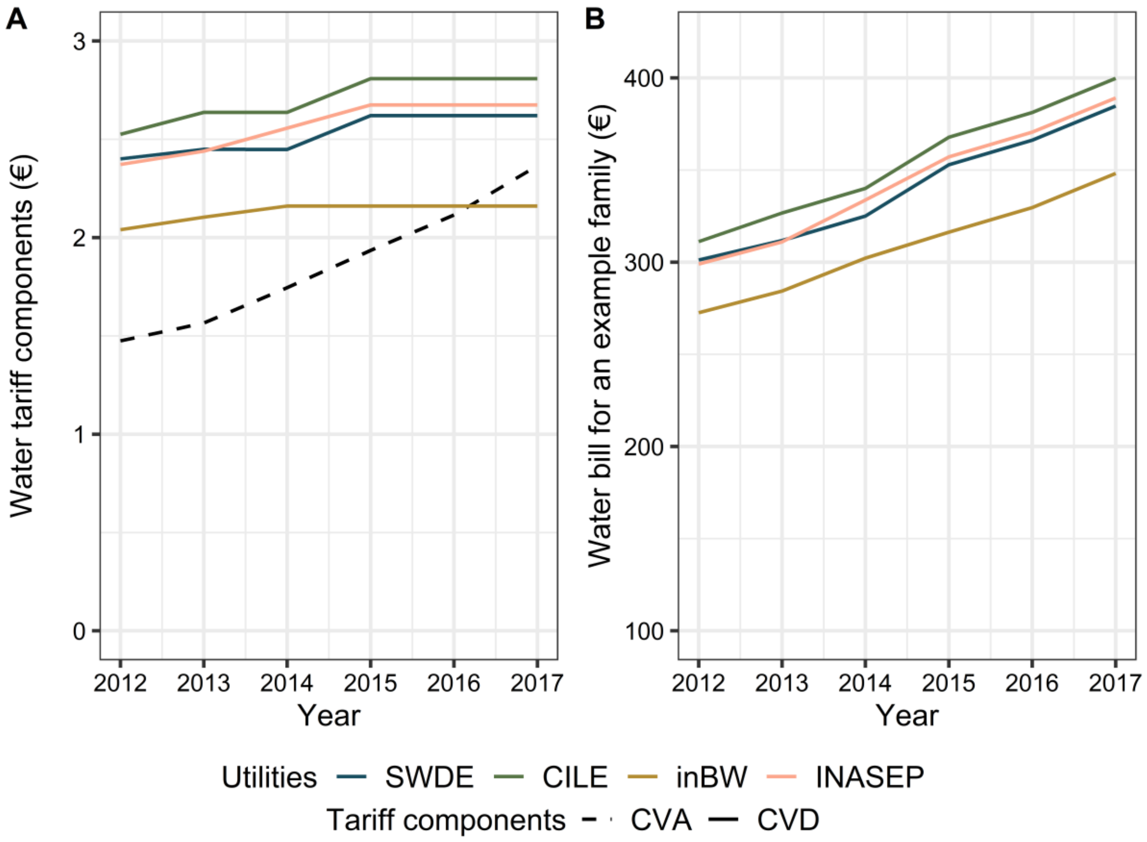
Commonly used plots

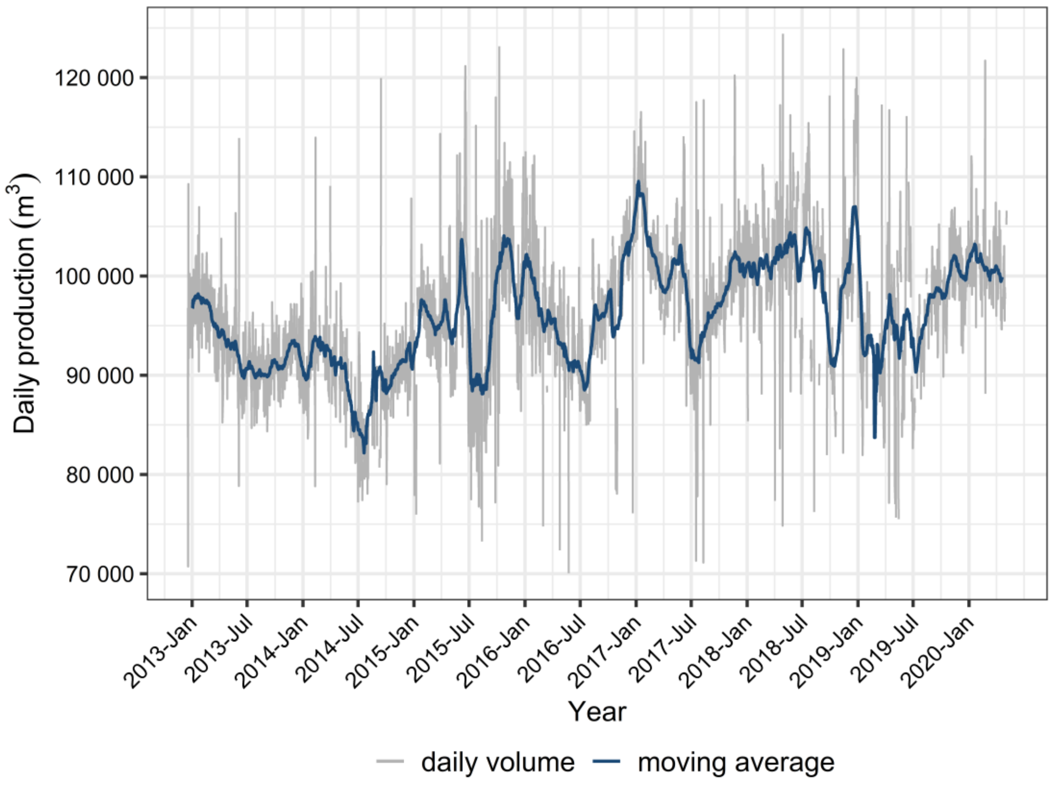




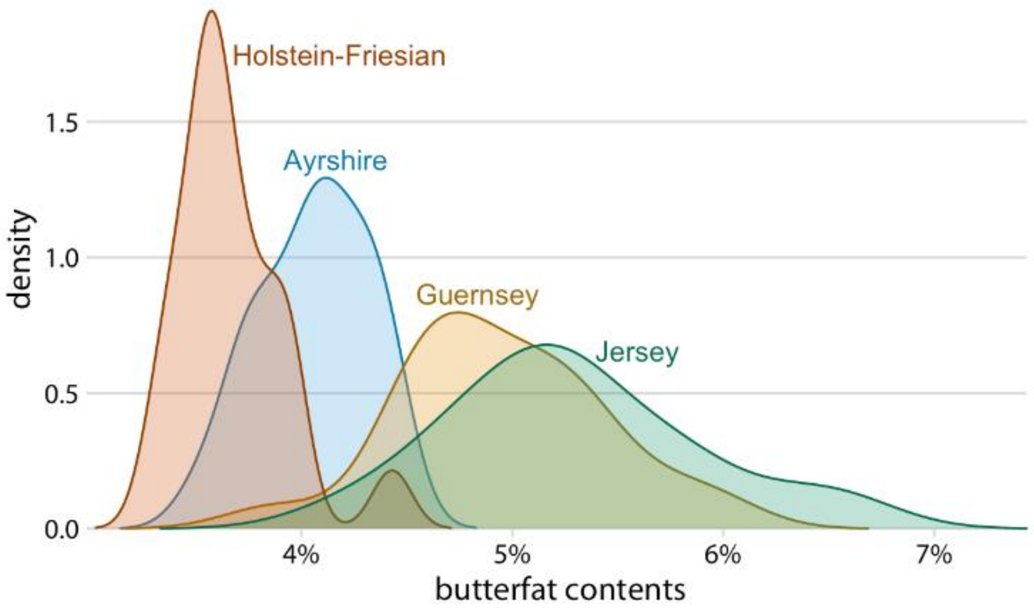
Wilke (2018)

Commonly used plots

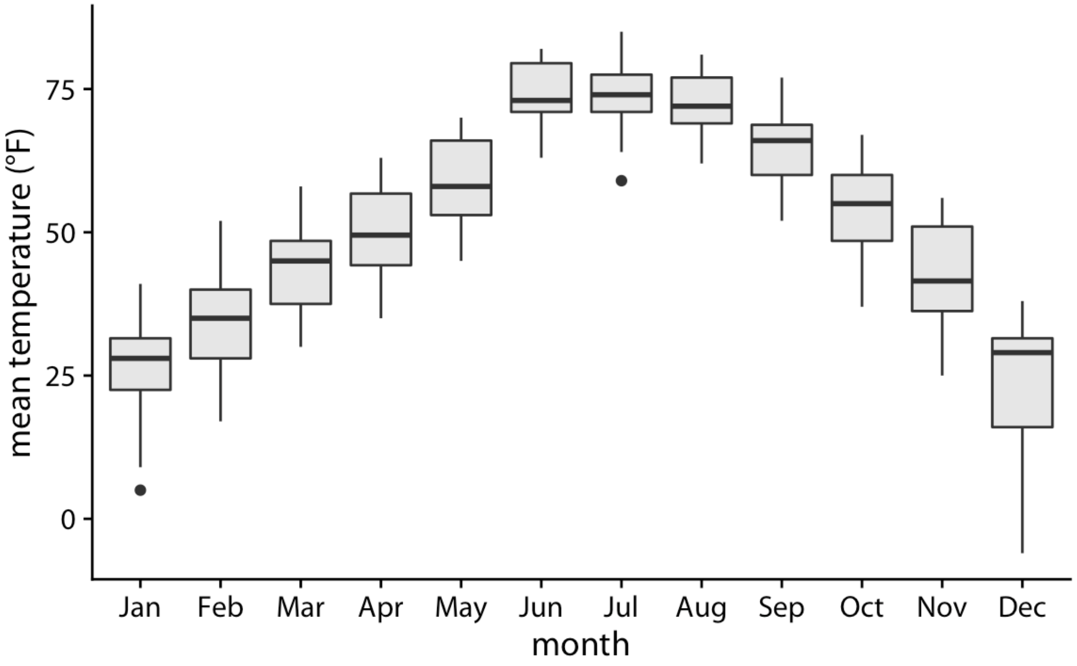




Commonly used plots

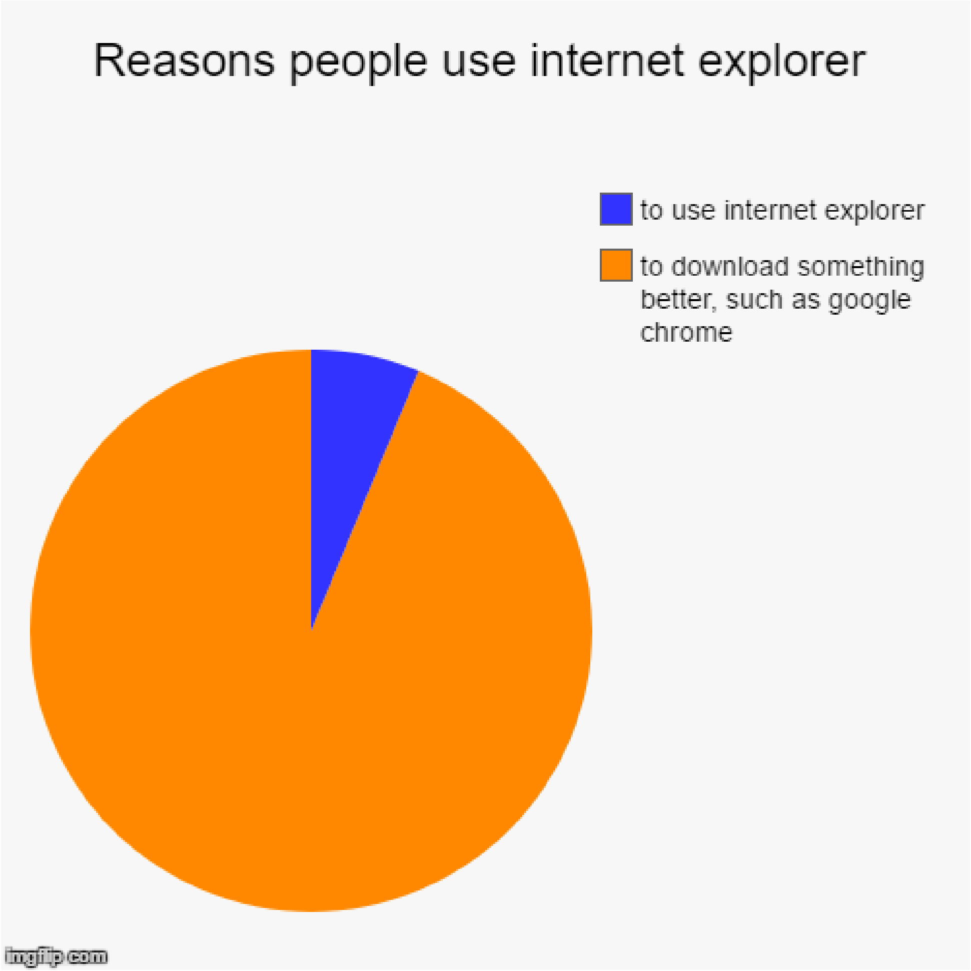


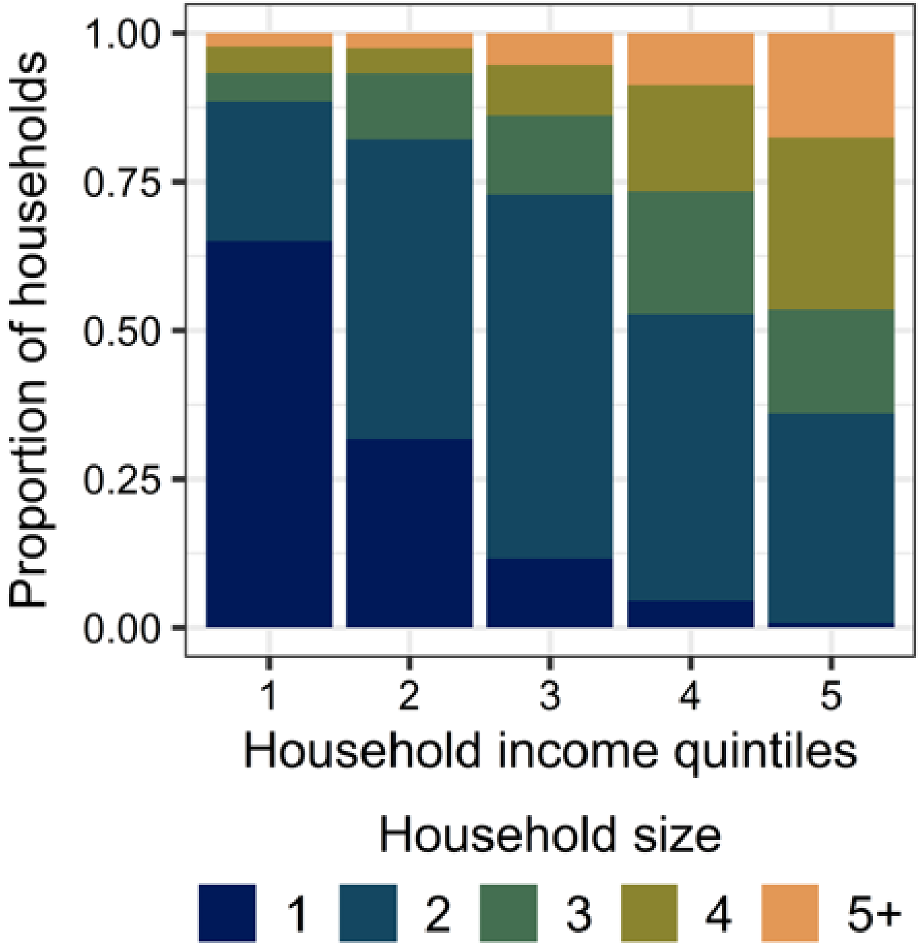
Wilke (2018)



