# Data Analytics for Psychology Business

## Session 2

Dirk Wulff & Rui Mata, FS 2021

The **sleepstudy** data set available in the lme4 package and stems from an experiment that examined reaction times of sleep deprived individuals over the course of a few days (individuals only got ca. 3h of sleep each day)…

CHALLENGE: What are the main insights one can get from those data? How would you visualize/analyse the data to get some insights from these data?

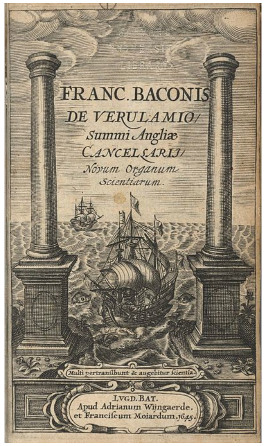
```
library(lme4)
```

## Goals

- Understand the nature of causal inference as the comparison of treatment to some counterfactual
- List different methods of causal inference (e.g., randomization/experiments, regression, regression discontinuity) and associated limitations
- Further familiarize us with regression/mixed-effects modeling

# Evidence-based decision making

Francis Bacon
(1561-1626)




1620


Causal inference in economics and marketing

Hal R. Varian

Varian, H. R. (2016). Causal inference in economics and marketing. Proceedings of the National Academy of Sciences of the United States of America, 113(27), 7310–7315. http://doi.org/10.1073/pnas.1510479113

Bacon suggests that one can draw up a list of all things in which the phenomenon to explain occurs, as well as a list of things in which it does not occur. Then one can rank the lists according to the degree in which the phenomenon occurs in each one. Then one should be able to deduce what factors match the occurrence of the phenomenon in one list and do not occur in the other list, and also what factors change in accordance with the way the data had been ranked.

"The critical step in any causal analysis is estimating the counterfactual—a prediction of what would have happened in the absence of the treatment"

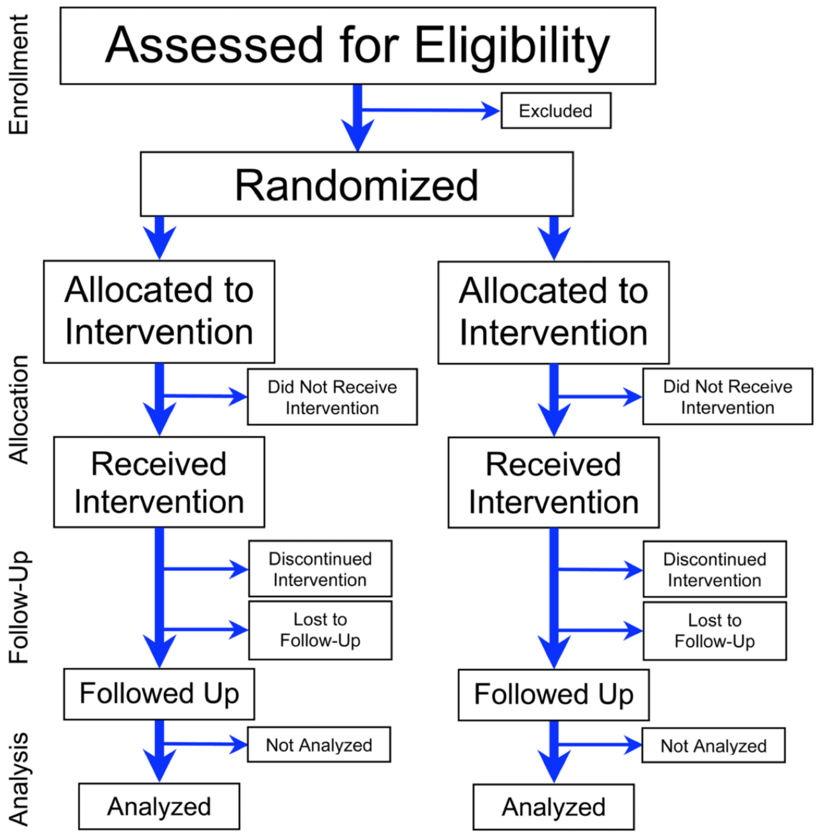
# The gold standard…

## Experiments/Randomised control trials (RCT)

A type of scientific experiment, where the people being studied are randomly allocated one or other of the different treatments under study. RCTs are considered the gold standard for a clinical trial. RCTs are often used to test the efficacy or effectiveness of various types of medical intervention and may provide information about adverse effects, such as drug reactions. Random assignment of intervention is done after subjects have been assessed for eligibility and recruited, but before the intervention to be studied begins.

