Overview of three big cohort projects

ELSA, TILDA, & SHARE

English Longitudinal Study of Ageing (ELSA)

- 20 years, 19k participants
- 10 waves, ± every 2 years
- samples representative to people aged 50+ living in private households in England
- roughly every other wave limited medical assessment and collection of biological data (blood, saliva)

English Longitudinal Study of Ageing (ELSA)

Topics

- Demographic – age, marital stat, children, education
- Economic – income, pension, employment, housing
- Health – physical & mental, health behaviours, height + weight, COVID
- Psychosocial – social isolation, well-being, self-perception, cognitive functions

elsa English Longitudinal Study of Ageing

How to access?

- https://www.elsa-project.ac.uk/accessing-elsa-data
- (free) access through UK Data Service
- account needed, affiliation with UCD helps

The Irish Longitudinal Study on Ageing (TILDA)

- 15 years, 20k+ observations
- 6 waves, ± every 2 years
- samples representative to Ireland population aged 50+; household response rate 62%
- medical assessment and collection of biological data in Waves 1&3

tilda
Staidéar Fadaimseartha na hÉireann um Dhul in Aois

The Irish Longitudinal Study on Ageing

The Irish Longitudinal Study on Ageing (TILDA)

Topics

- Demographic – migration
- Economic – similar to ELSA
- Health – cardiovascular, executive functions, gait, balance
- Psychosocial – similar to ELSA

tilda
Staidéar Fadaimseartha na hÉireann um Dhul in Aois

The Irish Longitudinal Study on Ageing

How to access?

- https://tilda.tcd.ie/data/accessing-data/
- on-site (Dublin Center) hot desk with full data
  - €10k fee for full data access if you have available budget
- online access with partial data through Irish Social Science Data Archive (ISSDA)
  - signed data access form must be submitted by email

Survey of **H**ealth, **A**geing and **R**etirement in **E**urope (SHARE)

- 20 years, 28 European countries and Israel, 600k+ observations
- 9 waves, ± every 2 years
- samples theoretically representative to country population aged 50+; different sampling methods
- country-specific drop-off questionnaires with unique variables

SHARE
SURVEY OF HEALTH, AGEING AND RETIREMENT IN EUROPE

Survey of Health, Ageing and Retirement in Europe (SHARE)

Topics

- largely overlapping with ELSA and TILDA
- specific SHARELIFE questionnaire in Wave 3 on timeline of close personal relationships, partnerships
- specific COVID questionnaires in Wave 8 and 9

SHARE
SURVEY OF HEALTH, AGEING AND RETIREMENT IN EUROPE

How to access?

- https://share-eric.eu/data/become-a-user
- free for research purposes
- data access form must be submitted and approved to create an account
- conditions of use have extensive citation requirements
  - they change periodically, check both on start and end of your project

Watch out for variable availability throughout the projects

Gateway to Global Ageing Data

- https://g2aging.org/hrd/overview
- neat project, which records and harmonizes compatible variables across 11 big projects
- engine can search for comparable variables present in datasets
- very useful when working with data from multiple projects

Advanced statistical analyses

SURVIVAL ANALYSIS

LINEAR MIXED MODEL (MULTILEVEL)