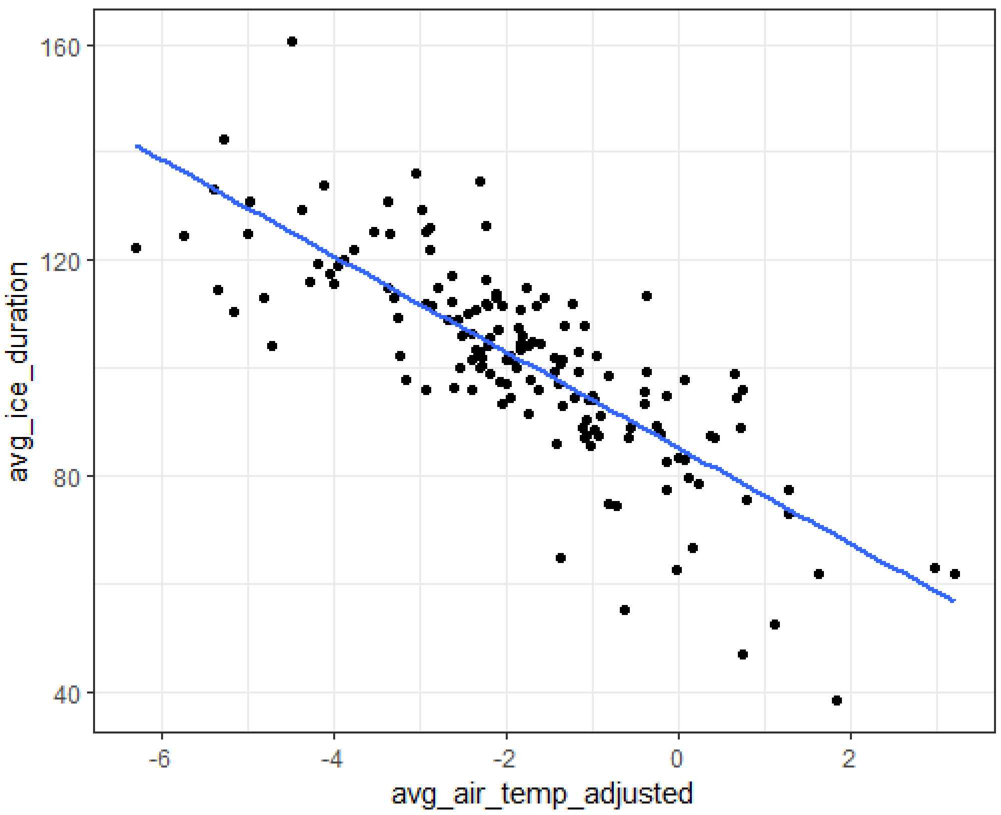
Wilke (2018)

