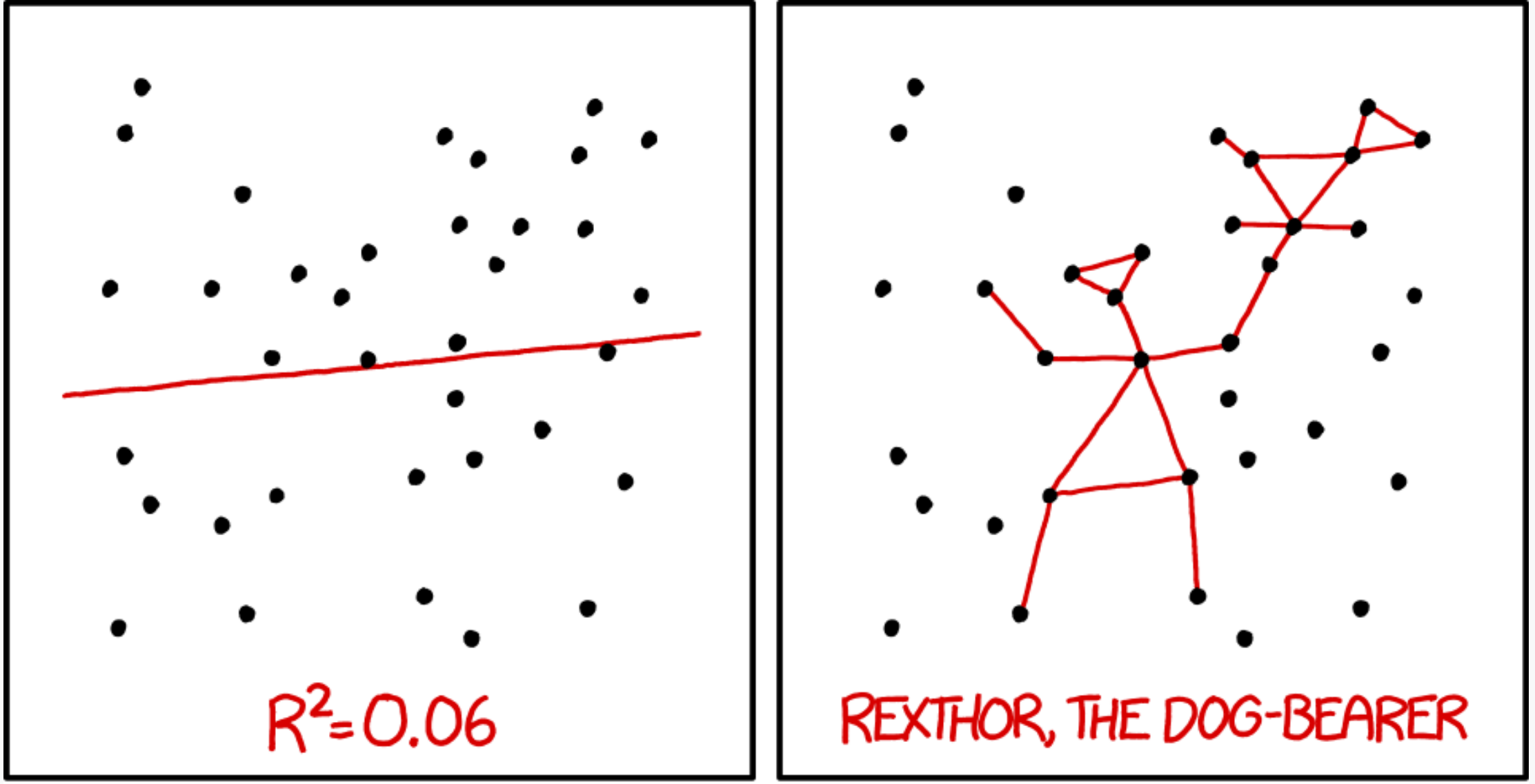
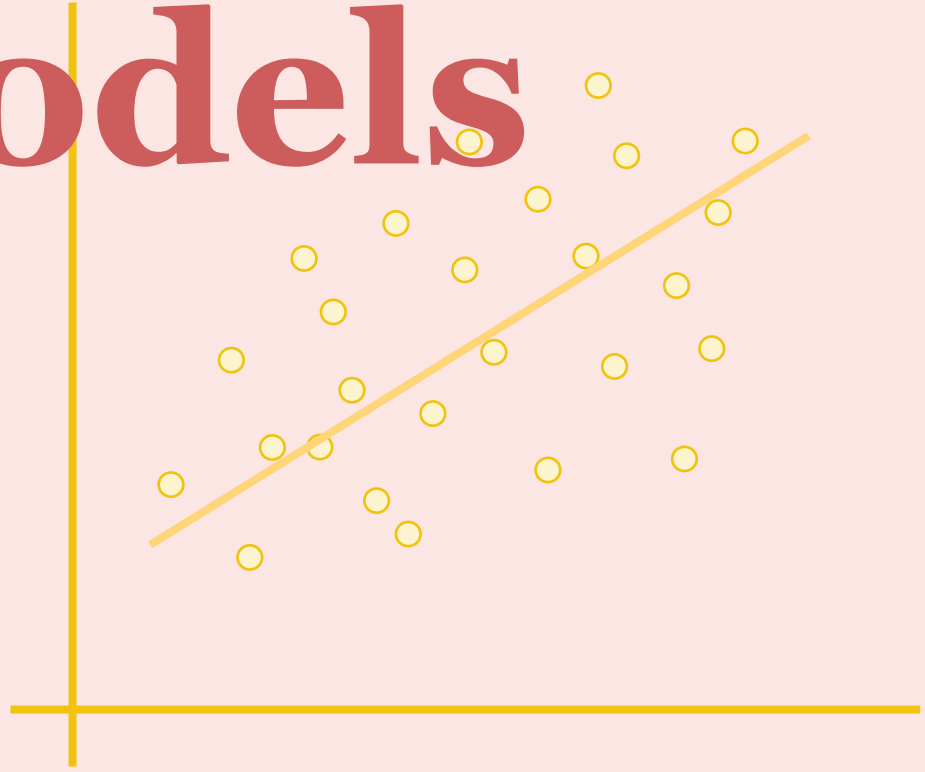