Tips & tricks for working with large data

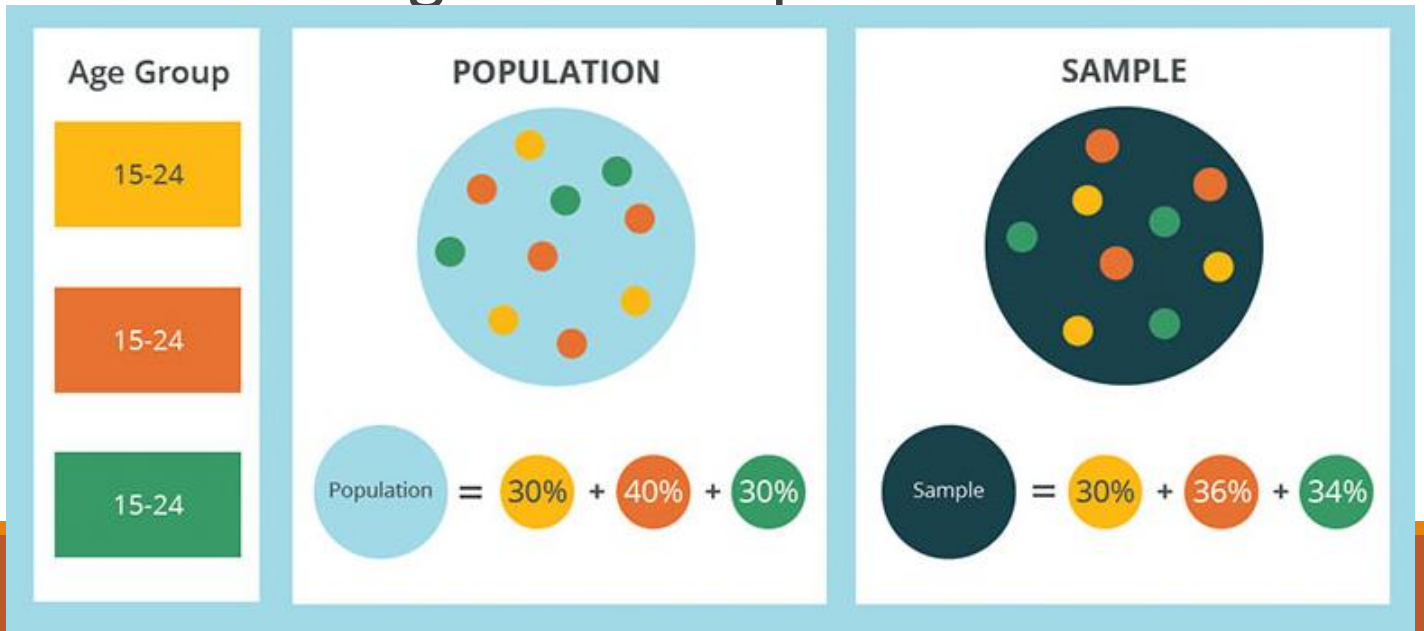
- always read the codebook / release guide first to get the sense of data
- do not rely on descriptions in data files
- often, recoded variables (e.g. total scale scores) are present

Table 9: Variables in the Technical Variables Module

Variable	Label
fam_resp	Family respondent
fin_resp	Financial respondent
hou_resp	Household respondent
*mn005_*	Single or couple interview
*mn016_*	Mother in household
*mn017_*	Father in household
*mn018_*	Mother-in-law in household
*mn019_*	Father-in-law in household
*mn024_*	Nursing home interview
*mn026_*	First respondent from couple or single
mn028	Eligible for dried blood spots collection (bs &gv_dbs)
mn029	Eligible for linkage
mn030	Eligible for social networks module (sn)
*mn031_*	Eligible for mini childhood module
*mn032_*	Eligible for social exclusion items
*mn038_*	Eligible for Accelerometry
*mn040_*	Need to ask consent question (ex123)
*mn041_*	Need to ask retirement info
*mn101_*	Questionnaire version (longitudinal vs. baseline)
*mn103_*	SHARELIFE life history interview (only w7)
*mn104_*	Household moved

Tips & tricks for working with large data

- have a strategy for dealing with participant missing data
- search for harmonized data if comparing countries
- weighting cases can support representativeness, but the logic of assigning your own weights is complex



Tips & tricks for working with large data

- learn how to merge datasets based on IDs and other variables
- learn how to switch wide and long data formats (Restructure in SPSS)