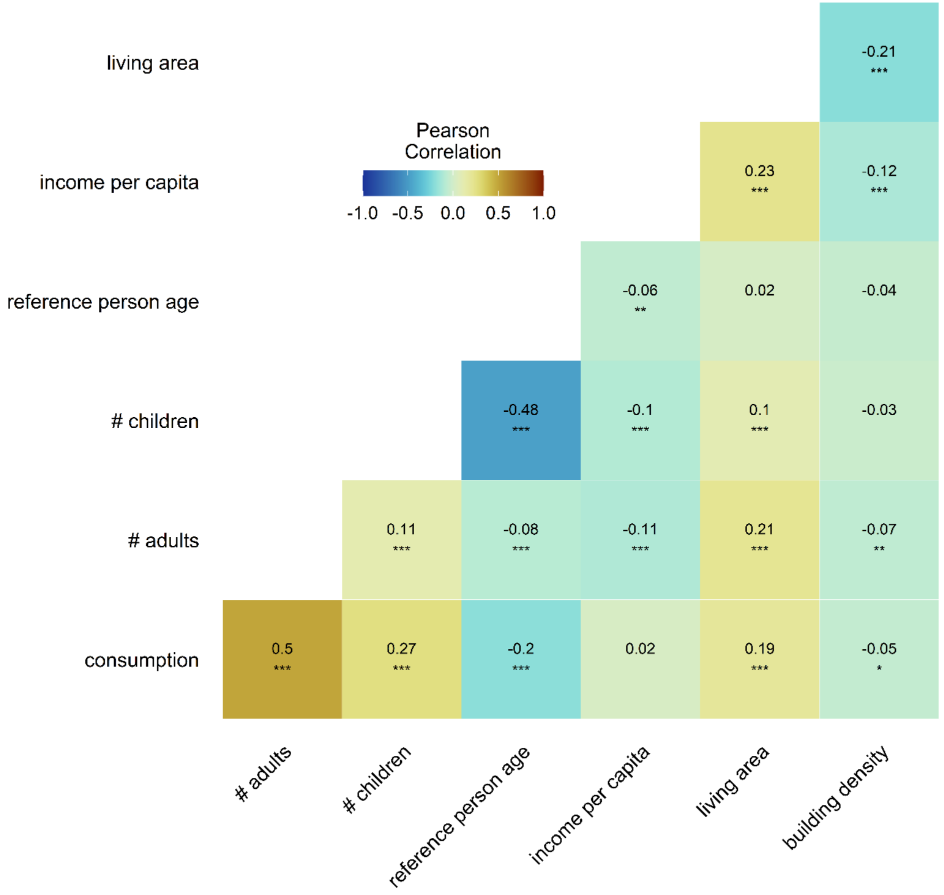
Commonly used plots



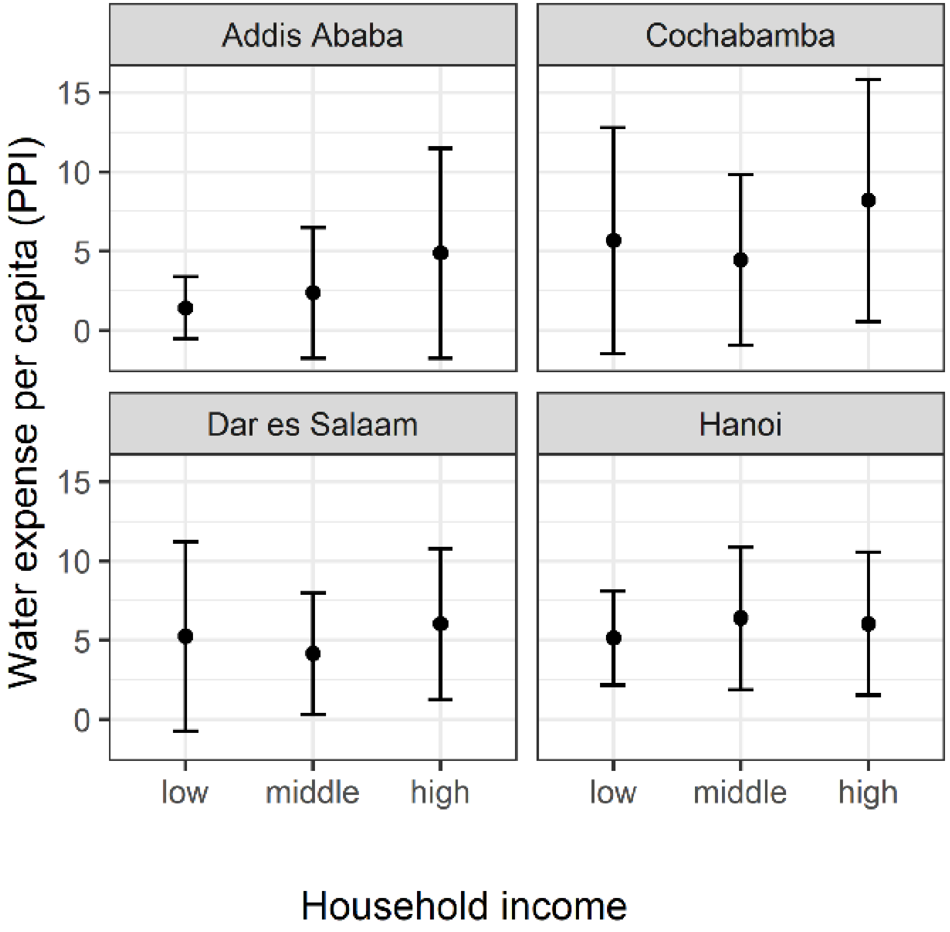


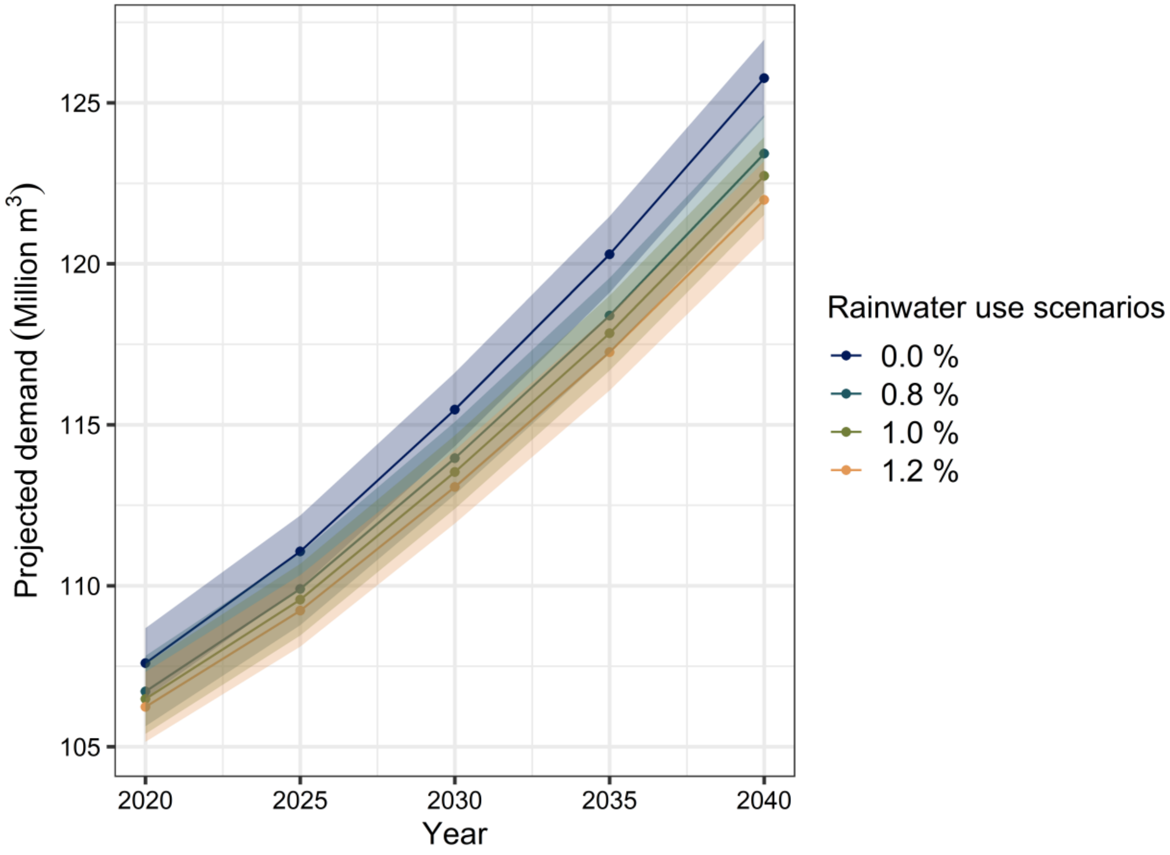
Commonly used plots





Commonly used plots

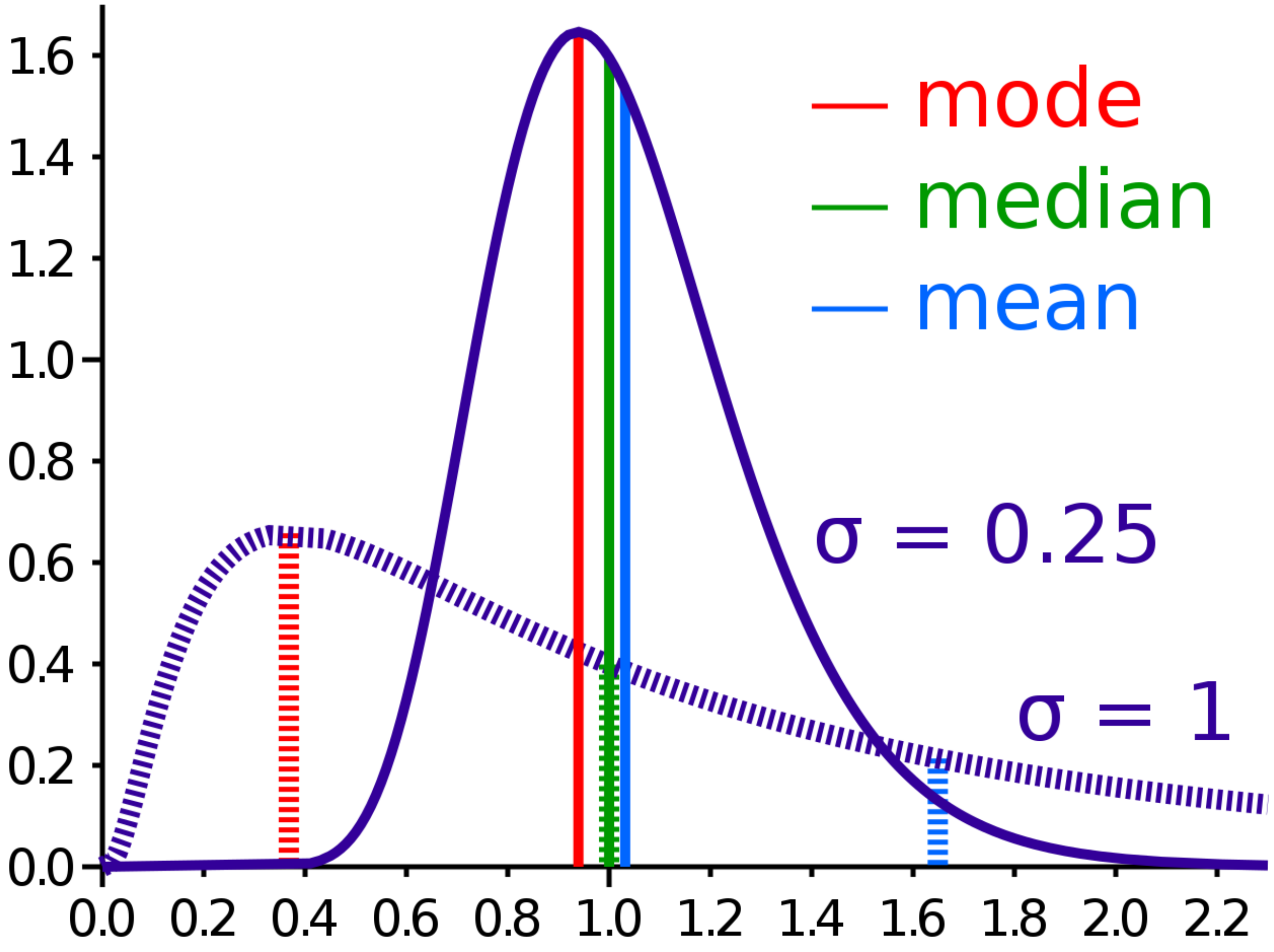




Descriptive statistics

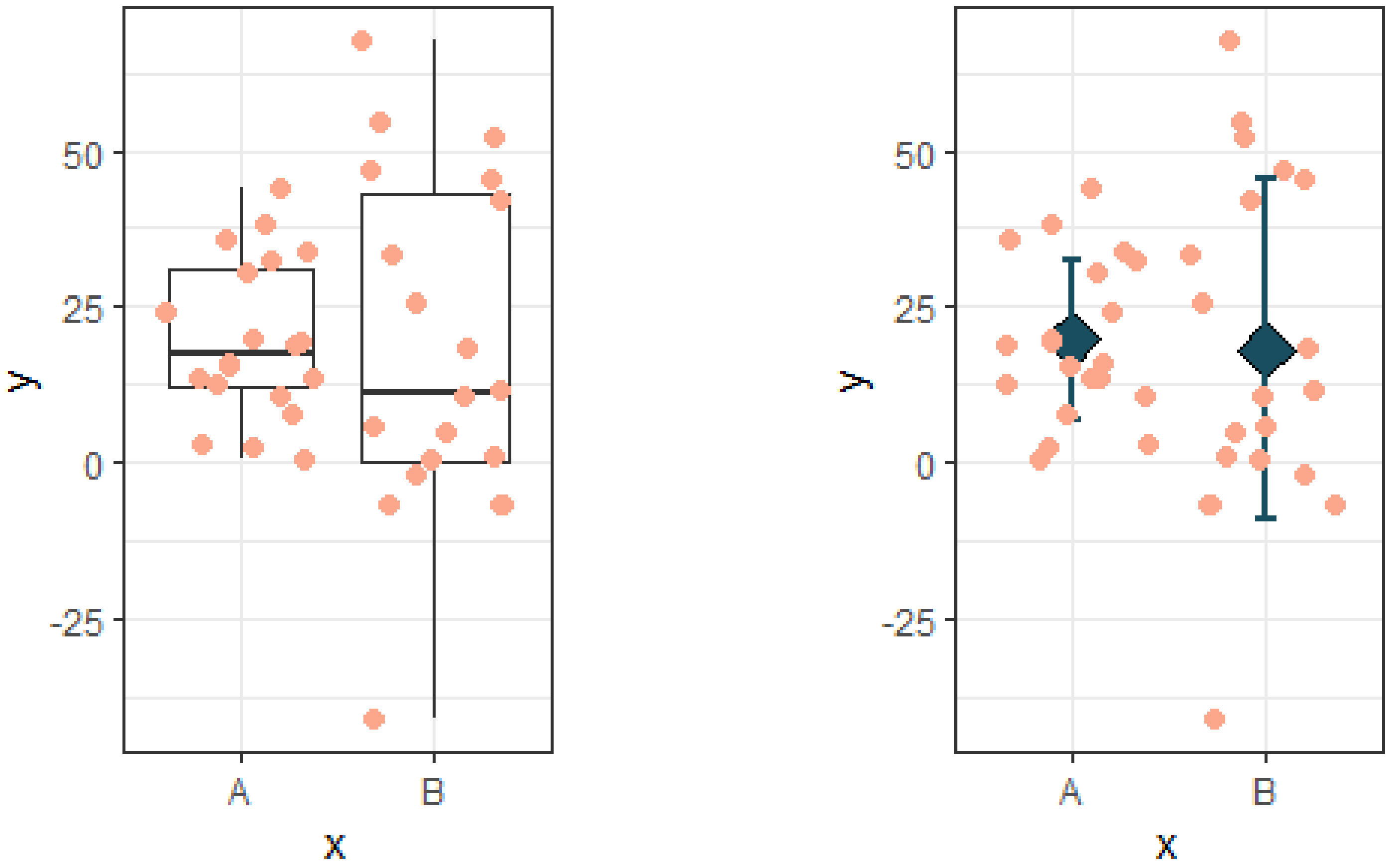
- Univariable
- Bivariable

Mean, median, mode



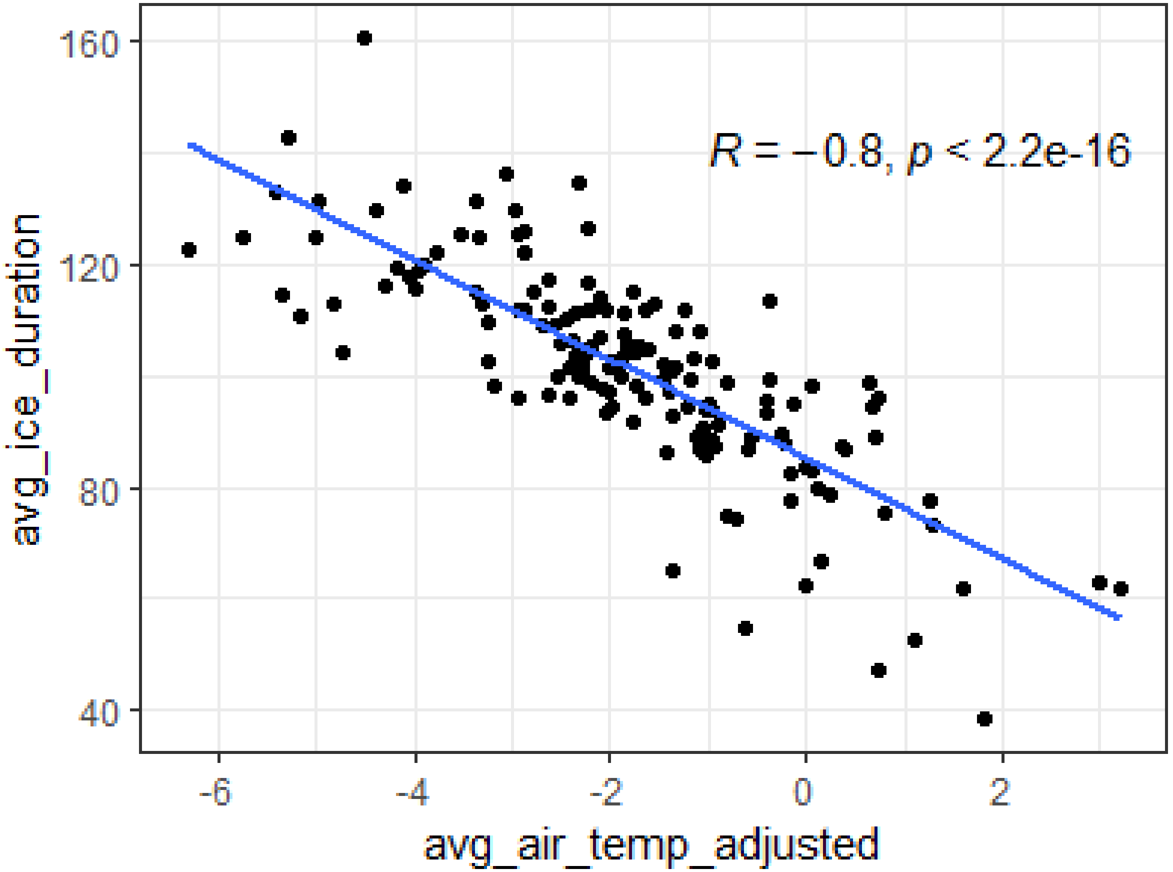
This Photo by Unknown Author is licensed under CC BY-SA

