

Statistical Inference with Linear & Logistic Regression

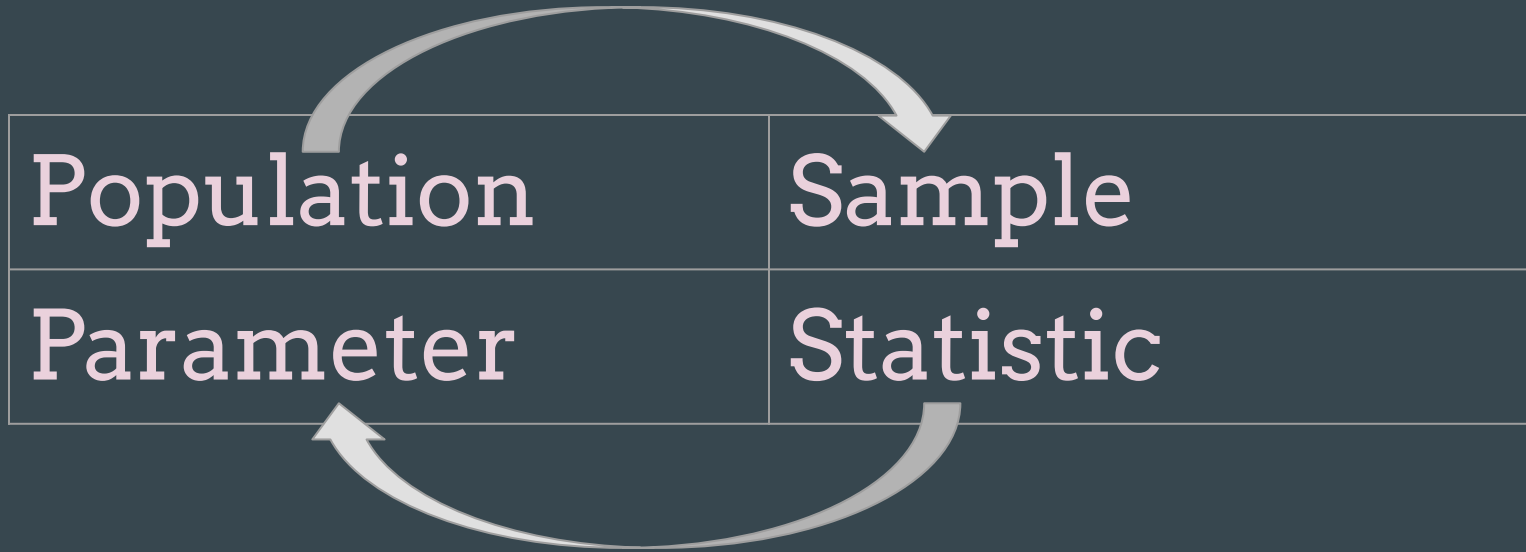


GW Libraries Workshop
Dan Kerchner ~ Fall 2024

go.gwu.edu/rstats

Super-Brief Review of Inference for Regression

High-Level Objective



Regression Model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n$$

(and an interaction term might look like $\beta_{12} X_1 X_2$)

Interpretation

β_0 = Mean Y when X_i values are 0

β_i = Mean change in Y for a 1-point increase in X_i ,
adjusting for other X variables

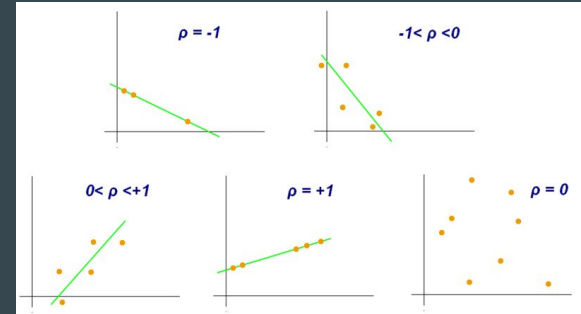
Correlation between dependent & independent variables

Pearson correlation (ρ) measures strength and direction (+/-) of linear association between X_i and Y . Ranges from -1 to 1.

