WELCOME TO NDACAN MONTHLY OFFICE HOURS!

NATIONAL DATA ARCHIVE ON CHILD ABUSE AND NEGLECT
DUKE UNIVERSITY, CORNELL UNIVERSITY, & UNIVERSITY OF CALIFORNIA: SAN FRANCISCO

NDACAN

Children's Bureau
An Office of the Administration for Children & Families

- The session will begin at 11am EST
  - 11:00 - 11:30am – LeaRn with NDACAN (Introduction to R)
  - 11:30 - 12:00pm – Office hours breakout sessions
- Please submit LeaRn questions to the Q&A box
- This session is being recorded.
- See ZOOM Help Center for connection issues: https://support.zoom.us/hc/en-us
  - If issues persist and solutions cannot be found through Zoom, contact Andres Arroyo at aa17@cornell.edu.

LEARN WITH NDACAN

Presented by Frank Edwards

MATERIALS FOR THIS COURSE

- Course Box folder (https://cornell.box.com/v/LeaRn-with-R-NDACAN-2024-2025) contains
  - Data (will be released as used in the lessons)
    - Census state-level data, 2015-2019
    - AFCARS state-aggregate data, 2015-2019
    - AFCARS (FAKE) individual-level data, 2016-2019
    - NYTD (FAKE) individual-level data, 2017 Cohort
  - Documentation/codebooks for the provided datasets
  - Slides used in each week's lesson
  - Exercises as that correspond to each week's lesson
  - An .R file that will have example, usable R code for each lesson – will be updated and appended with code from each lesson

WEEK 7: INTRODUCTION TO REGRESSION

April 18, 2025

DATA USED IN THIS WEEK'S EXAMPLE CODE

- AFCARS fake aggregated data ./data/afcars_aggreg_suppressed.csv
  - Simulated foster care data following the AFCARS structure
  - Can order full data from NDACAN:
    - https://www.ndacan.acf.hhs.gov/datasets/request-dataset.cfm
- Census data ./data/
  - Full data available from NIH / NCI SEER

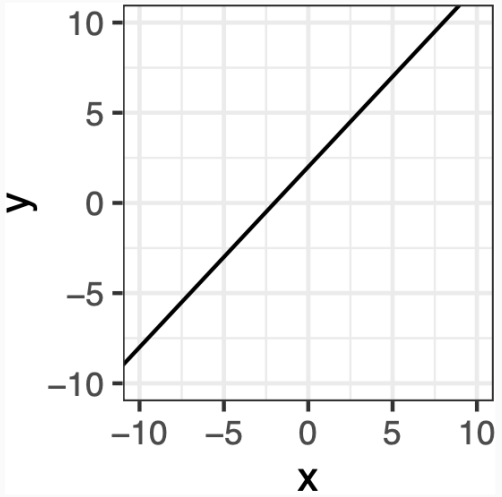
LINEAR REGRESSION 101

LINES

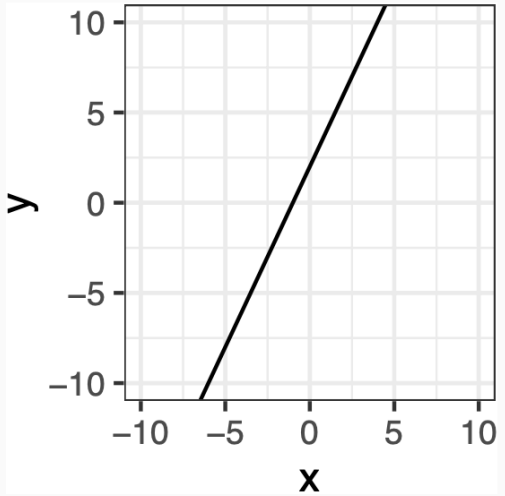
- We can define a line as: $y = mx + b$
- Where *m* is the slope and *b* is the y-intercept.

