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 8: INTRODUCTION TO ADVANCED MODELING

May 16, 2025

DATA USED IN THIS WEEK'S EXAMPLE CODE

- AFCARS fake aggregated data ./data/afcars_aggreg_suppressed.csv
- AFCARS fake individual-level data ./data/afcars_2016_2019_indv_fake.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

GENERALIZED LINEAR MODELS

FROM LINEAR REGRESSION

We define the expected value of y in a linear regression model as

$$E[y_i] = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$$

TO GENERALIZED LINEAR MODELS

We can define a generalized linear regression for the expectation of y as

$$g(E[y_i]) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$$

For a wide range of *link functions* g

TWO COMMONLY USED LINK FUNCTIONS

Logistic regression

$$logit(E[y_i]) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$$

Poisson regression

$$log(E[y_i]) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$$

CHANGES TO THE ERROR STRUCTURE

Linear regression

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$
$$\varepsilon_i \sim N(0, \sigma^2)$$

Logistic regression

$$p = logit^{-1}(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i})$$
$$y_i \sim Bernoulli(p)$$

Poisson regression

$$\lambda_i = e^{\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}}$$
$$y \sim Poisson(\lambda)$$

ESTIMATING GLMS IN R

We specify the distributional ‘family’ to be used with the glm() function

Logistic regression: glm(y ~ x, data = mydata, family = “binomial”)

Poisson regression: glm(y ~ x, data = mydata, family = “poisson”)

MULTILEVEL MODELS

FROM LINEAR REGRESSION

We can define a multilevel model with varying intercepts for cluster j as

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \delta_j + \varepsilon_i$$
$$\varepsilon \sim N(0, \sigma^2)$$
$$\delta \sim N(0, \sigma^2)$$

ESTIMATION IN R

The lme4 package provides a flexible set of multilevel models

Linear, multilevel slopes:

lmer(y ~ x + (1|j), data = mydata)

Logistic, multilevel slopes and intercepts:

glmer(y~x + (i|j), data = mydata, family = “binomial”)

Cheat sheet for lme4 formulas:
https://stats.stackexchange.com/questions/13166/rs-lmer-cheat-sheet

OVER TO RSTUDIO