Wide Format

Team	Points	Assists	Rebounds
A	88	12	22
B	91	17	28
C	99	24	30
D	94	28	31

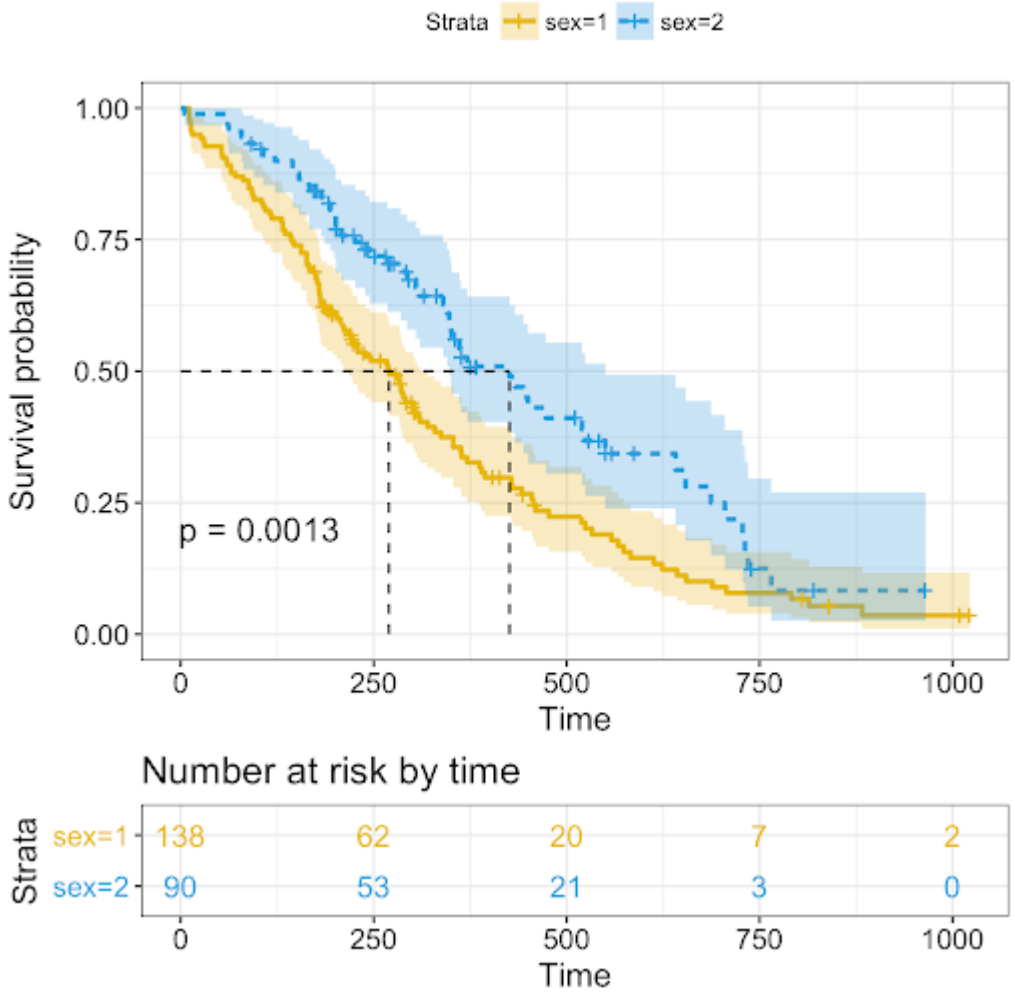
Long Format

Team	Variable	Value
A	Points	88
A	Assists	12
A	Rebounds	22
B	Points	91
B	Assists	17
B	Rebounds	28
C	Points	99
C	Assists	24
C	Rebounds	30
D	Points	94
D	Assists	28
D	Rebounds	31

Image source: https://www.statology.org/long-vs-wide-data/

Survival analysis

- Survival analysis operates with **time left** until a one-time event happens
- estimated odds are extrapolated across the whole observation period
- convertable to survival or hazard ratios
- Can handle multiple groups, or single-time covariates