I DON'T TRUST LINEAR REGRESSIONS WHEN IT'S HARDER TO GUESS THE DIRECTION OF THE CORRELATION FROM THE SCATTER PLOT THAN TO FIND NEW CONSTELLATIONS ON IT.

EDUC 7610

Chapter 1

Statistical Control & Linear Models



$$Y_i = \beta_0 + \beta_1 X_{1i} + \epsilon_i$$

Fall 2018

Tyson S. Barrett, PhD

Linear Models

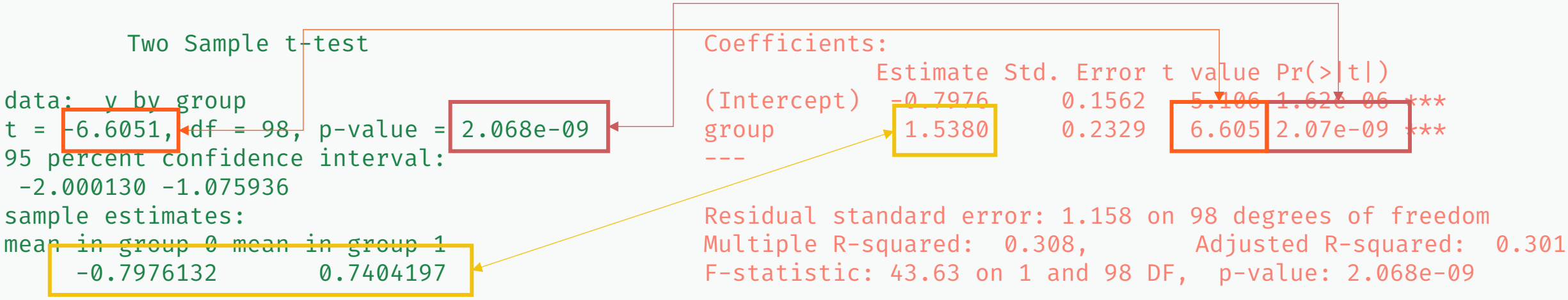
We will discuss how to use linear models to understand relationships in data and how this may generalize to the population

Everything you learned in EDUC 6600 applies here, but we will extend it (and probably simplify it somewhat)