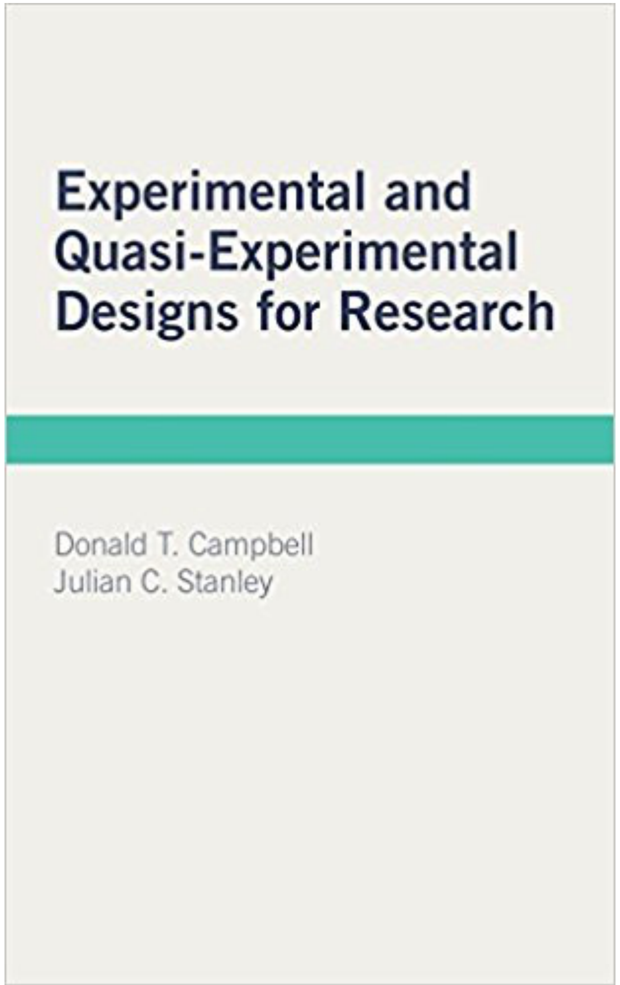
$$Y = B_0 + B_1 group$$

**Problem**: Experiments are not always feasible or ethical; samples may not be representative of the population leading to generalization failures

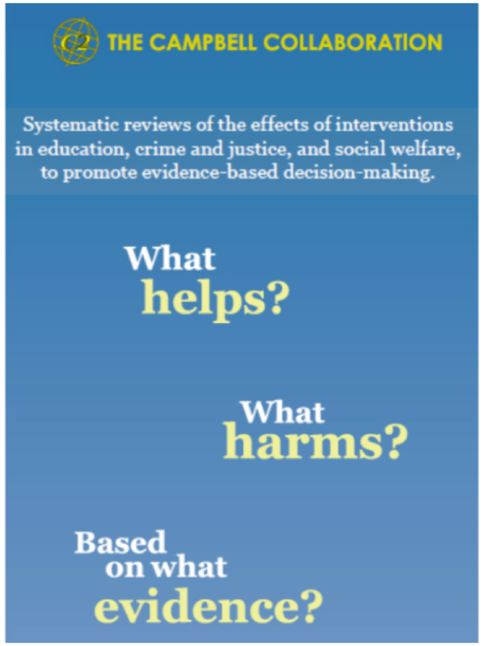



Shorter, E. (2011). A brief history of placebos and clinical trials in psychiatry. *Canadian Journal of Psychiatry, 56*(4), 193–197.