Survival analysis

- similar to logistic regression

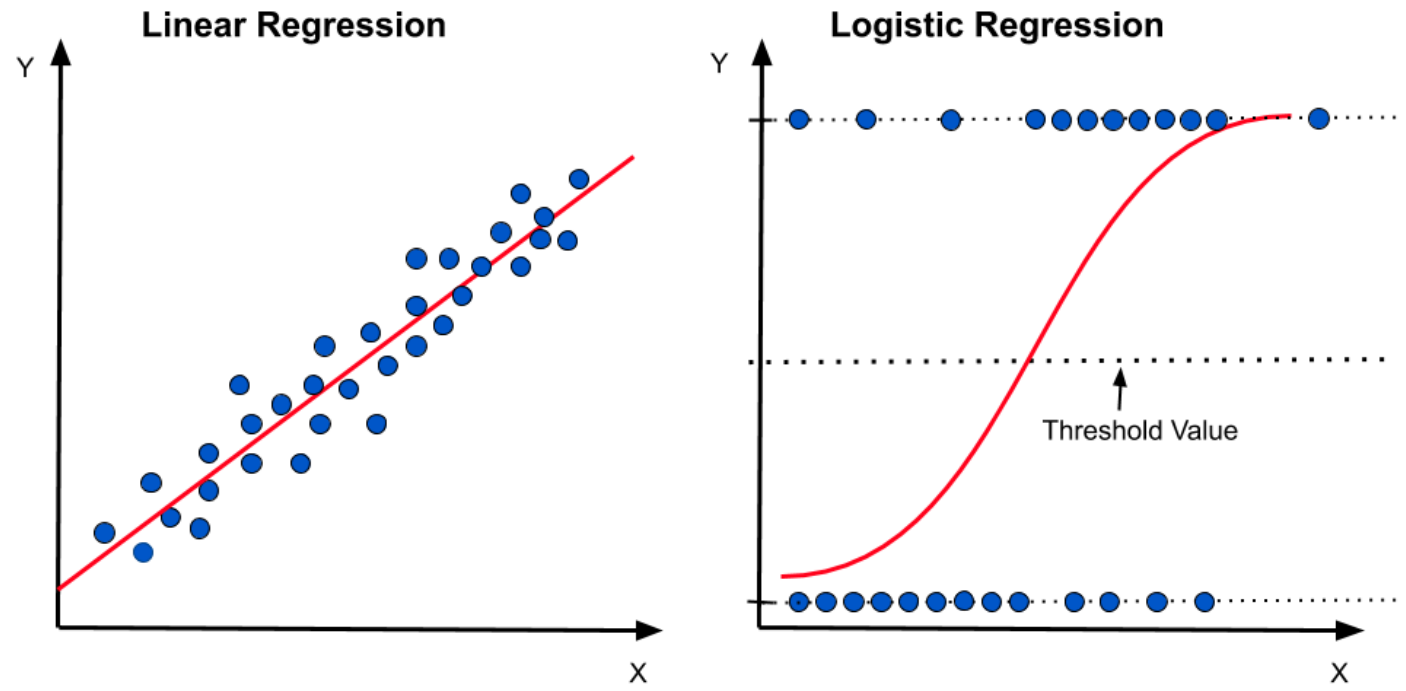


Image source: https://medium.com/@praveenraj.gowd/why-is-logistic-regression-called-logistic-regression-and-not-a-logistic-classification-5a418293040d