Dispersion

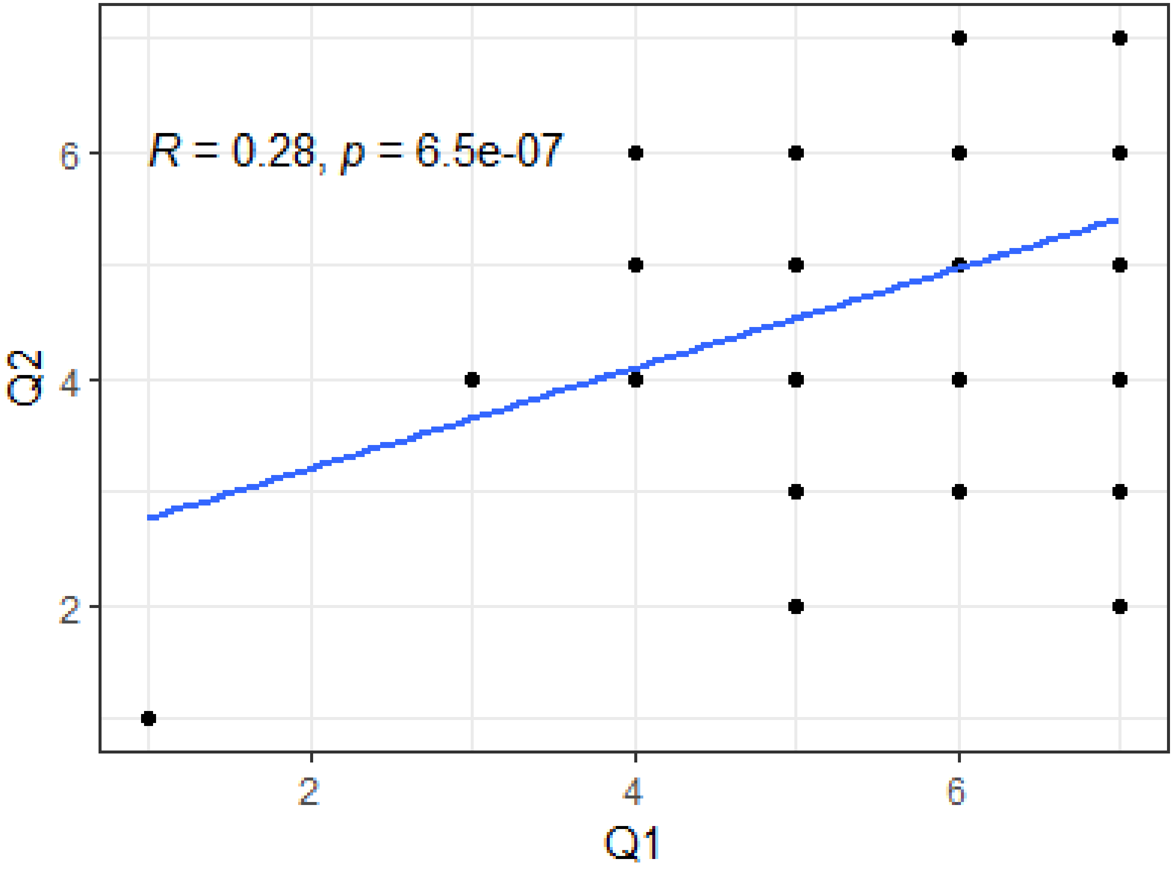


Associations

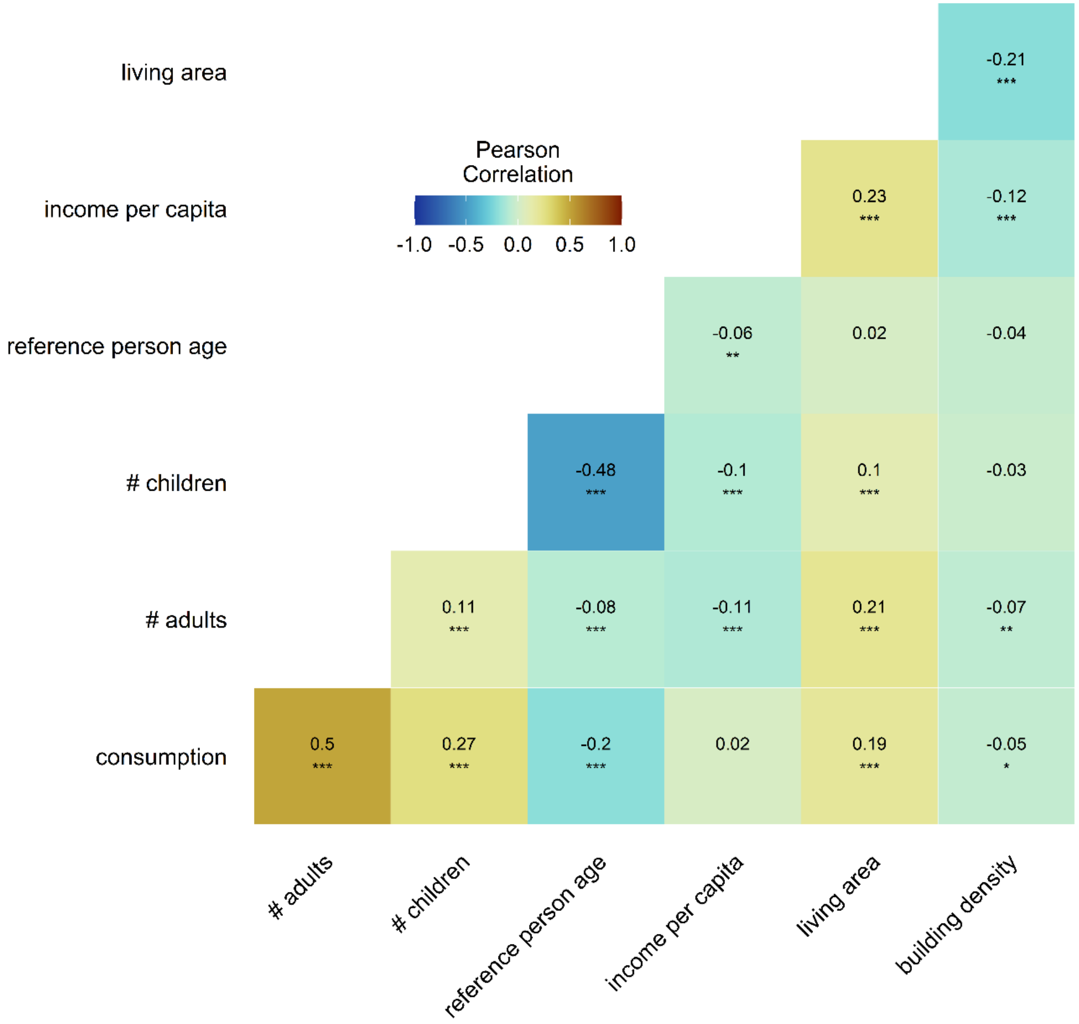
Pearson



Spearman

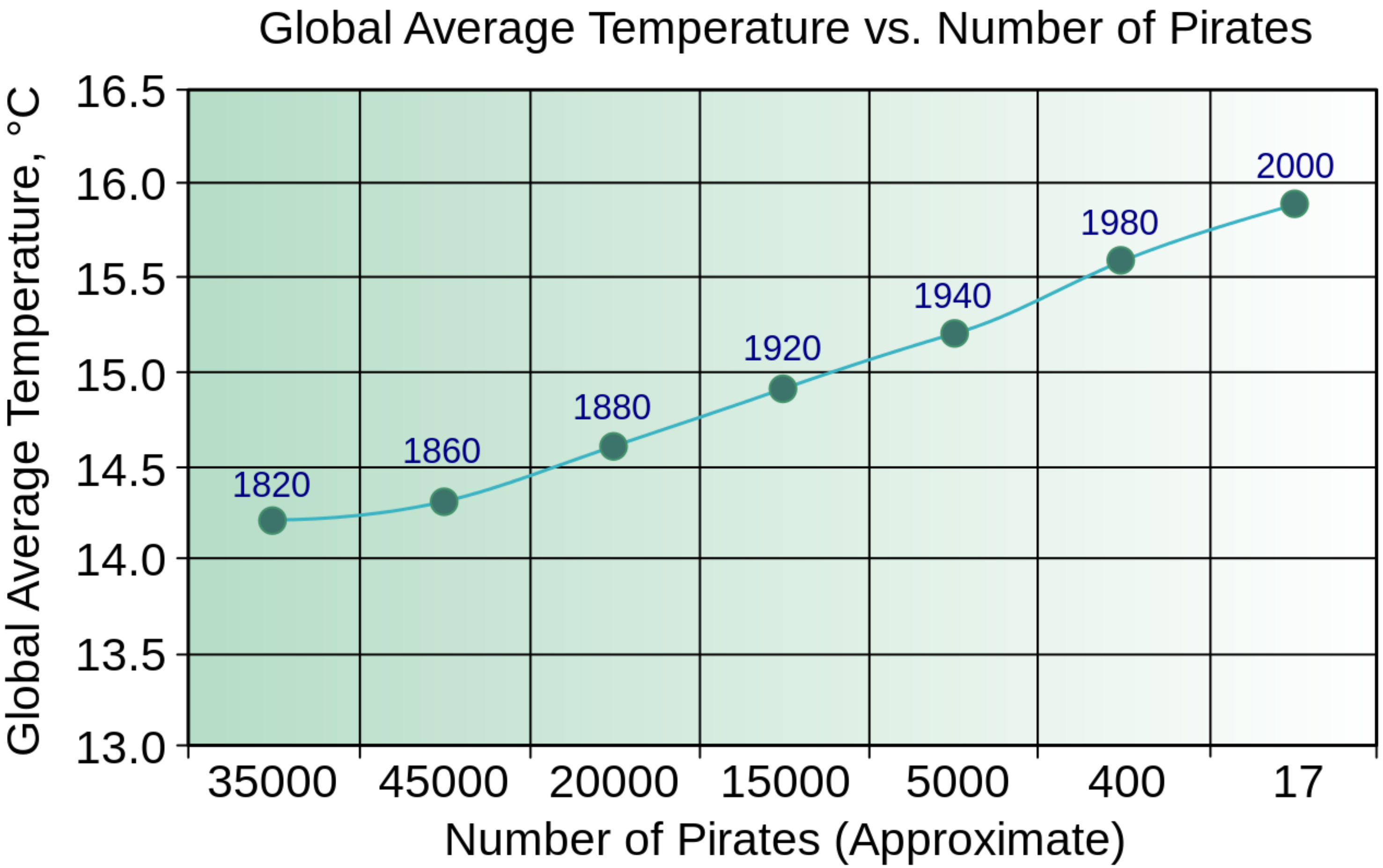


Associations



	# adults	# children	reference person age	income per capita	living area	building density
living area						-0.21 ***
income per capita					0.23 ***	-0.12 ***
reference person age				-0.06 **	0.02	-0.04
# children			-0.48 ***	-0.1 ***	0.1 ***	-0.03
# adults		0.11 ***	-0.08 ***	-0.11 ***	0.21 ***	-0.07 **
consumption	0.5 ***	0.27 ***	-0.2 ***	0.02	0.19 ***	-0.05 *

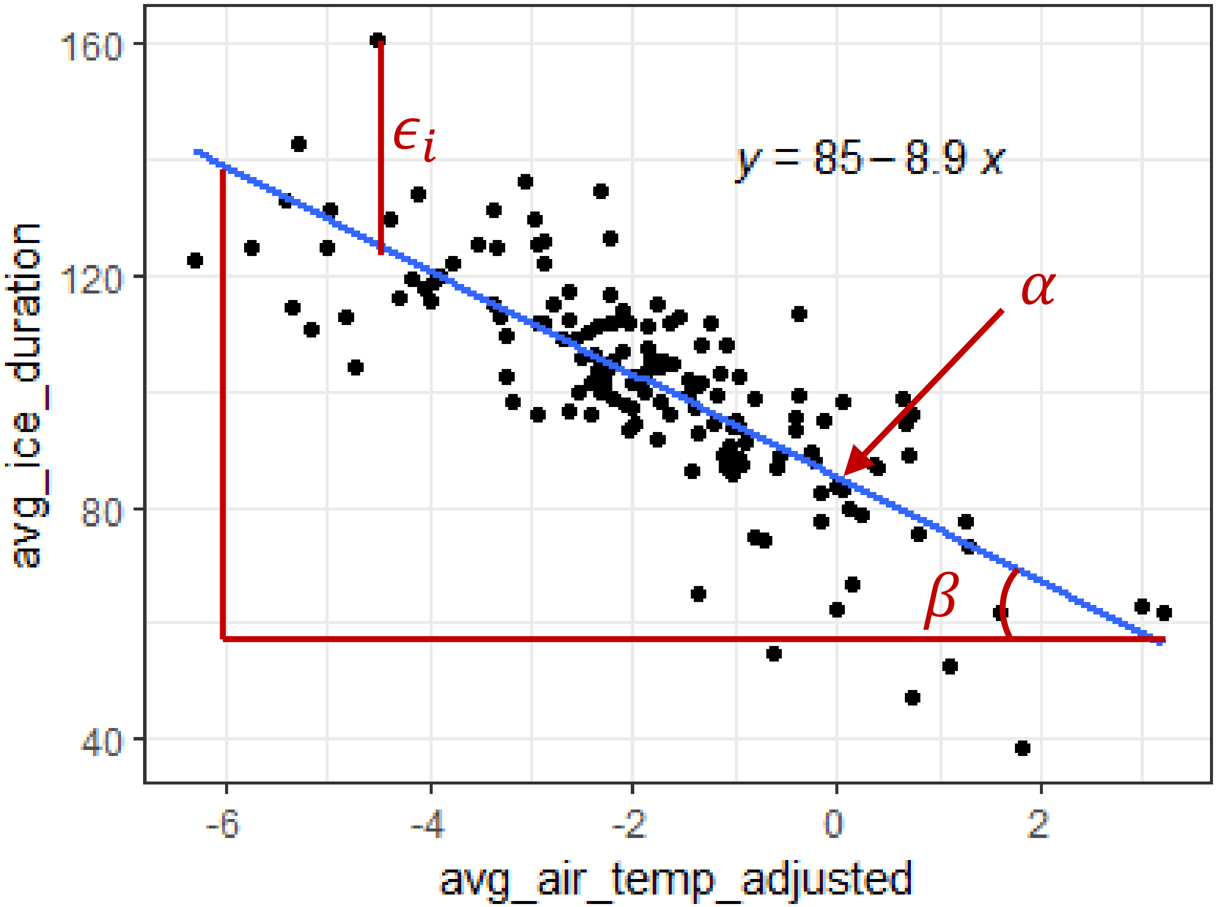
Correlation vs Causality



Inferential statistics

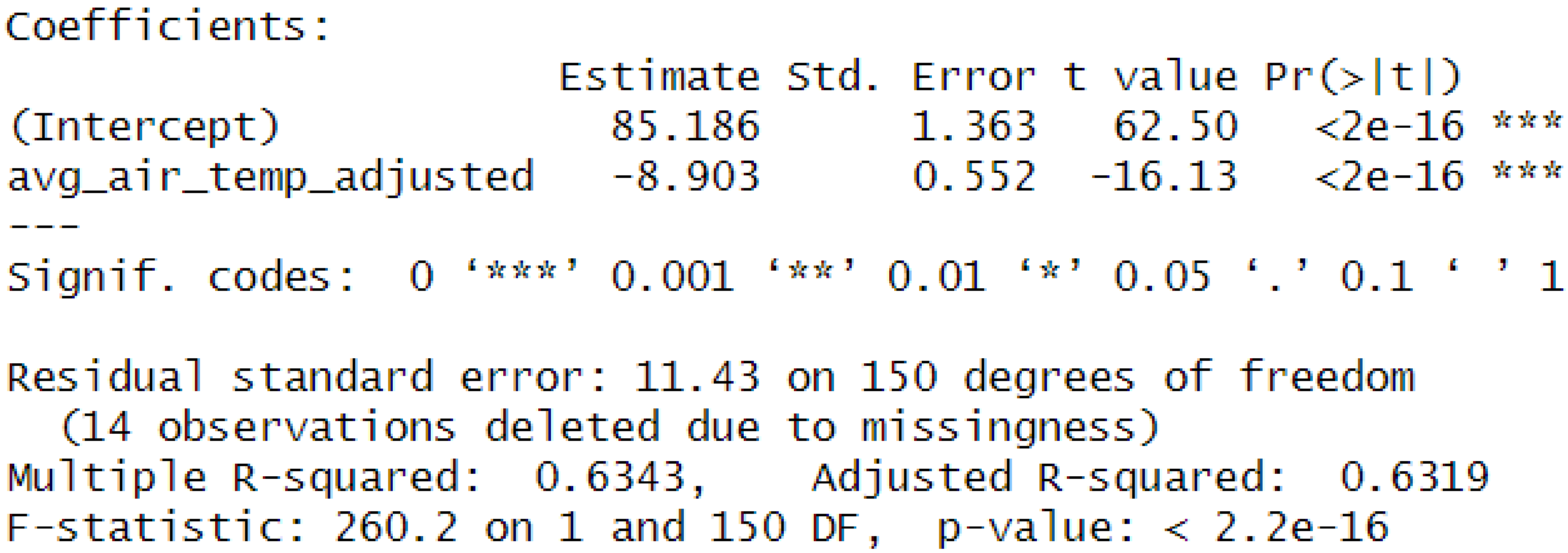
- Simple linear regression
- General linear regression