# But there are alternatives…




1963

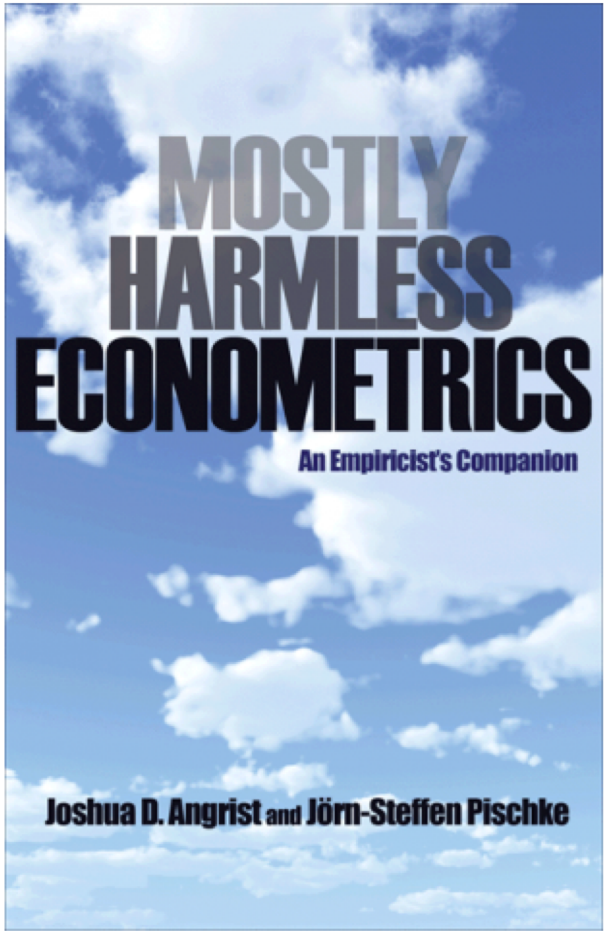



Donald Campbell
1916-1996

THE CAMPBELL COLLABORATION

“Through out, attention has been called to the possibility of **creatively** utilizing the idiosyncratic features of any specific research situation in designing unique tests of causal hypotheses.”

# “Furious Five” statistical methods for causal inference




- Randomisation
- Regression
- Instrumental variables
- Difference in differences
- Regression discontinuity

Angrist, J. D., & Pischke, J.-S. (2010). The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con out of Econometrics. *Journal of Economic Perspectives, 24*(2), 3–30. http://doi.org/10.1257/jep.24.2.3

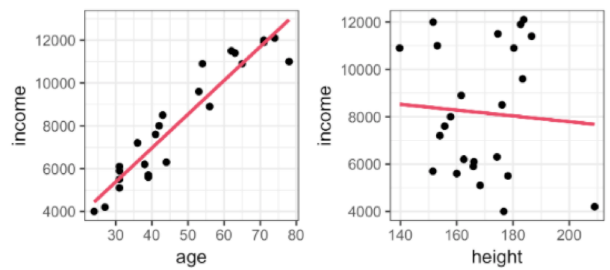
# Regression

Regression analysis is a set of statistical processes for estimating the relationships among variables. It includes many techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable (criterion) and one or more independent variables (predictors). More specifically, regression analysis helps one understand how the typical value of the dependent variable changes when any one of the independent variables is varied, while the other independent variables are fixed.

## Multiple Linear Regression

**Definition**: Multiple linear regression is a linear model with many predictors $x_1, x_2, \dots x_n$, and where the error term $\epsilon$ is Normally distributed.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n + \epsilon$$




| Parameter | Description | In words |
|---|---|---|
| $\beta_0$ | Intercept | When all x values are 0, what is the predicted value for y? |
| $\beta_1, \beta_2, \dots$ | Coefficient for $x_1, x_2, \dots$ | For every increase of 1 in coefficient for $x_1, x_2, \dots$ how does y change? |