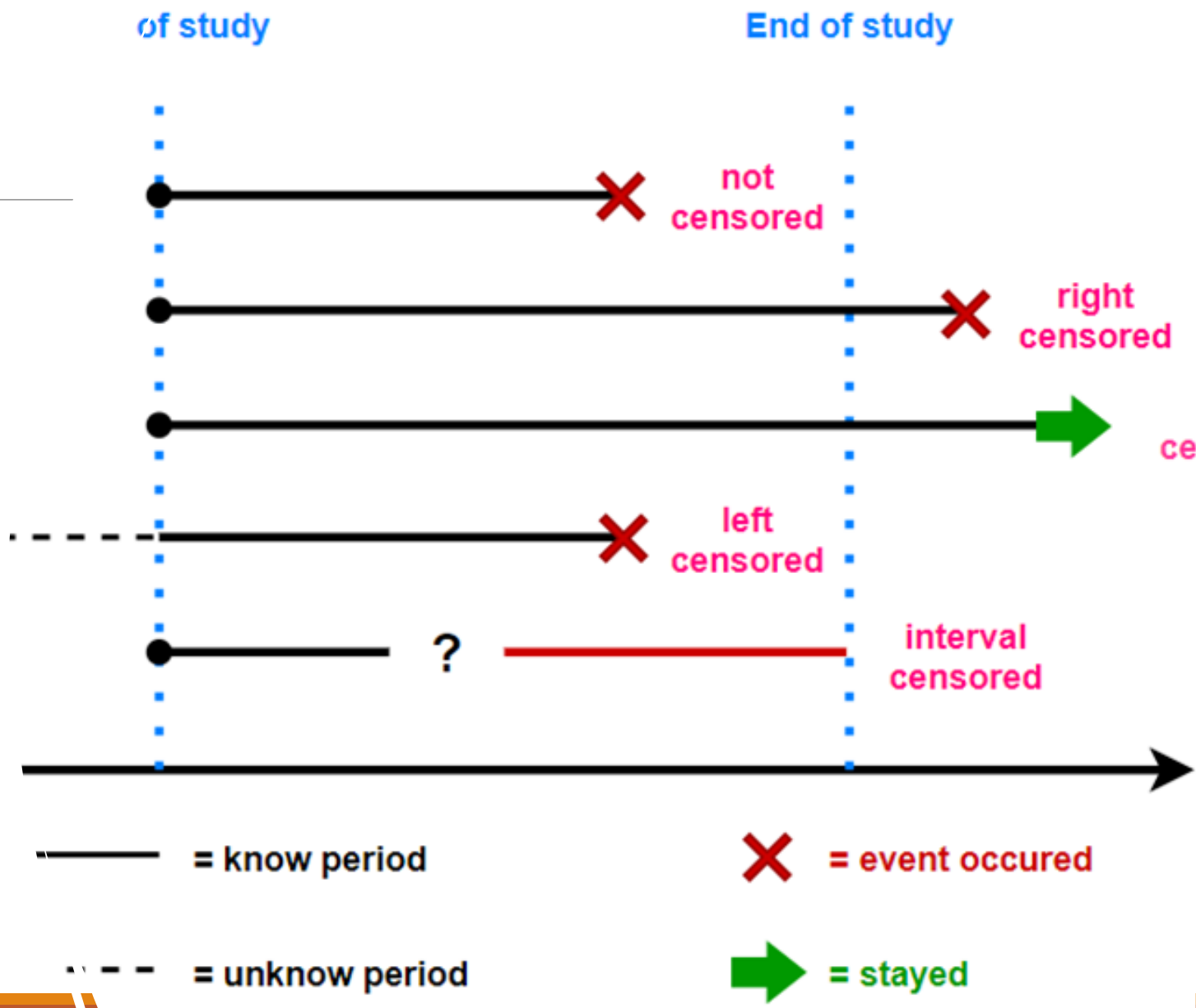
Survival analysis

- as in logreg, output are regression coefficients transformed into **Odds Ratios**
  - linreg coefficient(B) – if predictor value X increases by 1, by how many units does the Y change?
    - the coefficient value is additive; predicted value Y = intercept + X*(B)
  - logreg coefficient(also B) – if predictor value X increases by 1, how do the odds of Y happening change?
    - the OR value is multiplicative; predicted odds Y = exp(intercept) * X*exp(B)

Survival analysis - censorship

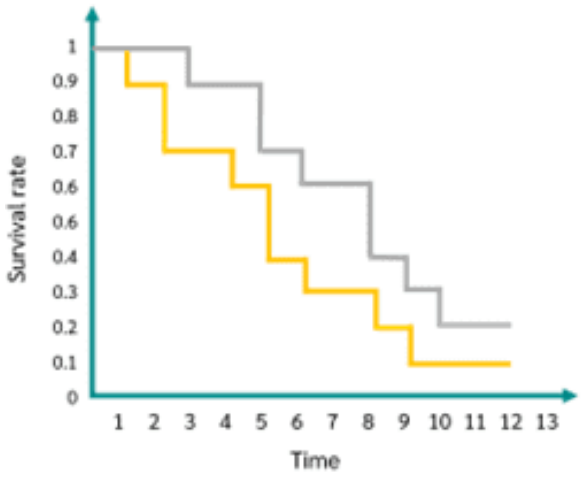
Ideally, we want clear starting and ending point of observing.

