WASHINGTON STATE UNIVERSITY

Analysis in Epidemiology

Contact Information & Course Website

- Office: Allen Center 311
- Office Hours: By Appointment
- Email: Eric.Lofgren@wsu.edu

Course material will be posted on GitHub at: https://github.com/epimodels/AnalysisInEpi

Generally I will try to have material up the day before class.

Purpose of this Course

- Build a functional vocabulary in regression models
- Let you engage with the literature, work on your own problems, etc.
- Know when and how to get help

Some Resources for the Curious

- *Modern Epidemiology 3rd Edition*
- stats.stackexchange.com
  - Generally speaking posting homework questions there is discouraged, but this can be a valuable resource for the future
- CISER
- The usual advice: Get help EARLY!

CISER
Center for Interdisciplinary Statistical
Education and Research

Questions?

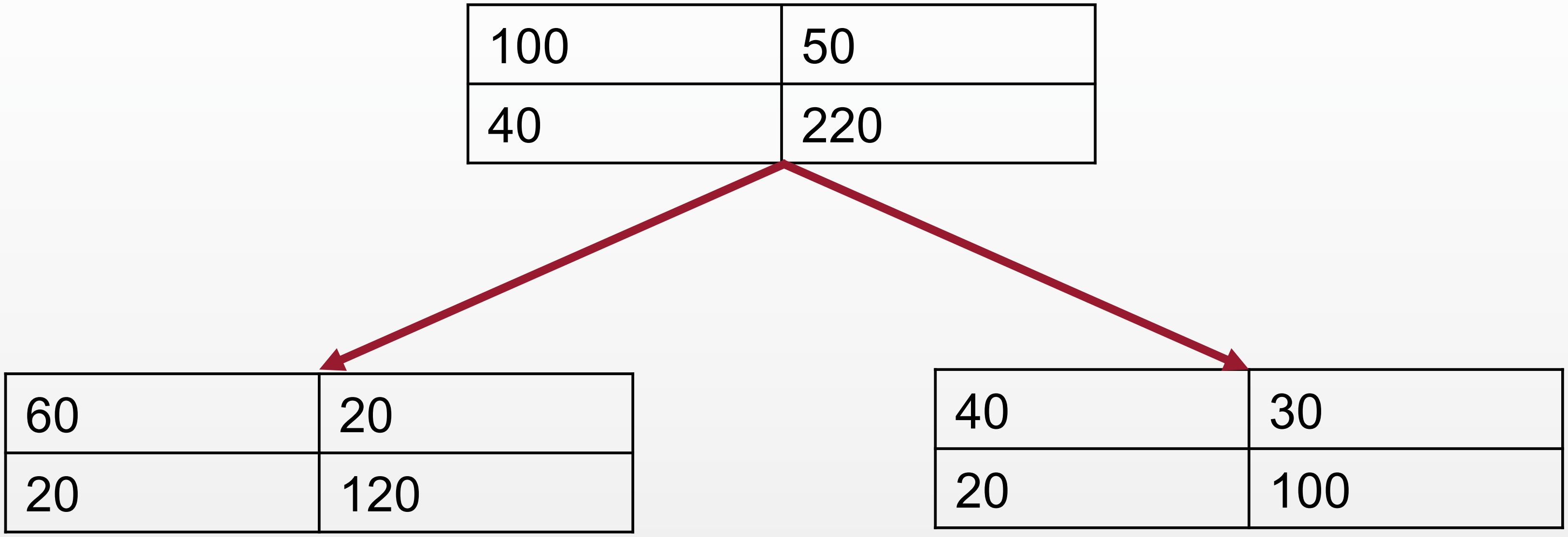
Types of Observational Data

- Most of these got covered in Module 1
- Binary/Continuous
- People/Time/Events
- The decision on *how* to model is largely dictated by *what* you are modeling

Why Regression?

- Module 1 taught you how to calculate measures of association with 2x2 tables, long division, etc.
- Adjusting for confounding using stratification
- All of this seemed to work well enough, can be done in Excel, on a whiteboard, etc. – why bother with regression?