Simple linear regression



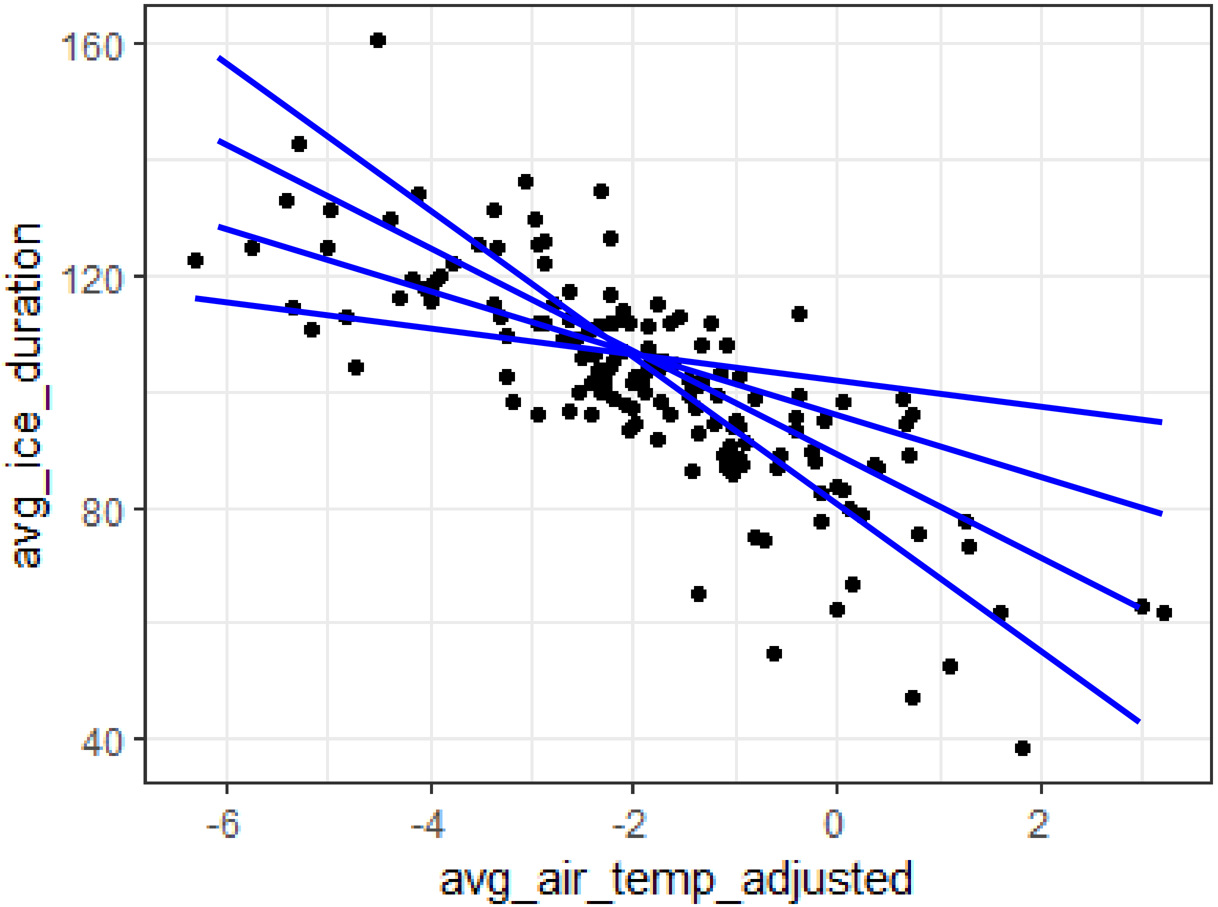
$$Y_i = \alpha + \beta X_i + \epsilon_i$$
$$\epsilon_i \sim N(0, \sigma^2)$$

Simple linear regression



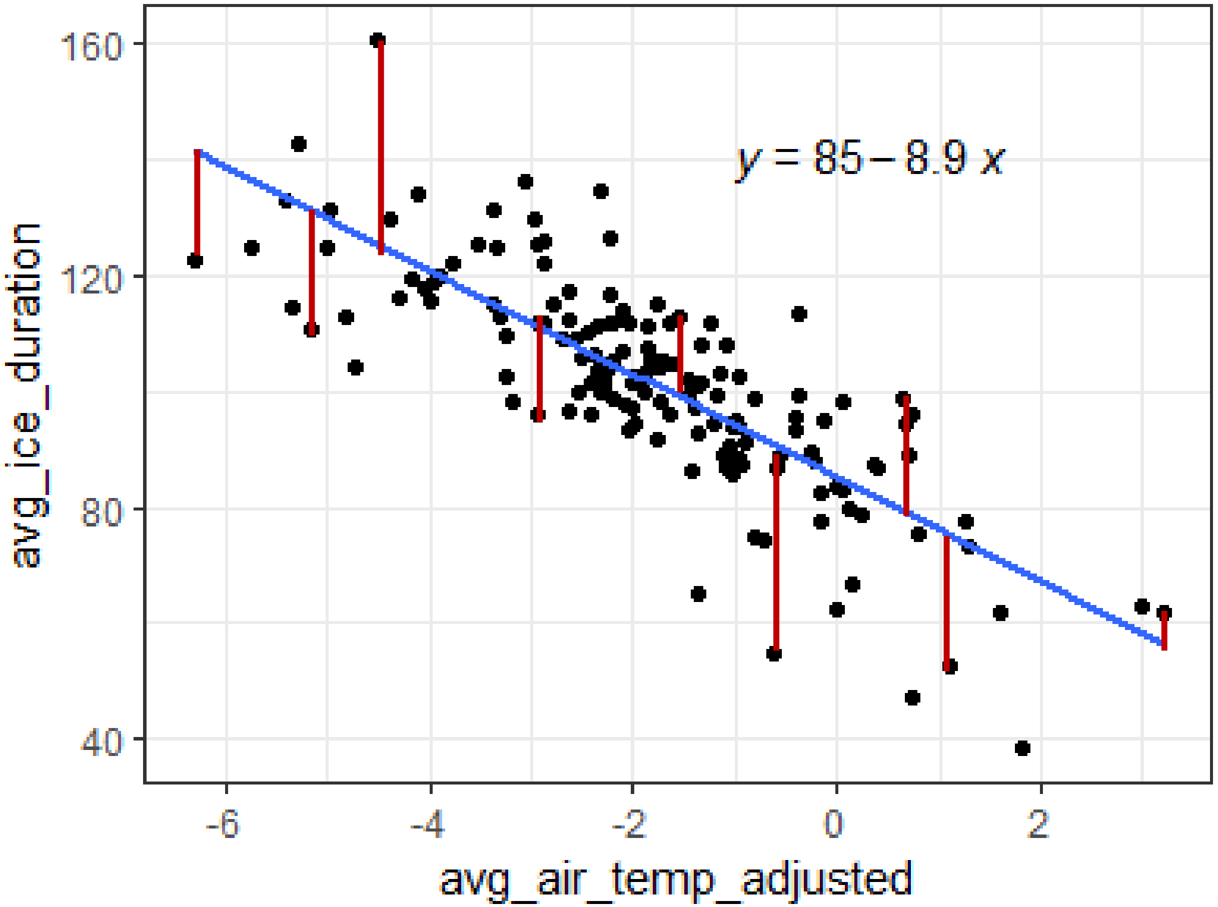
```
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)             85.186      1.363   62.50   <2e-16 ***
avg_air_temp_adjusted   -8.903      0.552  -16.13   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.43 on 150 degrees of freedom
  (14 observations deleted due to missingness)
Multiple R-squared:  0.6343,    Adjusted R-squared:  0.6319 
F-statistic: 260.2 on 1 and 150 DF,  p-value: < 2.2e-16
```

Simple linear regression



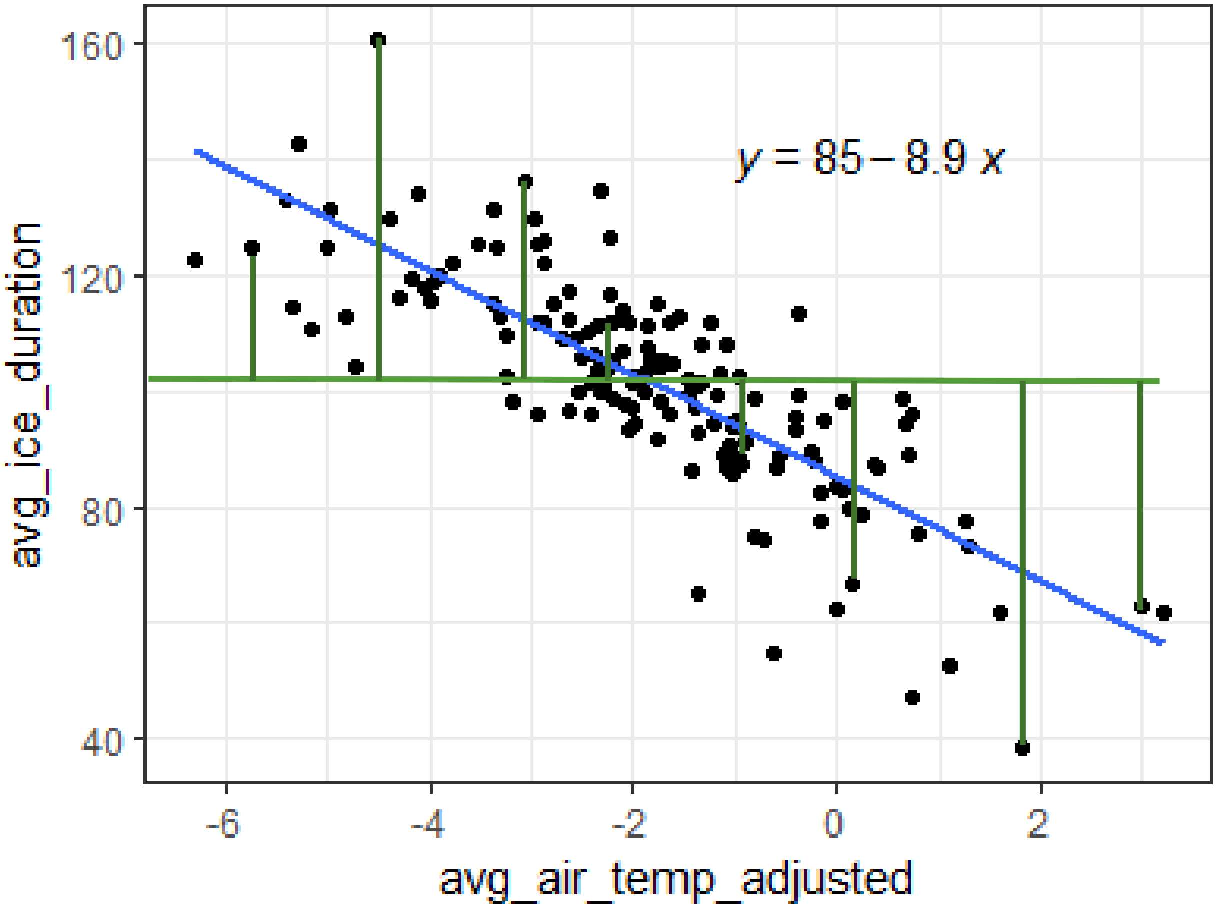
Which one?

Simple linear regression



sum of the squared residuals (SSR)