PLOTS

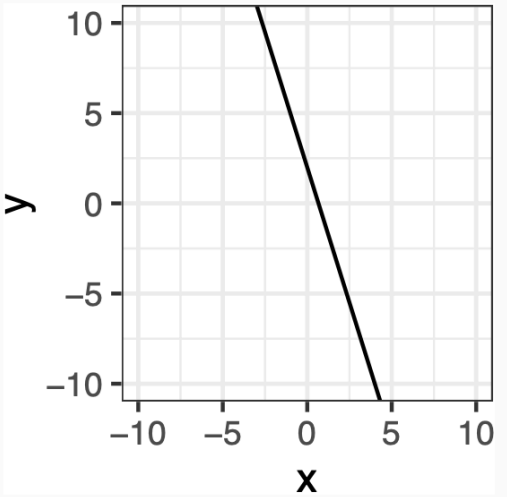
$y = x + 2$



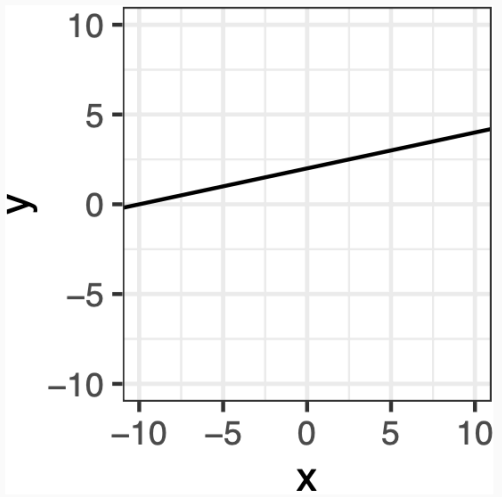
$y = 2x + 2$



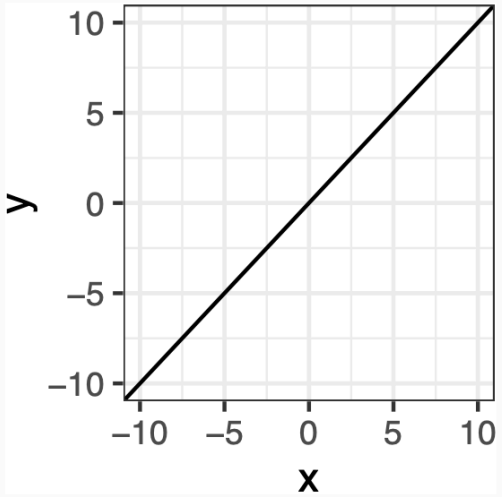
$y = -2x + 2$



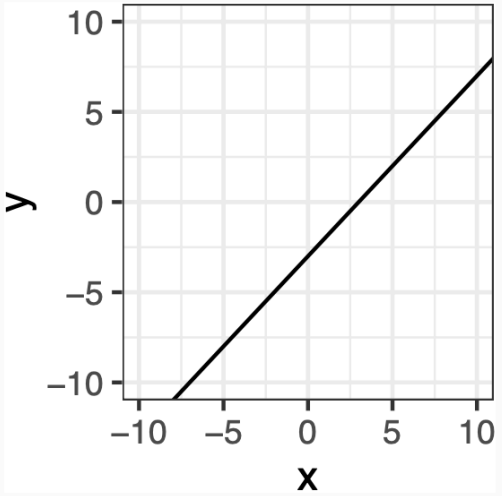
$y = 0.2x + 2$



$y = x + 0$



$y = x - 3$



THE LINEAR REGRESSION MODEL

We can describe the relationship between a predictor variable *x* and an outcome variable *y* with a line:

$$E(y) = mx + b$$

Where the parameter *m* describes the expected change in *y* when *x* changes by one unit.

The parameter *b* describes the expected value of *y* when $x = 0$

NOW WITH GREEK!

We can describe the relationship between a predictor variable *x* and an outcome variable *y* for unit *i* with the linear regression model:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$
$$\varepsilon \sim \text{Normal}(0, \sigma^2)$$

In plain English, these equations say: The variable y for unit i is equal to beta zero plus beta 1 times the variable x for unit i, plus an error term. That error term follows a Normal distribution with variance sigma squared.

REGRESSION ASSUMPTIONS FOR PREDICTION

- Correct functional form (linearity in y as a function of x)
- Errors follow a Normal distribution centered at zero (identically distributed errors)
- Errors are not correlated with each other (independent errors)

- *Additional assumptions are required for causal inference!!*

OVER TO RSTUDIO