Formula

$$income = 1628 + 147 \times age - 4.1 \times height + \epsilon$$

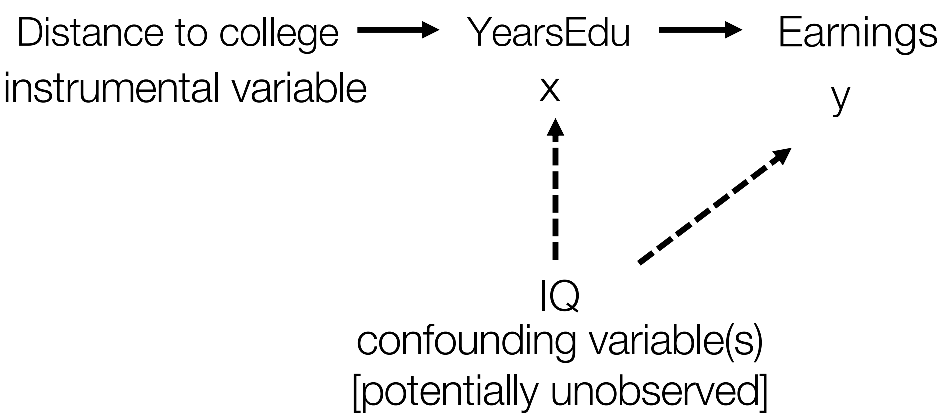
Coefficients

$$\beta_0 = 1628, \beta_{age} = 147, \beta_{weight} = -4.1$$

**Problem**: Correlation is not causation!

# Instrumental variables

The method of instrumental variables (IV) is used to estimate causal relationships when controlled experiments are not feasible or when a treatment is not successfully delivered to every unit in a randomized experiment. Intuitively, the method is used when an explanatory variable of interest is correlated with the error term, in which case ordinary least squares gives biased results. A valid instrument (instrumental variable) induces changes in the explanatory variable (x) but has no independent effect on the dependent variable (y), allowing a researcher to uncover the causal effect of the explanatory variable on the dependent variable.




Estimation though two-stage least squares.

Stage 1: generate predictions of YearsEdu:

YearsEdu_pred = $B_0$ + $B_1$ DisttoCollege + $Error_1$

Stage 2: test whether YearsEdu_pred is significantly associated with earnings:

Earnings = $B_2$ + $B_3$ YearsEdu_pred + $Error_2$

**Problem**: Good instrumental variables (i.e., that are correlated with x but not any confounding variables) are hard to find…

Angrist, J. D., & Krueger, A. B. (2001). Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments. *Journal of Economic Perspectives, 15*(4), 69–85.

# Instrumental variables

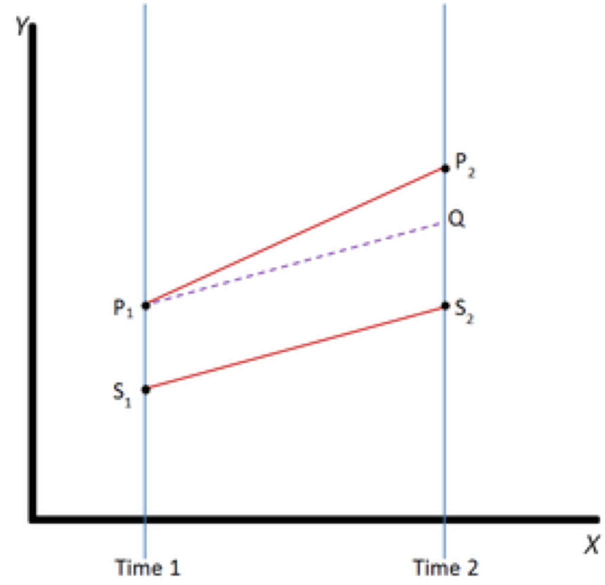
*Table 1*
**Examples of Studies That Use Instrumental Variables to Analyze Data From Natural and Randomized Experiments**

| *Outcome Variable* | *Endogenous Variable* | *Source of Instrumental Variable(s)* | *Reference* |
|---|---|---|---|
| | | *1. Natural Experiments* | |
| Labor supply | Disability insurance replacement rates | Region and time variation in benefit rules | Gruber (2000) |
| Labor supply | Fertility | Sibling-Sex composition | Angrist and Evans (1998) |
| Education, Labor supply | Out-of-wedlock fertility | Occurrence of twin births | Bronars and Grogger (1994) |
| Wages | Unemployment insurance tax rate | State laws | Anderson and Meyer (2000) |
| Earnings | Years of schooling | Region and time variation in school construction | Duflo (2001) |
| Earnings | Years of schooling | Proximity to college | Card (1995) |
| Earnings | Years of schooling | Quarter of birth | Angrist and Krueger (1991) |
| Earnings | Veteran status | Cohort dummies | Imbens and van der Klaauw (1995) |
| Earnings | Veteran status | Draft lottery number | Angrist (1990) |
| Achievement test scores | Class size | Discontinuities in class size due to maximum class-size rule | Angrist and Lavy (1999) |
| College enrollment | Financial aid | Discontinuities in financial aid formula | van der Klaauw (1996) |
| Health | Heart attack surgery | Proximity to cardiac care centers | McClellan, McNeil and Newhouse (1994) |
| Crime | Police | Electoral cycles | Levitt (1997) |
| Employment and Earnings | Length of prison sentence | Randomly assigned federal judges | Kling (1999) |
| Birth weight | Maternal smoking | State cigarette taxes | Evans and Ringel (1999) |