If relationship looks non-linear (but monotonic) then **Spearman** correlation should be considered.

$$H_0: \rho = 0$$

$$H_1: \rho \neq 0$$



Valid if joint distribution of X, Y is bivariate normal

Linear Regression Model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n$$

where Y is a continuous outcome

Interpretation

β_0 = Mean Y when X_i values are 0

β_i = Mean change in Y for a 1-point increase in X_i ,
adjusting for other X variables

Linear Regression Assumptions

- Observations are independent
- Linearity
- Homoscedasticity
- Normality

GLMs - Generalized Linear Models

$$g(\mu) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n$$

where $\mu = E(Y)$

Common link functions

$$g(\mu) = \mu$$

$$g(\mu) = \log(\mu)$$

$$g(\mu) = \log(\mu / (1 - \mu)) = \text{logit}(\mu)$$

(Binary) Logistic Regression Model

$$\log\left(\overset{\text{odds}}{\frac{p}{1-p}}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n$$

where p = Probability of $Y = 1$

$1-p$ = Probability of $Y = 0$

Interpretation

β_i = log OR (Odds Ratio) for having $Y = 1$ for a 1-point increase in X_i ,
adjusting for other predictors

e^{β_i} = OR for having $Y = 1$ for a 1-point increase in X_i

Binary Logistic Regression Assumptions

- Binary outcome (0, 1)
- Each predictor is linearly related to the log odds of the outcome

Inference for Regression Modeling

Confidence Interval

95% CI for $\beta = (0.44, 0.49)$

Hypothesis Testing

$H_0: \beta = \beta_0 \leftarrow$ Null Hypothesis

$H_A: \beta \neq \beta_0 \leftarrow$ Alternative Hypothesis

p-value: Chance that we are rejecting H_0 when we should not be

Goals



Today's Goal

- Learn to use R to read in data and conduct regression analysis and associated inference tests
 - Checking assumptions
 - Visualizing the data
 - Computing p-values, regression coefficients, confidence intervals, and odds ratios (for logistic models)

Today: 2 Scenarios

- Linear Regression (continuous outcome)
- Logistic Regression (categorical outcome)

(with both continuous and categorical predictors)

Today's Data Set #1: Siddarth et al., 2018

Siddarth P, Burggren AC, Eyre HA, Small GW, Merrill DA (2018) Sedentary behavior associated with reduced medial temporal lobe thickness in middle-aged and older adults. PLoS ONE 13(4): e0195549.

doi.org/10.1371/journal.pone.0195549

- Examined associations between sedentary behavior and medial temporal lobe (MTL) subregion integrity
- 35 non-demented middle-aged and older adults
- Measured physical activity levels w/questionnaire
- Measured MTL thickness w/MRI scan
- Adjusted for age

Sedentary behavior associated with reduced medial temporal lobe thickness in middle-aged and older adults

Prabha Siddarth^{1*}, Alison C. Burggren², Harris A. Eyre³, Gary W. Small¹, David A. Merrill¹

1 Semel Institute for Neuroscience and Human Behavior, UCLA, Los Angeles, CA, United States of America, 2 Center for Cognitive Neurosciences, UCLA, Los Angeles, CA, United States of America, 3 Discipline of Psychiatry, University of Adelaide, Adelaide, Australia