Linear Models

Most everything you've learned are specific approaches of linear models

1. T-tests



Linear Models

Most everything you've learned are specific approaches of linear models

1. T-tests
2. ANOVA

```
            Df Sum Sq Mean Sq F value    Pr(>F)
group        1  58.55   58.55   43.63 2.07e-09
Residuals   98 131.52    1.34
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.7976     0.1562  -5.106 1.62e-06 ***
group         1.5380     0.2329   6.605 2.07e-09 ***
---

Residual standard error: 1.158 on 98 degrees of freedom
Multiple R-squared:  0.308,       Adjusted R-squared:
0.301
F-statistic: 43.63 on 1 and 98 DF,  p-value: 2.068e-09
```

Linear Models

Most everything you've learned are specific approaches of linear models

1. T-tests
2. ANOVA
3. Correlation

```
t = 9.8972, df = 98, p-value < 2.2e-16
95 percent confidence interval:
 0.5929508 0.7932781
sample estimates:
      cor
0.7070244
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.548e-17  7.108e-02   0.000        1
xs          7.070e-01  7.144e-02   9.897   <2e-16 ***
---

Residual standard error: 0.7108 on 98 degrees of
freedom
Multiple R-squared:  0.4999
F-statistic: 97.95 on 1 and 98 DF,  p-value: < 2.2e-16
```

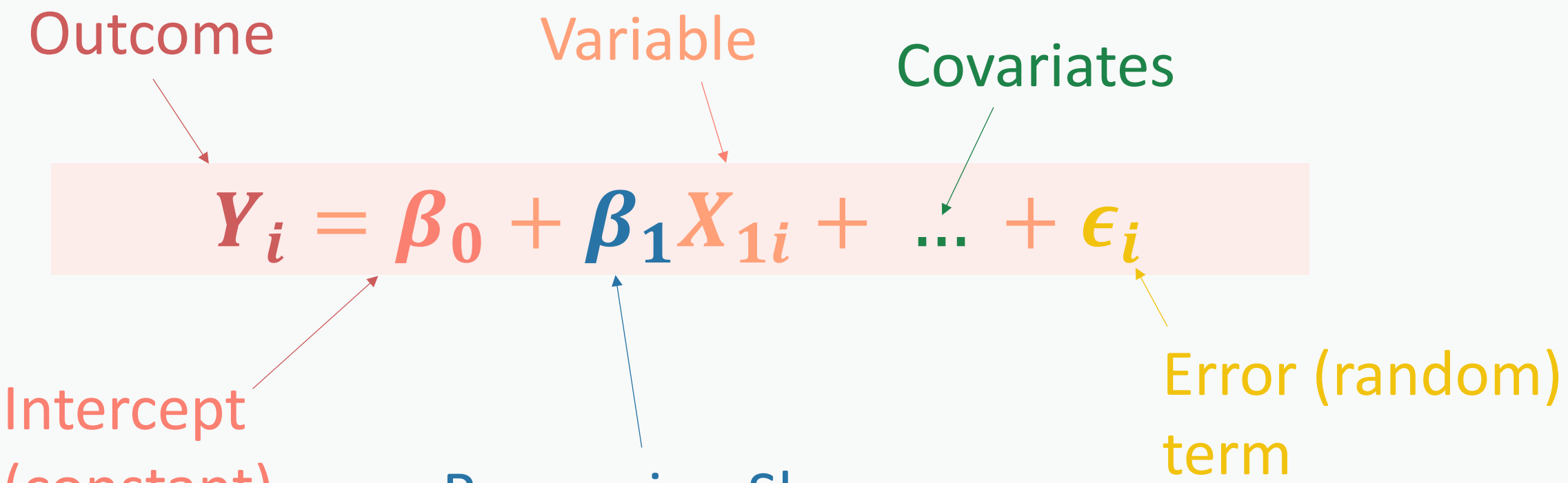
So what are linear models?

“Model” is a simplification of how something works

“Linear” says the relationships are linear

Here, we use **regression analysis (Ordinary Least Squares)** to build linear models

So what are linear models?