Problems with generalizing out of observed interval.

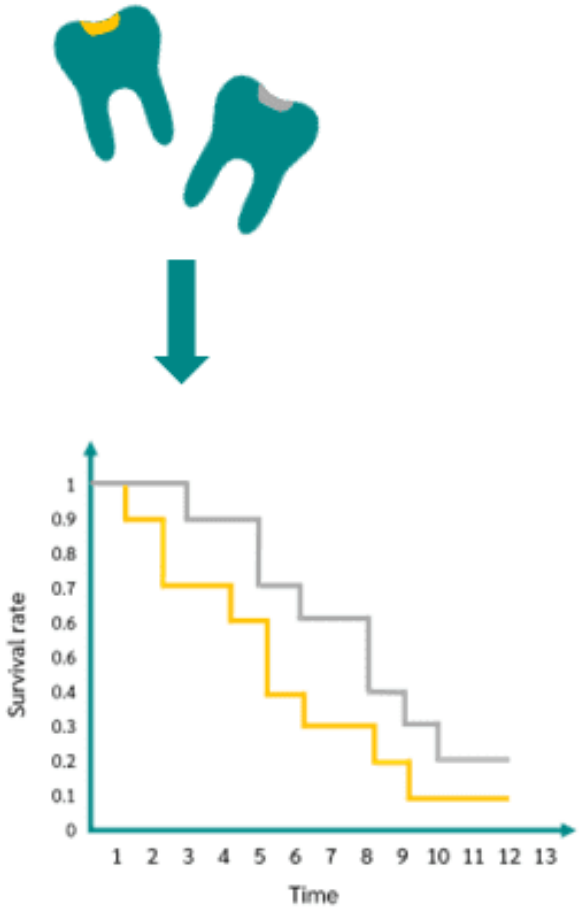


Survival analysis types

Kaplan Meier Curve



Log Rank Test



Cox Regression

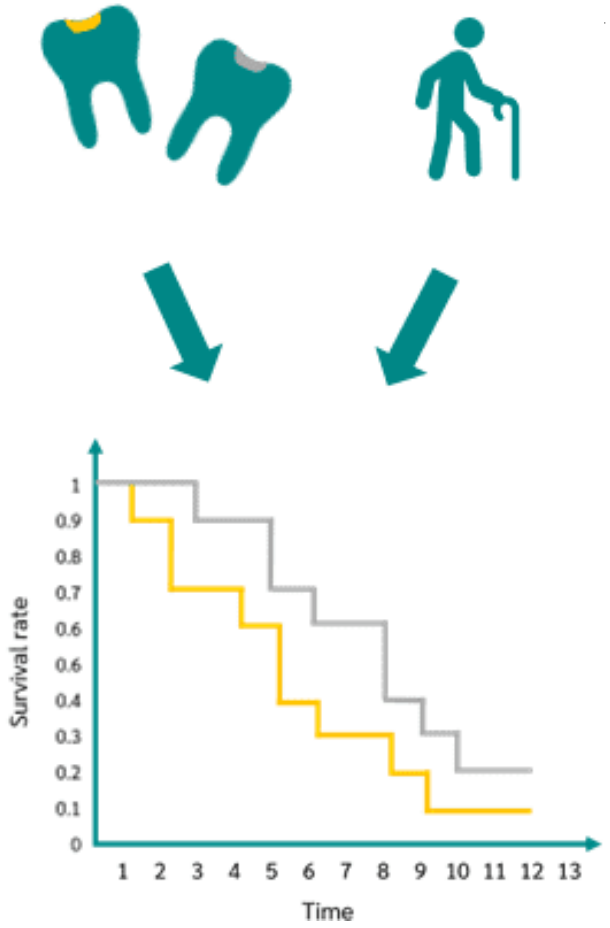


Image source: https://datatab.net/tutorial/survival-analysis

Cox regression results example

Factor	Group	N	HR	90% CI		*P* value
Sex	Male	195		1.000 (Reference)		
	Female	960	1.115	0.975	1.276	0.1816
Age	age < 40	203		1.000 (Reference)		
	40 ≤ age < 60	815	1.043	0.912	1.193	0.6087
	age ≥ 60	137	1.066	0.879	1.294	0.5861
family history	Non history	115		1.000 (Reference)		
	HCC	596	0.915	0.814	1.029	0.2130
	Cirrhosis	341	0.971	0.812	1.162	0.7908
	HBV	103	0.952	0.801	1.133	0.6441
Compensation status	Decompensation	345		1.000 (Reference)		
	Compensation	810	1.114	0.998	1.244	0.1058
time lengths from cirrhosis treatment to HCC onset	Non-treated	627		1.000 (Reference)		
	Total	528	0.810	0.721	0.911	0.0004
	LAM	102	1.045	0.873	1.250	0.6877
	ADV	181	0.792	0.687	0.914	0.0072
	ETV	83	0.716	0.585	0.877	0.0068
	LAM+ADV	95	0.822	0.684	0.989	0.0808
	Others	67	0.731	0.589	0.906	0.0166

Source: Bi, Jingfeng & Zhang, Zheng & Qin, Enqiang & Hou, Jun & Liu, Shuiwen & Liu, Zengmin & Li, Shuo & Wei, Zhenman & Zhong, Yanwei. (2017). Nucleoside analogs treatment delay the onset of hepatocellular carcinoma in patients with HBV-related cirrhosis. Oncotarget. 8. 10.18632/oncotarget.18075.

Linear mixed (multilevel) model

- very flexible, allows for designs with multiple measurements per participant (e.g., longitudinal)
- works similarly to regression, yields regression coefficients with the same meaning
- can also be used for nonlinear associations or categorical dependent variables, those models are called generalized

Linear mixed (multilevel) model

Many statistical tests in psychology are simplified cases of complex mathematical modeling techniques