* psiddarth@mednet.ucla.edu

Abstract

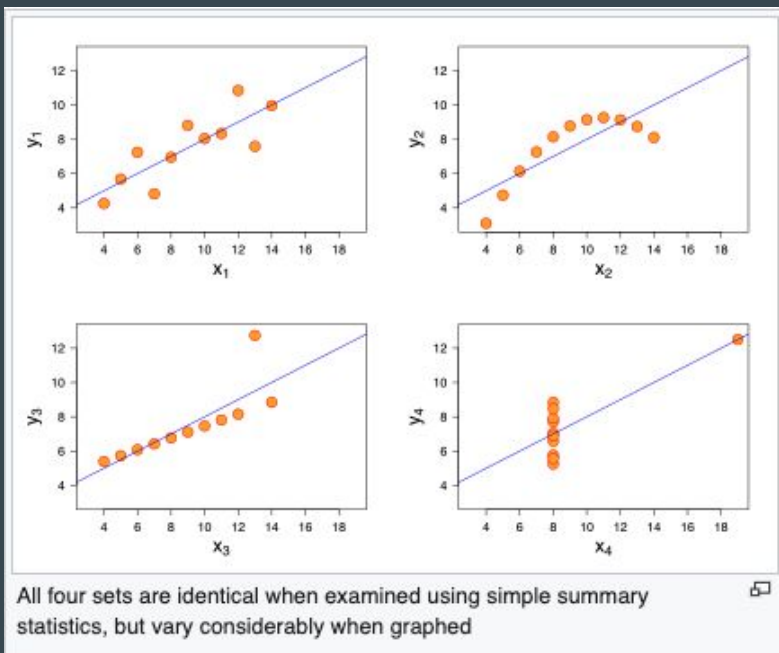
Atrophy of the medial temporal lobe (MTL) occurs with aging, resulting in impaired episodic memory. Aerobic fitness is positively correlated with total hippocampal volume, a heavily studied memory-critical region within the MTL. However, research on associations between sedentary behavior and MTL subregion integrity is limited. Here we explore associations between thickness of the MTL and its subregions (namely CA1, CA23DG, fusiform gyrus, subiculum, parahippocampal, perirhinal and entorhinal cortex), physical activity, and sedentary behavior. We assessed 35 non-demented middle-aged and older adults (25 women, 10 men; 45–75 years) using the International Physical Activity Questionnaire for older adults, which quantifies physical activity levels in MET-equivalent units and asks about the average number of hours spent sitting per day. All participants had high resolution MRI scans performed on a Siemens Allegra 3T MRI scanner, which allows for detailed investigation of the MTL. Controlling for age, total MTL thickness correlated inversely with hours of sitting/day ($r = -0.37$, $p = 0.03$). In MTL subregion analysis, parahippocampal ($r = -0.45$, $p = 0.007$), entorhinal ($r = -0.33$, $p = 0.05$) cortical and subiculum ($r = -0.36$, $p = .04$) thicknesses correlated

Today's Data Set #2: Framingham Heart Study

- framinghamheartstudy.org
- Long-term prospective study of the etiology of cardiovascular disease among a population of subjects in Framingham, MA
- Began in 1948 with 5,209 subjects
- Is the source of the term "risk factor"
- Over 3,000 peer-reviewed papers published based on this study
- Participants were each followed for a total of 24 years for cardiovascular events (heart attack, stroke, death, etc.)

First, visualize the data! Is a linear model the right one?

Anscombe's Quartet:



Some Handy R Links

Some R packages & functions for regression++

- **lme4** - Linear mixed-effects models
- **MASS** - Model selection
- **ts()**, **arima()** - Time series object, model (part of **stats**)
- **forecast** - Forecasting for time series and linear models
- **mice** - imputation of missing data
- **medflex** - mediation analysis
- **splines2** - regression spline basis functions
- **survival** - survival analysis; JM - joint modeling
- AND MANY, MANY, MANY MORE

Tutorials

- RStudio R paths: education.rstudio.com/learn/
- Data Carpentry & Software Carpentry:
 - datacarpentry.org and software-carpentry.org
- LinkedIn Learning @ GW: go.gwu.edu/linkedinlearning
- r-tutor.com/r-introduction & r-tutor.com/elementary-statistics
- UCLA Data Analysis Examples: stats.idre.ucla.edu/other/dae/
- Visualizing regression results:
worldbank.github.io/r-econ-visual-library/RegressionCoef.html
- R Graph Gallery (w/code): r-graph-gallery.com

Books you can access for free


- Free books online - Hadley Wickham:
 - R for Data Science r4ds.had.co.nz
 - Advanced R adv-r.hadley.nz/
- Through your GW library privileges:

ADVANCED SEARCH

Search for: ☐ Catalog + Articles ☒ Catalog ☐ Articles

Subject ▼ contains ▼ R (Computer programming language)

Reference Links

- R language (CRAN): r-project.org
- R package search: r-universe.dev 
- R search engine: rseek.org
- rstudio.com
 - Cheat Sheets! rstudio.com/resources/cheatsheets
- stackoverflow.com

Statistics+R help @ GW

R-Statistics Appointments: go.gwu.edu/dataconsulting

Also...

Appointments with me: calendly.com/kerchner

Coding consultations

(Python, R, git, etc.): calendly.com/gwul-coding/

Thanks!

Dan Kerchner

kerchner@gwu.edu