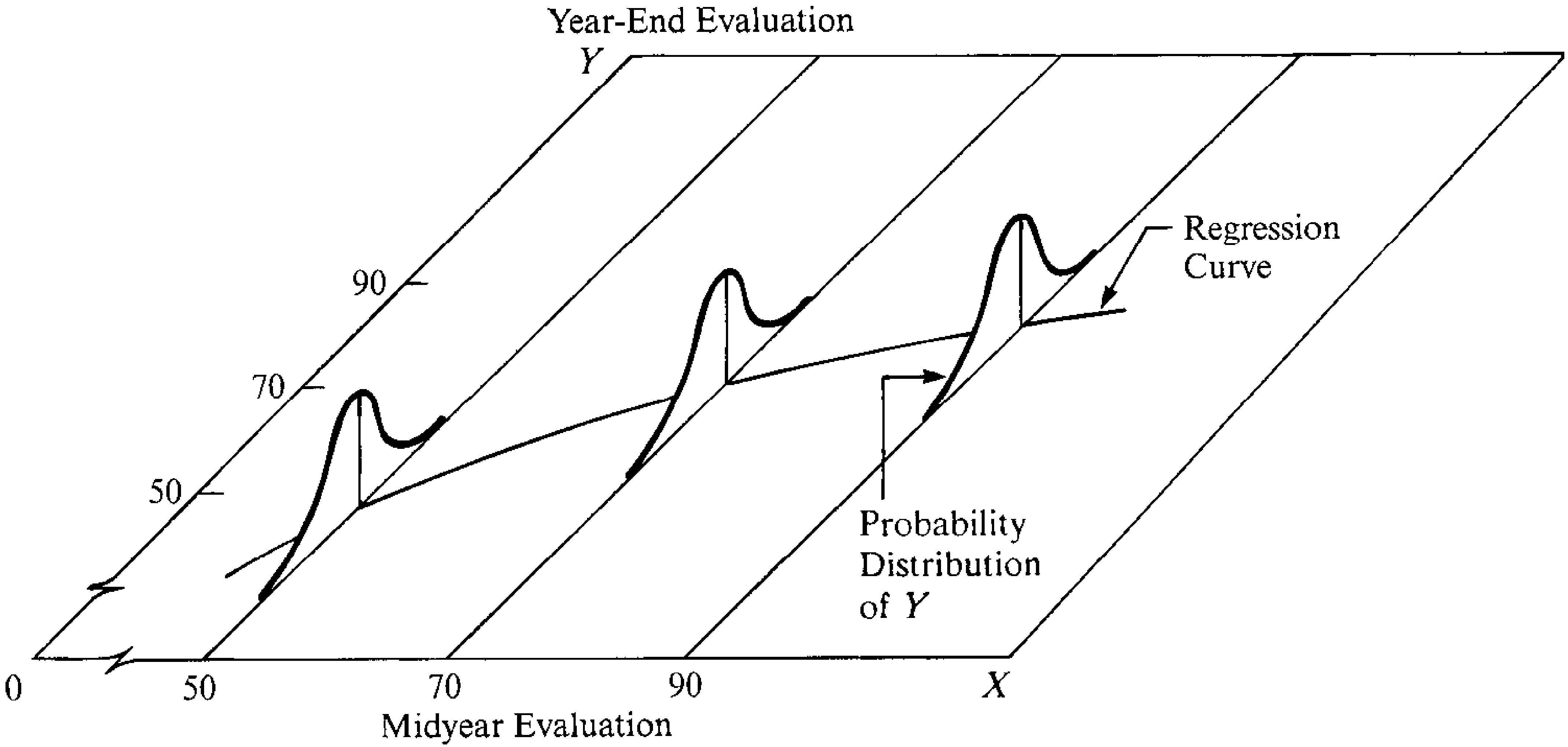
Goodness of fit



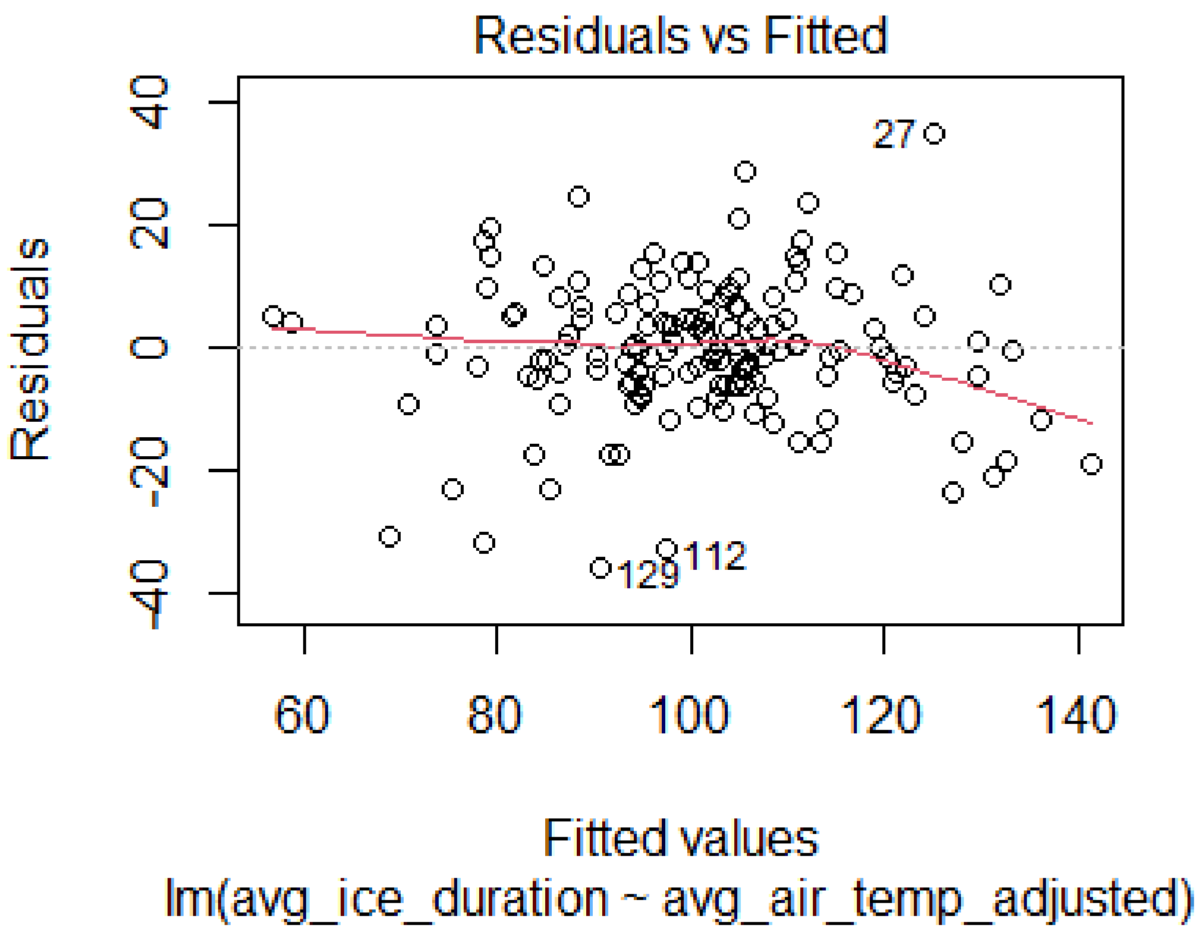
$$R^2 = \frac{TSS - SSR}{TSS} = 1 - \frac{SSR}{TSS}$$

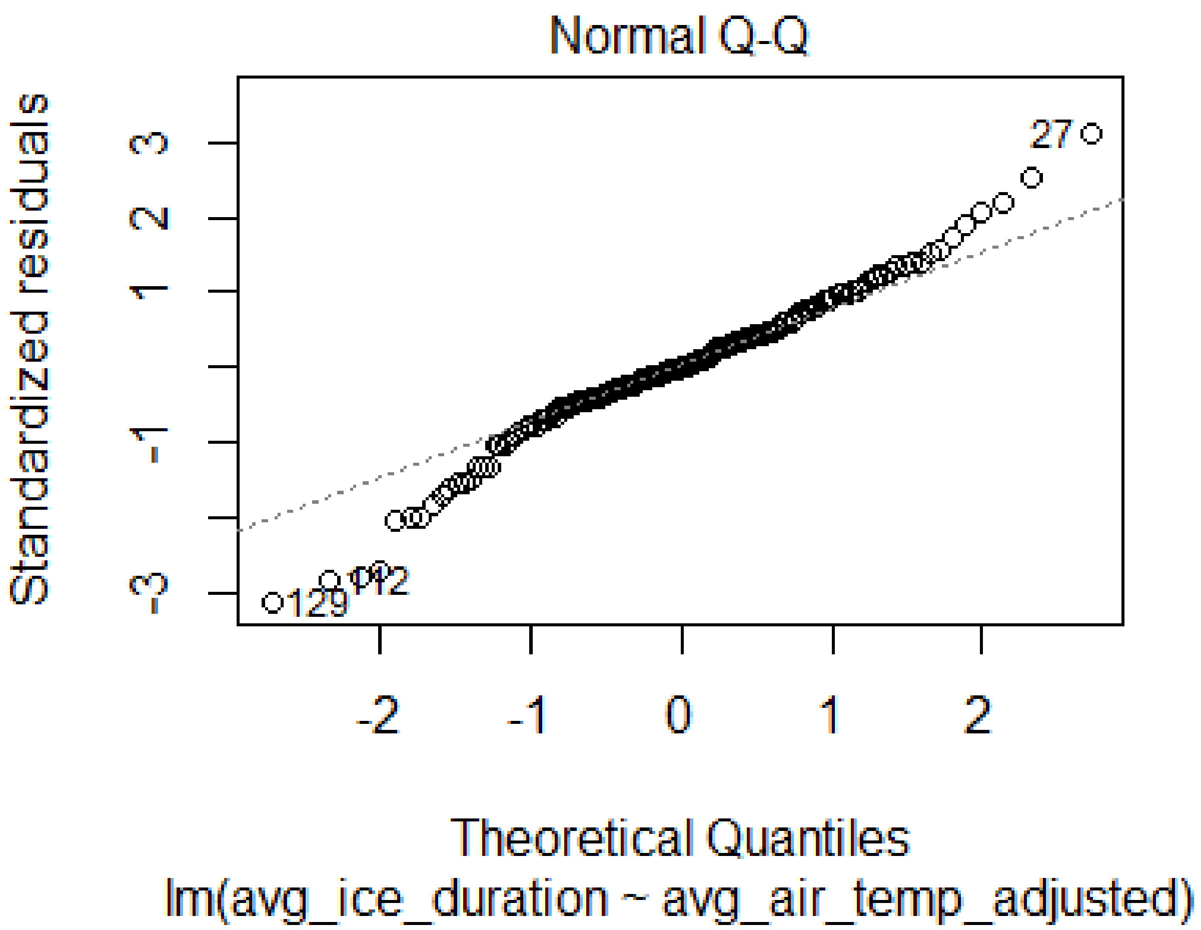
- TSS: Total sum of squares
- SSR: Sum squared residuals

Assumptions



Assumptions





General Linear Regression

$$Y_i = \alpha + \beta_1 X_{i1} + \beta_1 X_{i1}^2 + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \epsilon_i$$

$$\epsilon_i \sim N(0, \sigma^2)$$

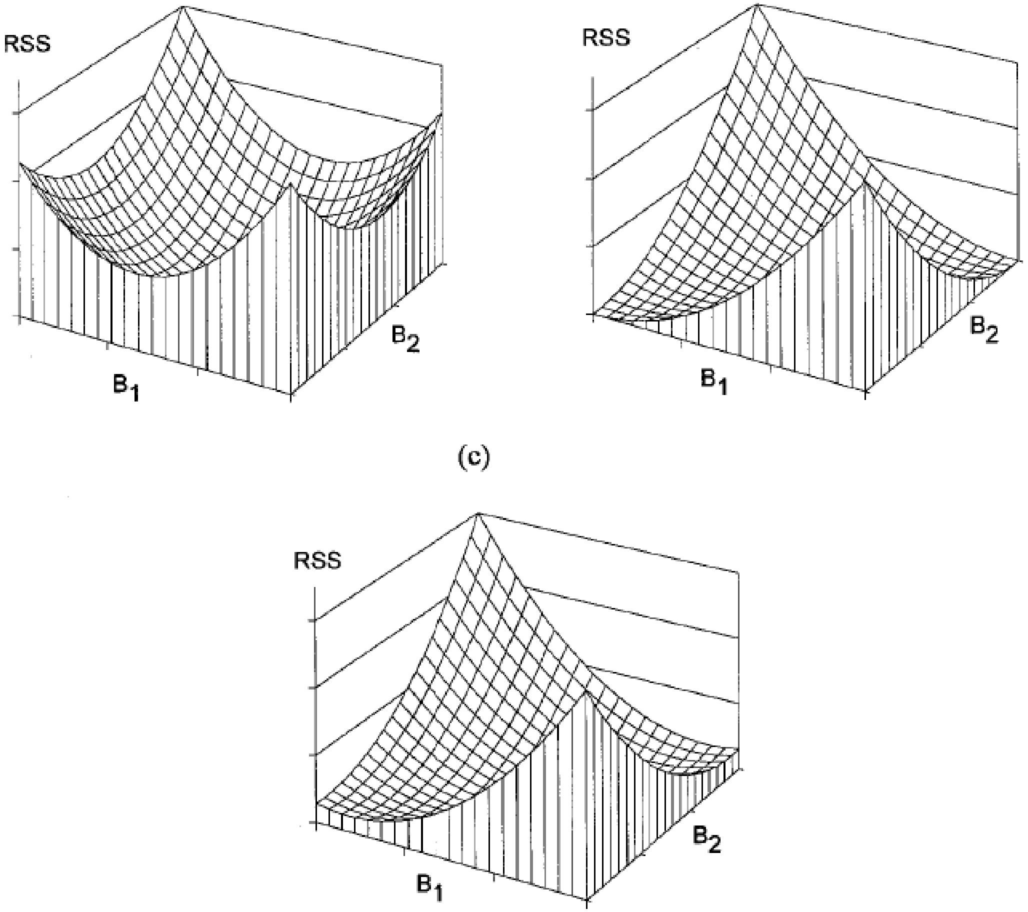
Example:

Gasoline consumption/Distance ~ Car Make + Car Age + Driver Age + Driver Gender + # of breaks per minute + ….

The more the better?

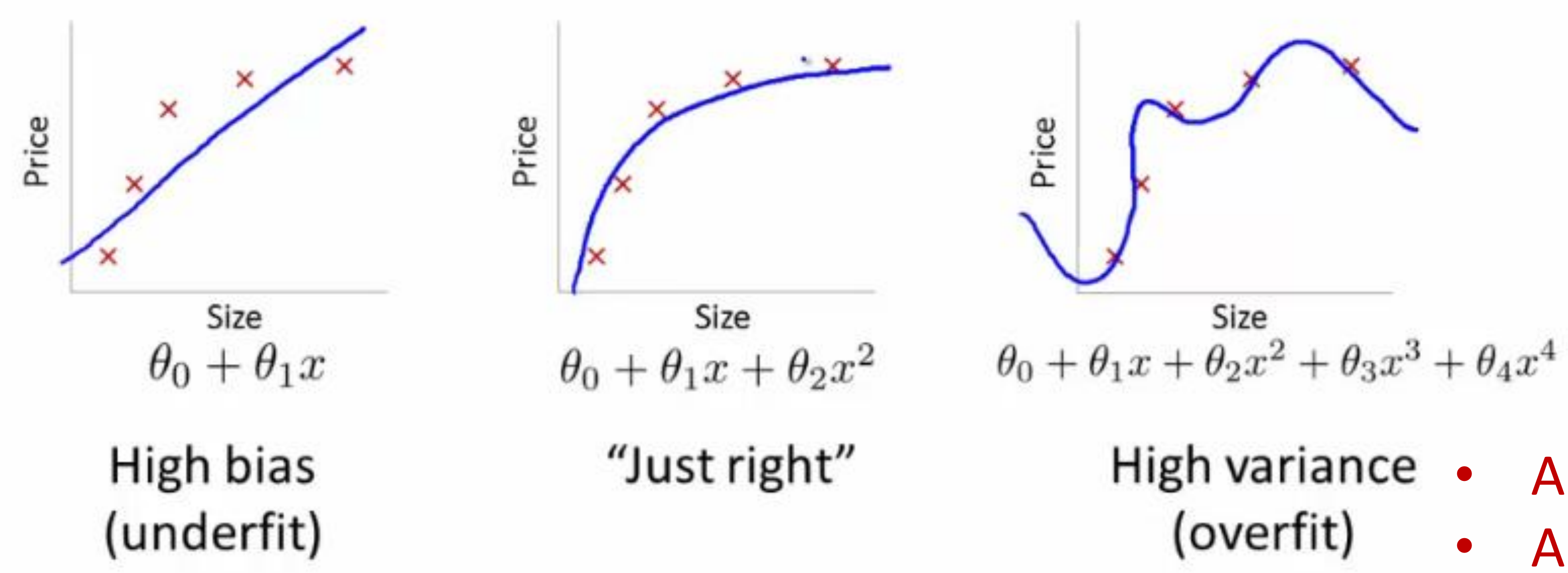
Multicollinearity

Variance Inflation Factor



(c)

Overfitting vs underfitting



- Adjusted R^2
- AIC
- BIC
- Predictive power

Categorical predictors

Gasoline consumption/Distance ~ Car Make + Car Age + Driver Age + Driver Gender + # of breaks per minute + ….