Stratification

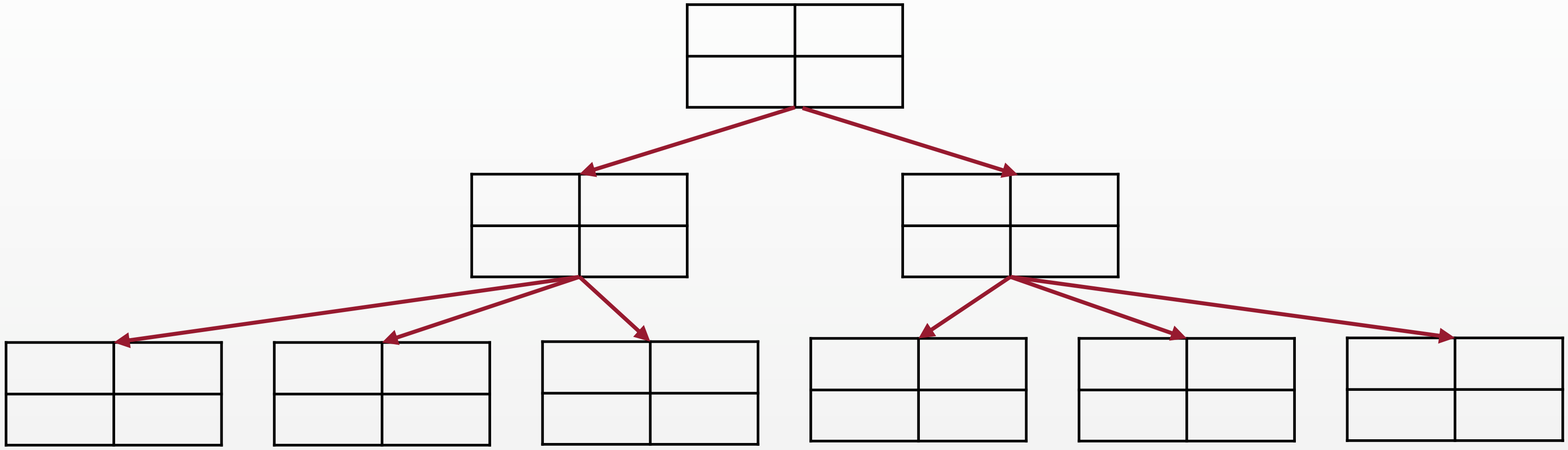


100	50
40	220

60	20
20	120

40	30
20	100

Stratification



And this is just for two variables…what about 16?

Strengths of Regression

- Can handle adjustment by *many* variables
- Can handle non-categorical data
- Can smooth/spackle over empty cells
  - If you know what happens to 26 year olds and 28 year olds, you can guess what happens to 27 year olds
- You can predict
  - Given m, X and b, solve for Y
  - Regression is the foundational toolset of machine learning/data science

What Regression Isn’t

Regression Can't…

- Automatically fix bad data collection
- Control for bias that it (or you) don't know about
- Solve your sample size problems for you
- …solve *any* of your problems **for** you – regression is a tool, and a dumb one at that

Assumptions and Problems of Regression

- Positivity: An individual has a non-zero probability of having any combination of parameter values
  - Regression assumes cells with 0's happened by chance – what if those cells are impossible?
- Model misspecification: Missing confounders, the wrong distribution, etc. will give you the wrong answer
  - This is, I would argue, the biggest problem in observational epi
- Nonidentifiability: Two (or more) combinations of parameters are equally supported by the data, and there is no "best fit"
- Others we will discuss as the class goes on
- There are *more* assumptions necessary for causal inference, which is beyond the scope of this module

Reading a Regression Equation

- Regression is essentially progressively more complex versions of y = mx +b

$$Y = \underbrace{\beta_0 + \beta_1 A + \varepsilon}_{\text{Linear Predictor}}$$

$$Y = \alpha + \beta_1 A + \gamma \boldsymbol{Z}$$

What’s a Link Function?

- A link function is a function that describes the relationship between Y and the rest of the equation
- Linear Predictor: $\mathbf{X}\beta$
- Link function: $g(\mathrm{Y}) = \mathbf{X}\beta$
- Identity: $\mathrm{Y} = \mathbf{X}\beta$
- Log: $\ln(\mathrm{Y}) = \mathbf{X}\beta$
- Logit: $\frac{Y}{1-Y} = \mathbf{X}\beta$

What is a Distribution?

- Linear regression assumes things came from a normal distribution
- This is *often* not true
- Other distributions are common
- Binomial: Binary data
- Poisson/Negative Binomial: Counts and rates
- Exponential/Weibull/Gamma: Time
- When unspecified, it is often assumed to be normal

Least Squares and Maximum Likelihood

- Two ways to estimate the best fitting parameter
- Linear regression often uses least squares
- Most of the other models we will discuss use some form of maximum likelihood