Generalized Linear (Mixed) Model

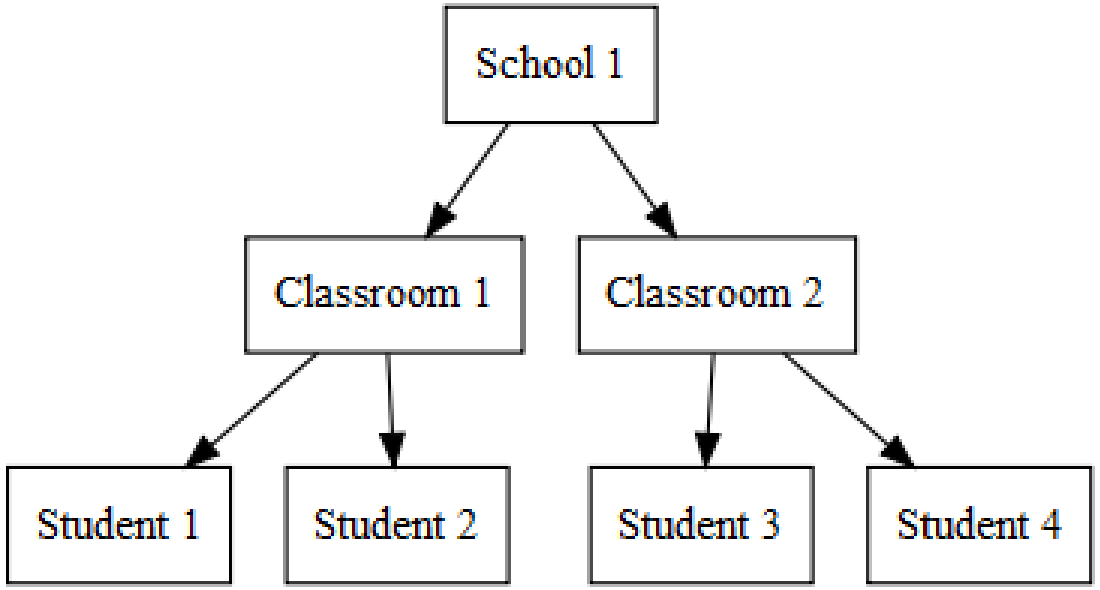
Regression

ANOVA

t-test

Multilevel models are similar to regression

- simple linear regression:
  - Y ~ b0 + b1*X
- simple linear mixed model
  - Y ~ b0 + b1*X + (1|Z)
    - Z is random effect factor, across which we want to „average“ the results, usually grouping variable – participant ID, strata, etc.



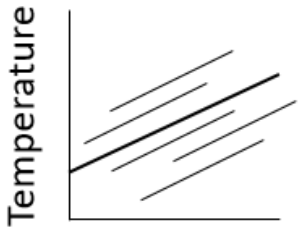
Linear mixed model results example

```
Formula: testScore ~ bodyLength2 + (1 | mountainRange)
   Data: dragons

Random effects:
 Groups        Name        Variance Std.Dev.
 mountainRange (Intercept) 339.7    18.43
 Residual                  223.8    14.96
Number of obs: 480, groups:  mountainRange, 8

Fixed effects:
            Estimate Std. Error t value
(Intercept)  50.3860     6.5517   7.690
bodyLength2   0.5377     1.2750   0.422
```

The random effect part tells you how much variance you find among levels of your grouping factor(s), plus the residual variance

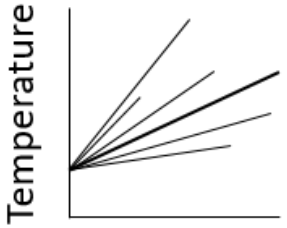
The fixed effect part is very similar to a linear model output:
Intercept and error
Slope estimate and error

Image source: https://ourcodingclub.github.io/tutorials/mixed-models/



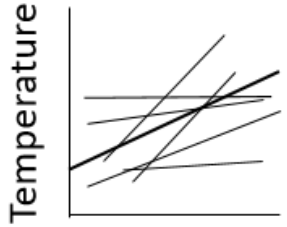
(1|SiteID)

- Different sites have different intercepts
- The slope of temperature over time is the same for all sites



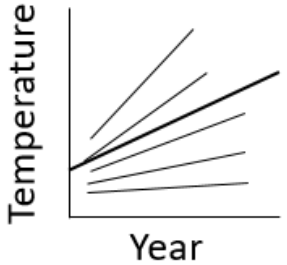
(0 + Year|SiteID)

- The intercept is the same for all sites.
- The slope of temperature over time varies among sites



(0 + Year|SiteID) + (1|SiteID)

- No covariance between the intercept and the slope.
- Statistically easier to compute
- Is this realistic or not? If there is no reason to expect it then why model it? If you have a massive dataset then may as well model the covariance.



(1 + Year|SiteID) + (1|SiteID)

- If there is a positive covariance, then a lower intercept would mean that the slope for that site would be lower.
- Note: if the correlation value reported is 1 or -1 then it suggests there are issues calculating this in the model and it can't estimate it well.

Image source: https://ninamclean.weebly.com/statistics-and-other-bits-and-pieces/random-intercepts-and-slopes-guide

Linear mixed (multilevel) model

- (RE)ML estimator, nested models easily comparable
- sensitive to misspecification of covariances
- able to construct complex models
- gives fixed effects, then the variability of dependent variable across groups (random effects)

Final Comparison

Survival Analysis

- best to model continuous time until the occurence of a one-time event
- bad flexibility of the model, cannot handle groups within groups well or time-varying covariates; expects proportional hazard ratios
- requires specific data structure

Multilevel Analysis

- great flexibility in predicting dependent variable across multiple observations and/or groups; quite benevolent about missing data
- needs clearly defined assumptions regarding the covariance structure
- requires long data structure