Dummy variables

$$X_i = \begin{cases} 1\ if\ Female \\ 0\ if\ male \end{cases}$$

Male: $Y_i = \alpha + \epsilon_i$

Female: $Y_i = \alpha + \beta_1 + \epsilon_i$

Categorical predictors

Gasoline consumption/Distance ~ Car Make + Car Age + Driver Age + Driver Gender + # of breaks per minute + ….

Dummy variables (Toyota, VinFast, Mercedes)

$$X_{i1} = \begin{cases} 1 \; if \; VinFast \\ 0 \; if \; other \end{cases} \qquad X_{i2} = \begin{cases} 1 \; if \; Mercedes \\ 0 \; if \; other \end{cases}$$

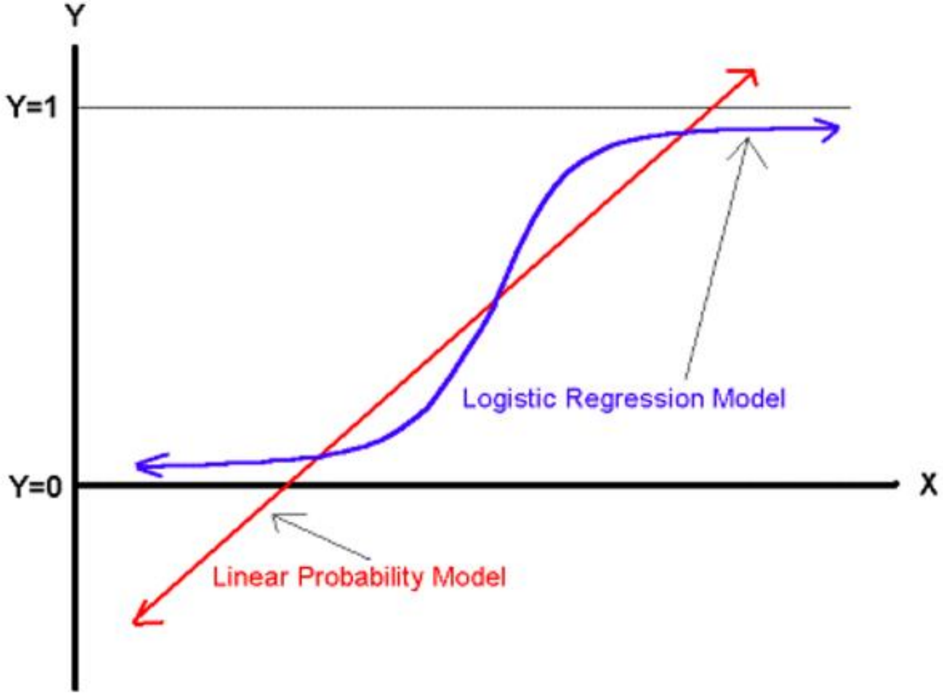
$$Y_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i$$

Toyota: $Y_i = \alpha + \epsilon_i$ VinFast: $Y_i = \alpha + \beta_1 + \epsilon_i$ Mercedes: $Y_i = \alpha + \beta_2 + \epsilon_i$

Advanced topics in quantitative methods

Generalized linear model

- General linear regression
  - Y: continuous/quantitative
- When Y is qualitative/discrete?
  - Binary (Yes/No): Logistic regression
  - Nominal: Multinominal logistic regression
  - Count data: Poisson regression



Multilevel model/Mixed effect model

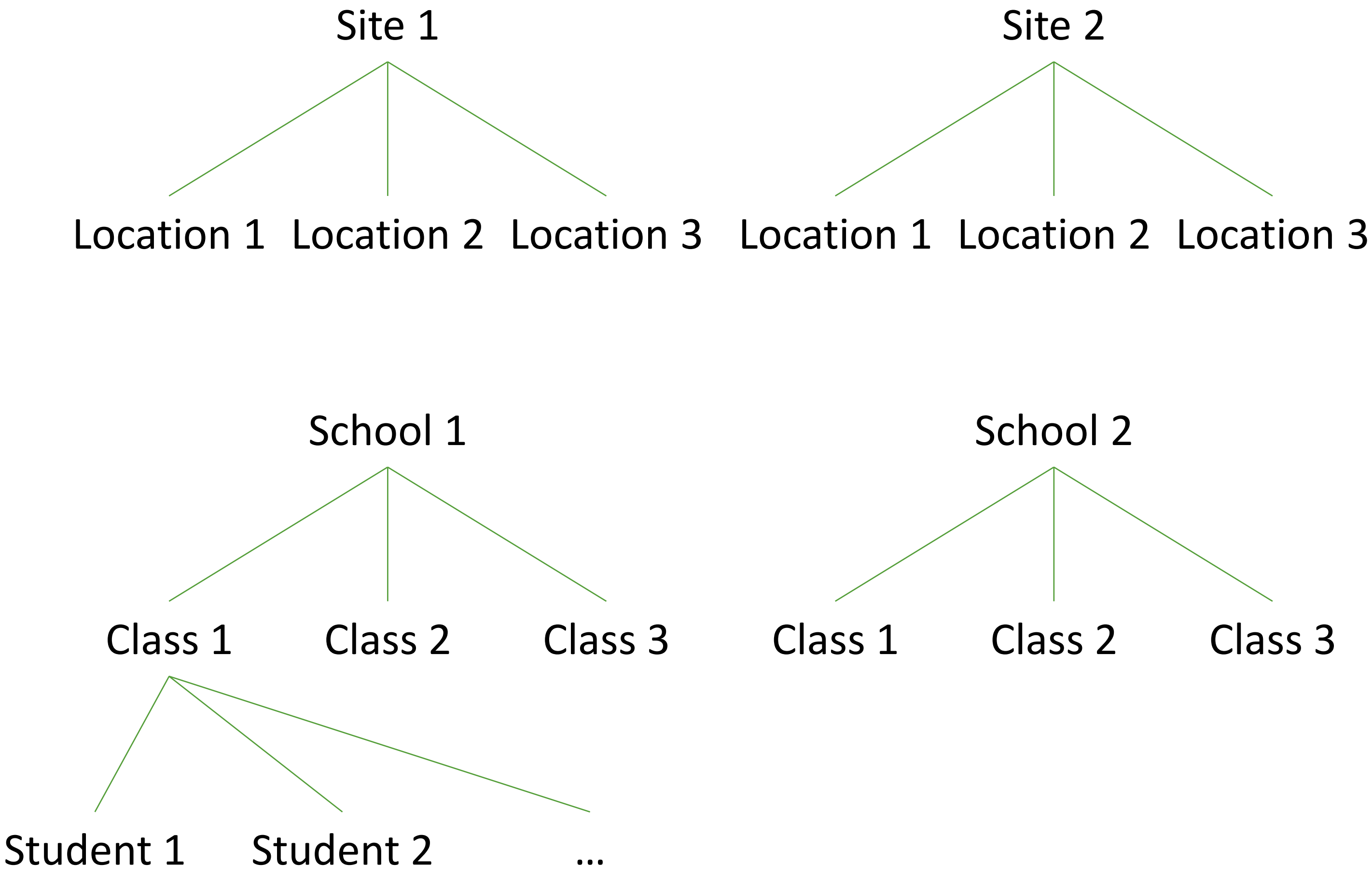
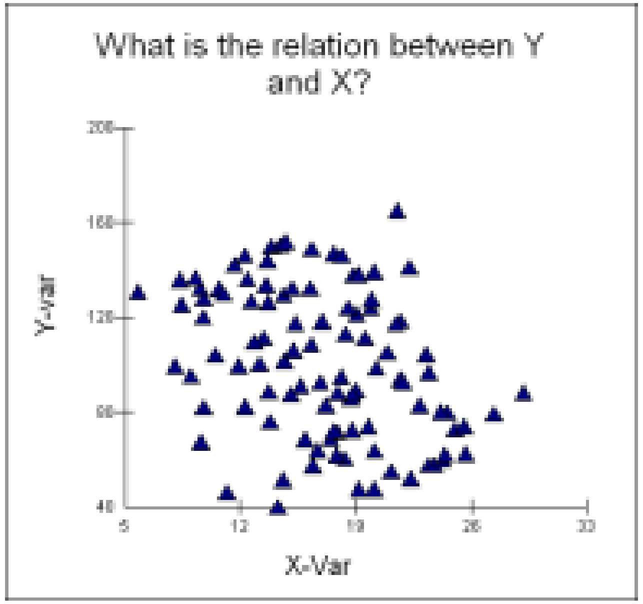
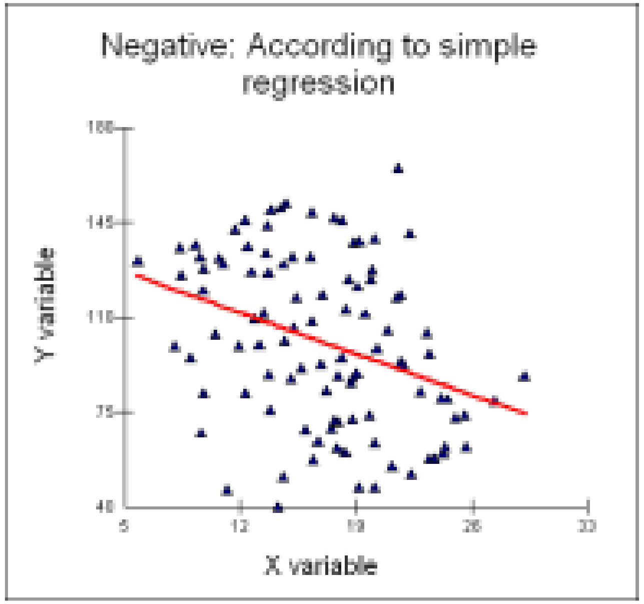


Image by Chelsea Parlett-Pelleriti

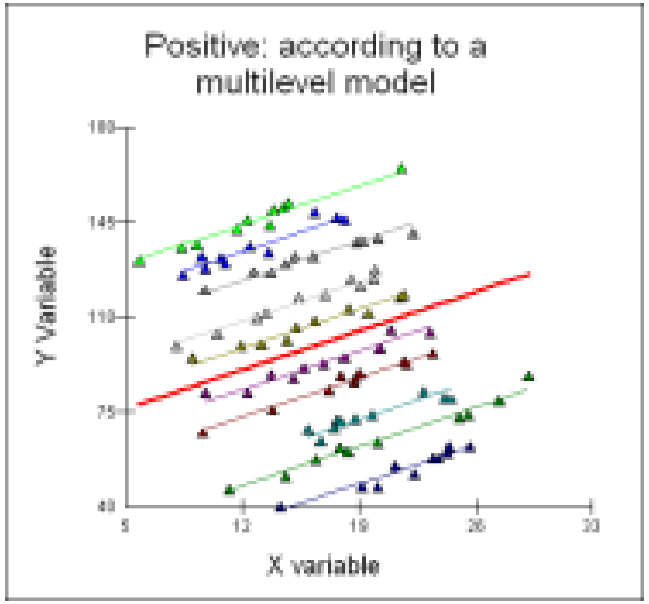
Multilevel model/Mixed effect model



What is the relation between Y and X?



Negative: According to simple regression

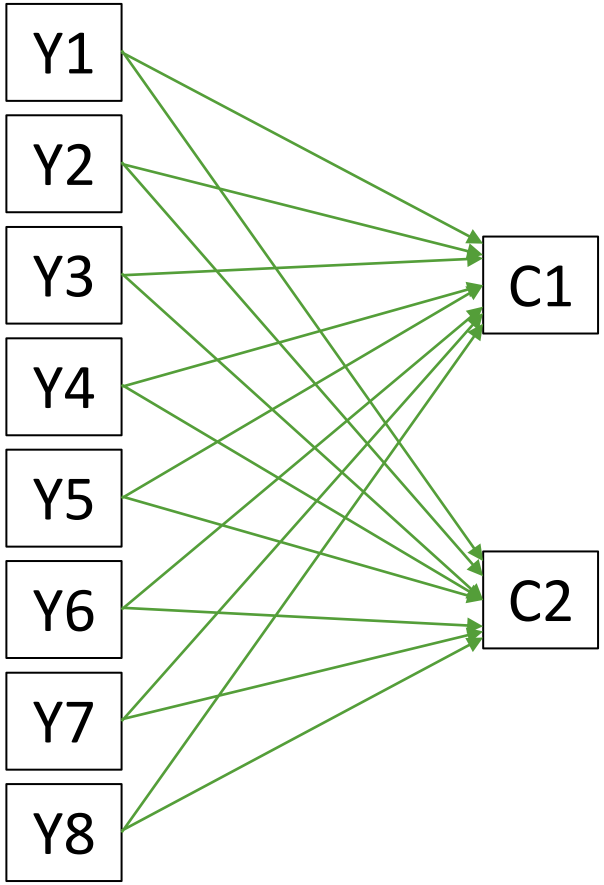


Positive: according to a multilevel model

Image by Chelsea Parlett-Pelleriti

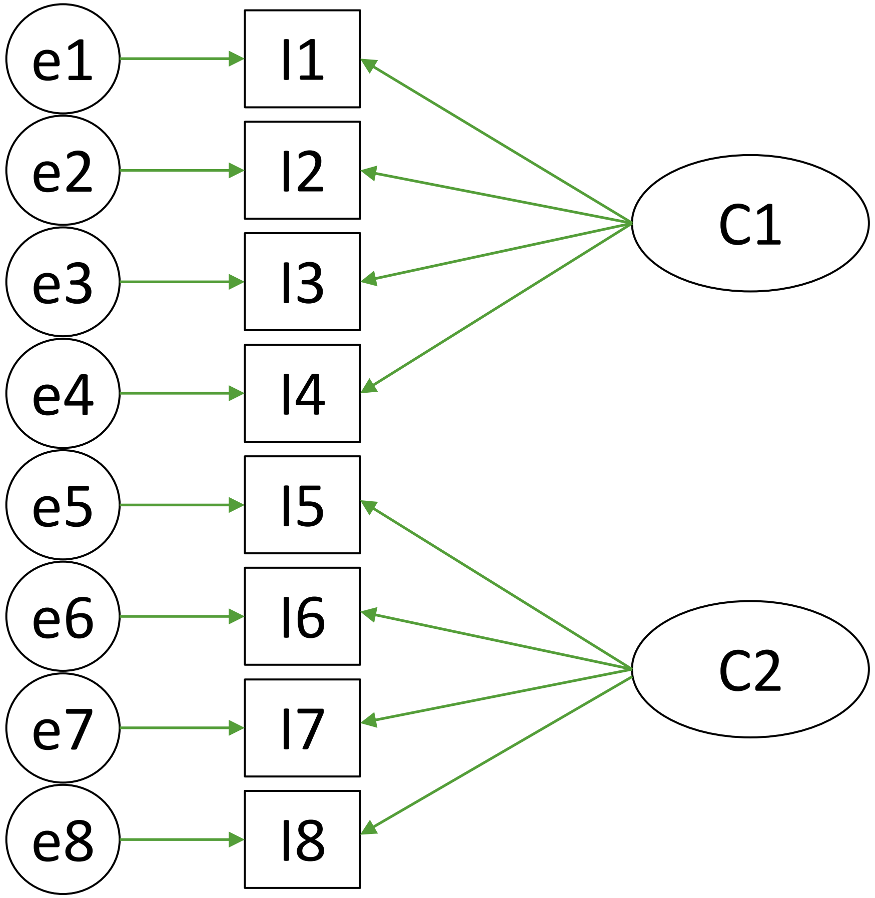
PCA

- Multidimensional reduction method
- No dependent/independent variables
- Exploratory method
- PCA – linear regression