Angrist, J. D., & Krueger, A. B. (2001). Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments. *Journal of Economic Perspectives, 15*(4), 69–85.

# Difference in differences

Difference in differences (DID or DD) is a statistical technique used in the social sciences that attempts to mimic an experimental research design using observational study data, by studying the differential effect of a treatment on a 'treatment group' versus a 'control group' in a natural experiment. It calculates the effect of a treatment on an outcome by comparing the average change over time in the outcome variable for the treatment group, compared to the average change over time for the control group. Although it is intended to mitigate the effects of extraneous factors and selection bias, depending on how the treatment group is chosen, this method may still be subject to certain biases (e.g., omitted variable bias).



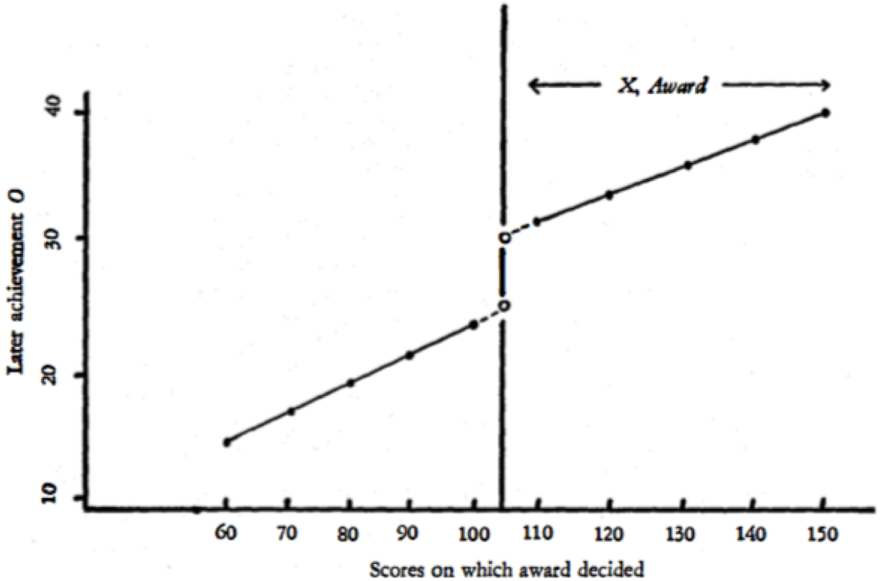

$$Y = B_0 + B_1 Group + B_2 Time + B_3\, Group*Time$$

**Problem**: Assumption that the change in outcomes from pre- to post-intervention in the control group (S) is a good proxy for the (counterfactual) change in untreated potential outcomes in the treated group (P) may not be warranted; choice of treatment/control groups is crucial (an additional trick may be *matching* on observables)…


Bertrand, M., Duflo, E., & Mullainathan, S. (2004). How Much Should We Trust Differences-in-Differences Estimates? The Quarterly Journal of Economics, 119(1), 249–275.

# Regression discontinuity

A regression discontinuity design (RDD) is a quasi-experimental pretest-posttest design that elicits the causal effects of interventions by assigning a cutoff or threshold above or below which an intervention is assigned. By comparing observations lying closely on either side of the threshold, it is possible to estimate the average treatment effect in environments in which randomization is unfeasible. RDD was first applied by Donald Thistlethwaite and Donald Campbell to the evaluation of scholarship programs.