Requirements for linear models

1. Participants (rows in a rectangular data matrix)
2. Measures (in rectangular data matrix form)
3. Each variable a single column of numbers
4. A single dependent variable
5. Dependent variable is numeric (otherwise, GLMs)
6. Other assumptions for statistical inference

Tidy Data

Tidy Data

The idea of having rows be observations and columns are variables

The ideal for data prep for all analyses (especially in R)

Pre-print: http://vita.had.co.nz/papers/tidy-data.pdf

Journal of Statistical Software
MMMMMM YYYY, Volume VV, Issue II. *http://www.jstatsoft.org/*

Tidy Data

Hadley Wickham
RStudio

Abstract

A huge amount of effort is spent cleaning data to get it ready for analysis, but there has been little research on how to make data cleaning as easy and effective as possible. This paper tackles a small, but important, component of data cleaning: data tidying. Tidy datasets are easy to manipulate, model and visualise, and have a specific structure: each variable is a column, each observation is a row, and each type of observational unit is a table. This framework makes it easy to tidy messy datasets because only a small set of tools are needed to deal with a wide range of un-tidy datasets. This structure also makes it easier to develop tidy tools for data analysis, tools that both input and output tidy datasets. The advantages of a consistent data structure and matching tools are demonstrated with a case study free from mundane data manipulation chores.

Keywords: data cleaning, data tidying, relational databases, R.

1. Introduction

It is often said that 80% of data analysis is spent on the process of cleaning and preparing the data (Dasu and Johnson 2003). Data preparation is not just a first step, but must be repeated many over the course of analysis as new problems come to light or new data is collected. Despite the amount of time it takes, there has been surprisingly little research on how to clean data well. Part of the challenge is the breadth of activities it encompasses: from outlier checking, to date parsing, to missing value imputation. To get a handle on the problem, this paper focusses on a small, but important, aspect of data cleaning that I call data **tidying**: structuring datasets to facilitate analysis.

The principles of tidy data provide a standard way to organise data values within a dataset. A standard makes initial data cleaning easier because you don't need to start from scratch and reinvent the wheel every time. The tidy data standard has been designed to facilitate initial exploration and analysis of the data, and to simplify the development of data analysis tools that work well together. Current tools often require translation. You have to spend time

Tidy Data

Some pragmatic guidelines to have tidy data

1 Be Consistent
2 Choose good names for things
3 Write dates as YYYY-MM-DD
4 Put just one thing in a cell
5 Make it a rectangle
6 Create a data dictionary

Data organization in spreadsheets

Karl W. Broman *
Department of Biostatistics & Medical Informatics, University of Wisconsin-Madison
and
Kara H. Woo
Information School, University of Washington

August 19, 2017

Abstract

Spreadsheets are widely used software tools for data entry, storage, analysis, and visualization. Focusing on the data entry and storage aspects, this paper offers practical recommendations for organizing spreadsheet data to reduce errors and ease later analyses. The basic principles are: be consistent, write dates like YYYY-MM-DD, don't leave any cells empty, put just one thing in a cell, organize the data as a single rectangle (with subjects as rows and variables as columns, and with a single header row), create a data dictionary, don't include calculations in the raw data files, don't use font color or highlighting as data, choose good names for things, make backups, use data validation to avoid data entry errors, and save the data in plain text file.

Keywords: data management, data organization, spreadsheets, Microsoft Excel

*The authors thank Lance Waller, Lincoln Mullen, and Jenny Bryan for their comments to improve the manuscript.

Linear models are flexible

1. Predictor can be experimentally manipulated or naturally occurring
2. A dependent variable in one model can be a predictor in another
3. Predictors may be dichotomous, multi-categorical or continuous
4. Predictors can be correlated
5. Predictors may interact
6. Can be extended to handle curvilinear relations
7. Assumptions are not all that limiting

Some Vocabulary

1. Predictor = the variable that predicts the outcome
2. Independent Variable (IV) = the predictor(s) of interest
3. Covariates = the predictors that are not considered IV
4. Control = a class of methods that remove (or control for) the effect of another variable

Controlling for Confounders

Five Types of Control

1. Random assignment on the independent variable
2. Manipulation of covariates
3. Other types of randomization
4. Exclusion of cases
5. **Statistical control**

Experimental Control

Statistical Control

i.e., “Adjust for”, “Correct for”, “Hold constant”, “Partial out”

Linear models allow us to control for the effects of covariates

We’ll discuss more about how this works, but essentially makes everyone statistically equal on each covariate

Can assess how much of the differences in the outcome are uniquely attributable to the IV (once we take out the effect of the covariates)