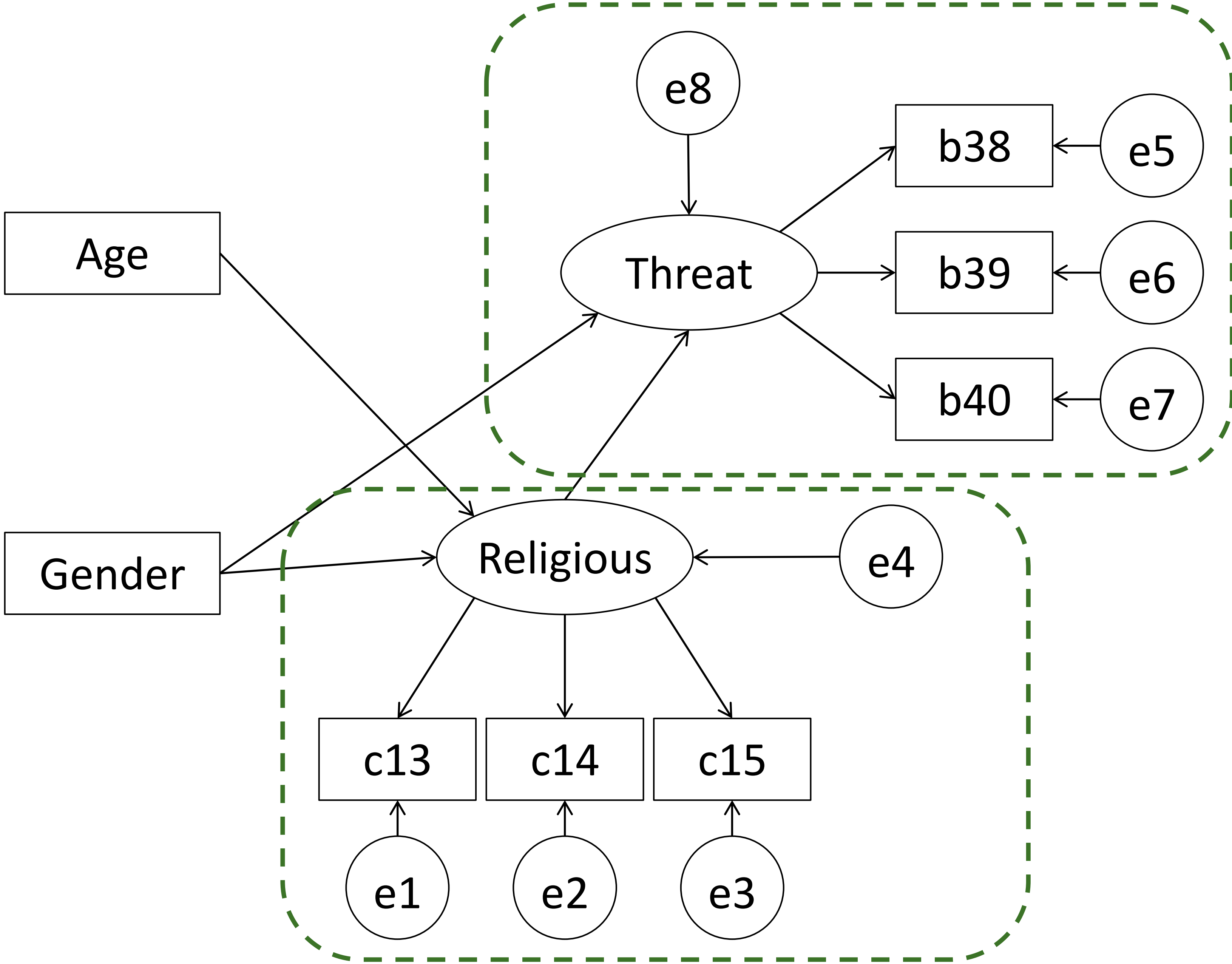
Factor Analysis

- Latent constructs
- Questionnaires
- Exploratory/Confirmatory



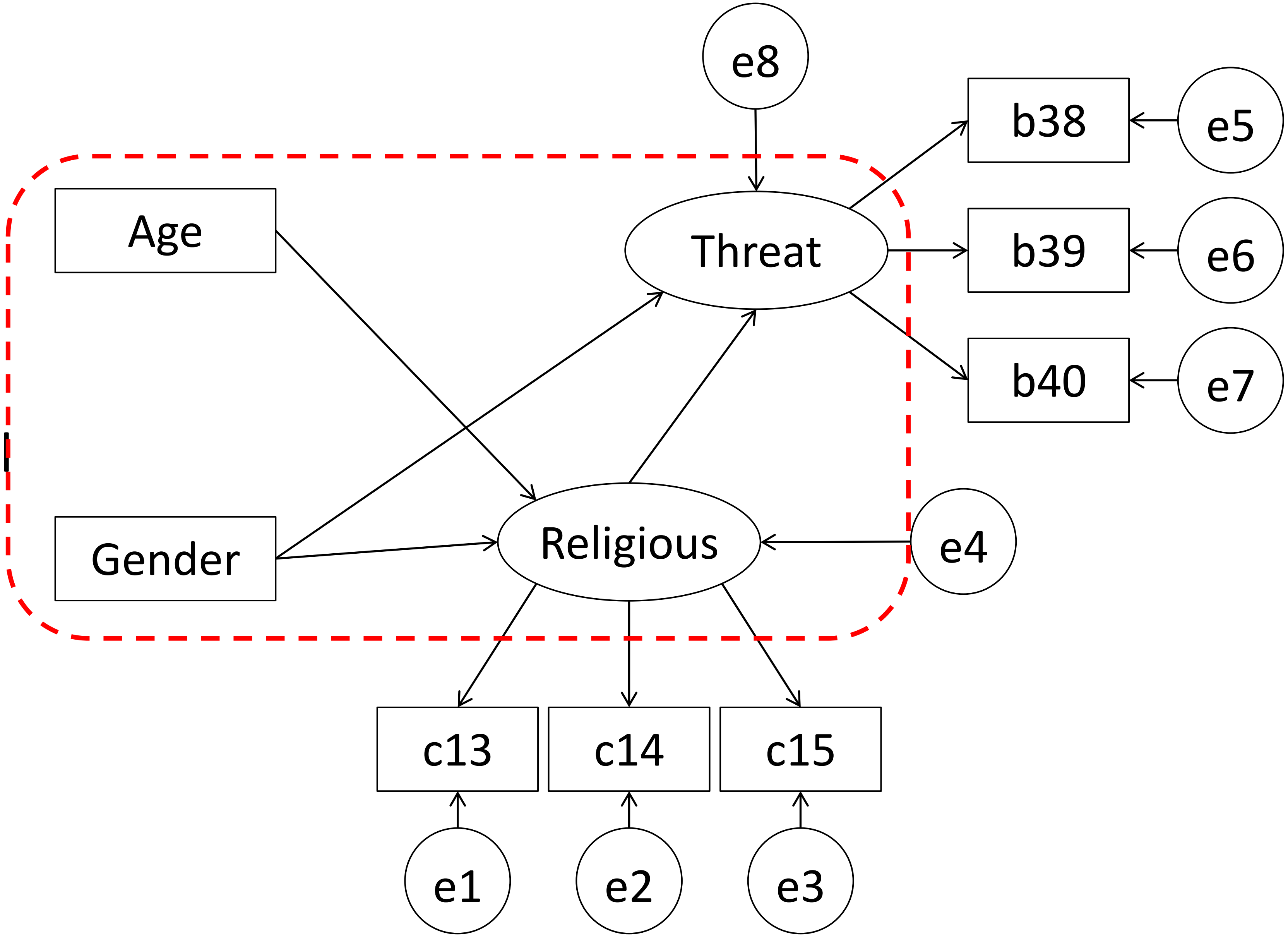
SEM

- Latent vs manifest
- Endogenous vs exogenous
- Measurement model vs structural model
- Direct vs indirect effects

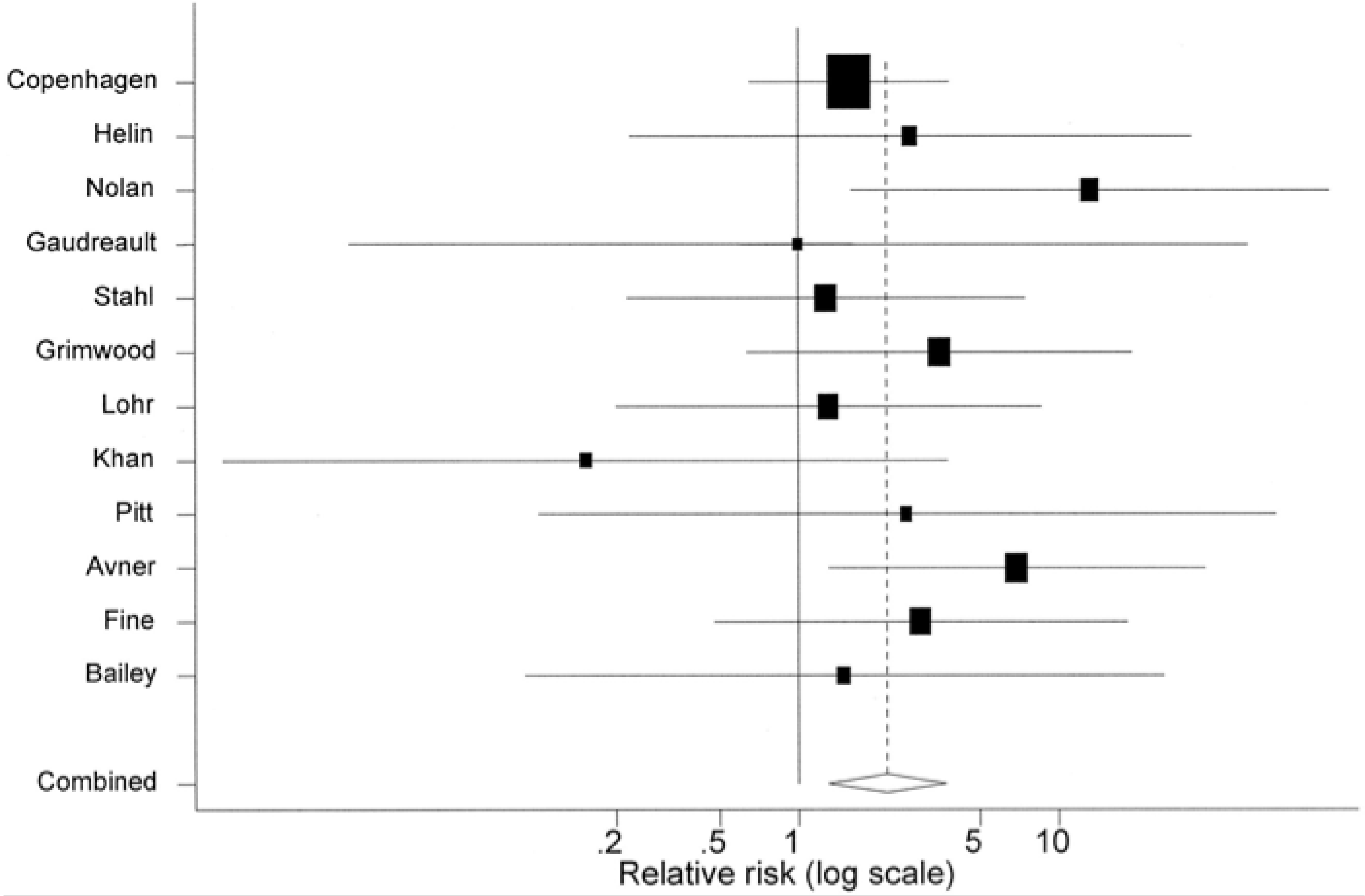


SEM

- Latent vs manifest
- Endogenous vs exogenous
- Measurement model vs structural model
- Direct vs indirect effects



Meta Analysis


Copenhagen
Helin
Nolan
Gaudreault
Stahl
Grimwood
Lohr
Khan
Pitt
Avner
Fine
Bailey
Combined
Relative risk (log scale)