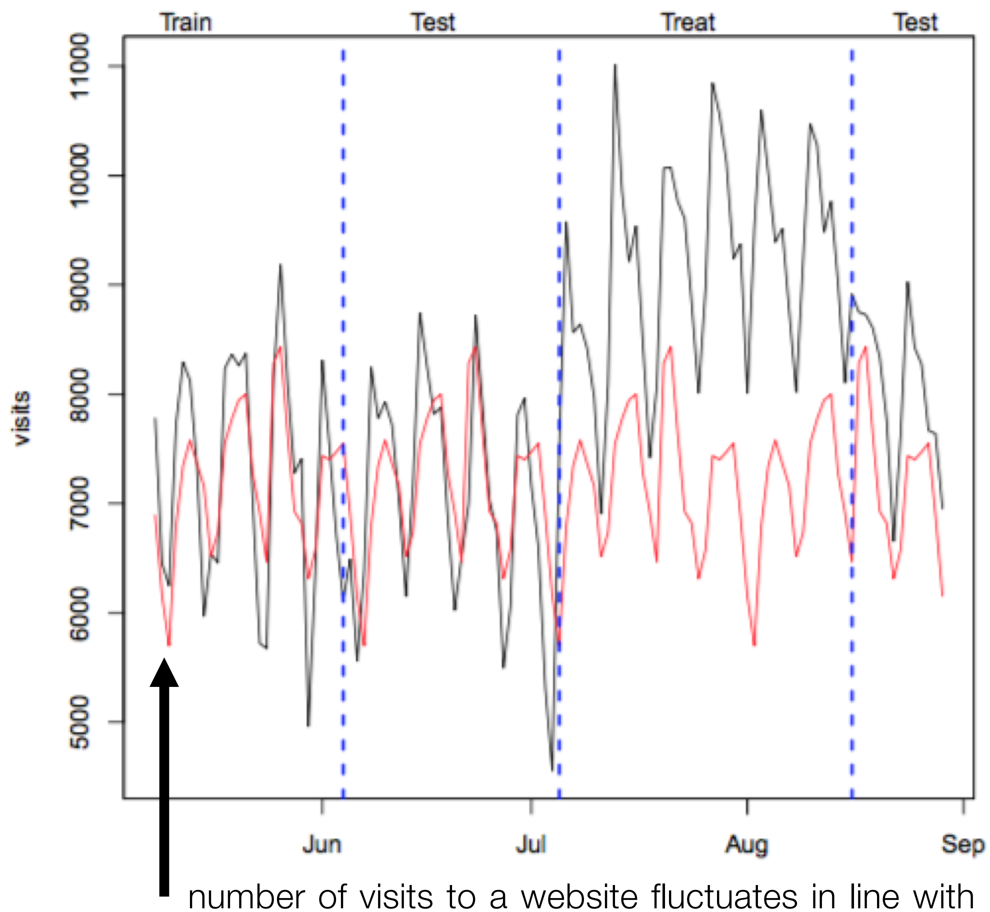

$Y = B_0 + B_1 Score + B_2 Award$

**Problem**: Assumption that the individuals just below the cutoff are not systematically different from those just above can be wrong (e.g., individuals just above the threshold could try harder); the estimation may not generalise to observations away from the cutoff (e.g., awards could have different results at different levels of ability).

Lee, D. S., & Lemieux, T. (2010). Regression Discontinuity Designs in Economics. *Journal of Economic Literature, 48*(2), 281–355. http://doi.org/10.1257/jel.48.2.281

# New developments using machine learning…

## Using models as the control group (Train-test-treat-compare)




number of visits to a website fluctuates in line with number of searches
(the model - red line - is trained to capture these fluctuations)

An online advertiser might ask "if I increase my ad expenditure by some amount, how many extra sales do I generate?"

A predictive statistical model (based on number of "searches" about topics related to the subject matter of the website) is estimated during the training period and its predictive performance is assessed during the test period. The extrapolation of the model during the treat period (red line) serves as a counterfactual. This counterfactual is compared with the actual outcome (black line), and the difference is the estimated treatment effect. When the treatment is ended, the outcome returns to something close to the original level.

Varian, H. R. (2016). Causal inference in economics and marketing. *Proceedings of the National Academy of Sciences of the United States of America, 113*(27), 7310–7315. http://doi.org/10.1073/pnas.1510479113

# Summary

"The critical step in any causal analysis is estimating the counterfactual—a prediction of what would have happened in the absence of the treatment"

There are many types of causal inference analyses that can be (and are) used in the behavioural sciences - in psychology, experiments and multiple regression from observational data are the most commonly used inference methods.

It is helpful to be aware of other methods (e.g., instrumental variables, regression discontinuity, difference in differences) and, more importantly, "the possibility of **creatively** utilizing the idiosyncratic features of any research situation in designing tests of causal hypotheses".

# Regression: Mixed-effects approach

A mixed model is a statistical model containing both **fixed** effects (same across groups) and **random** effects (varying across groups).

These models are useful in a wide variety of disciplines in the physical, biological and social sciences. They are particularly useful in settings where repeated measurements are made on the same statistical units (longitudinal study), or where measurements are made on clusters of related statistical units. Because of their advantage in dealing with missing values, mixed effects models are often preferred over more traditional approaches.

# Reasons for using mixed-effects: Nested structures

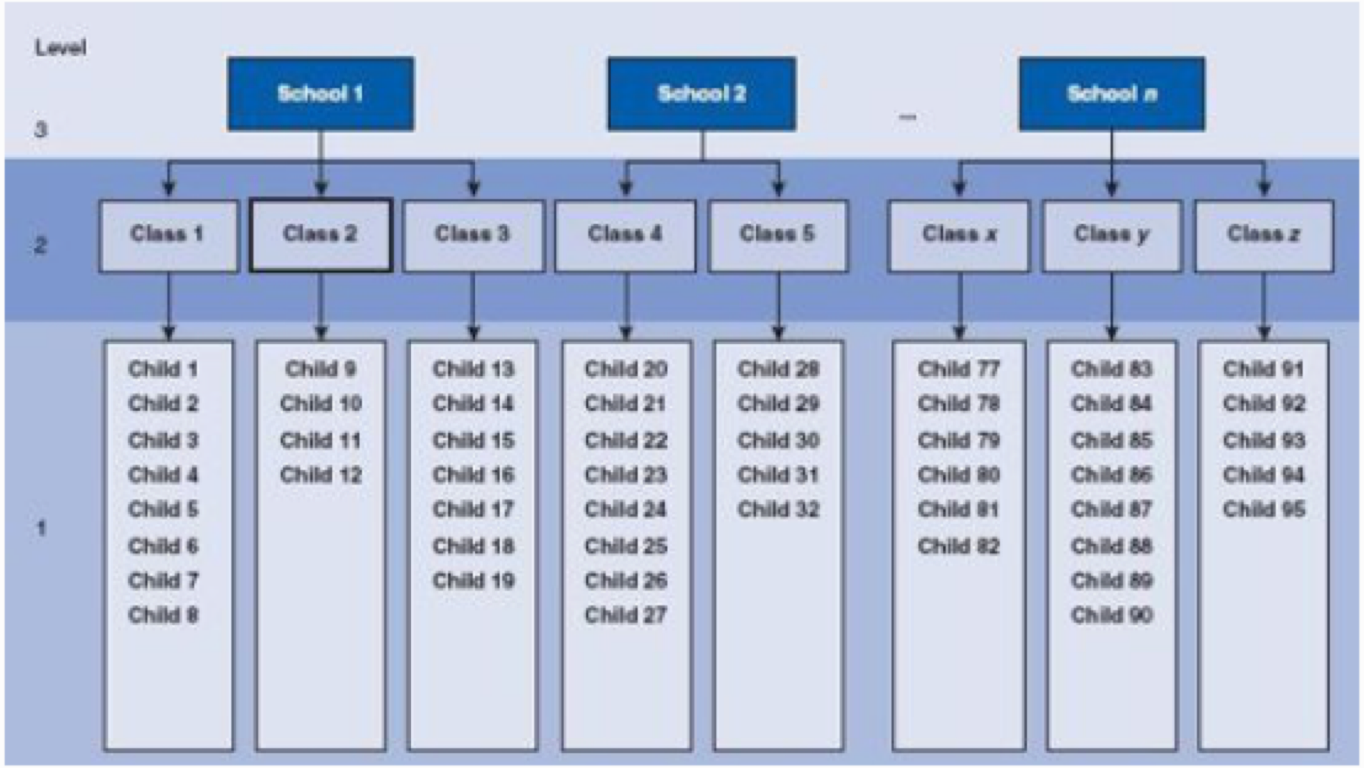
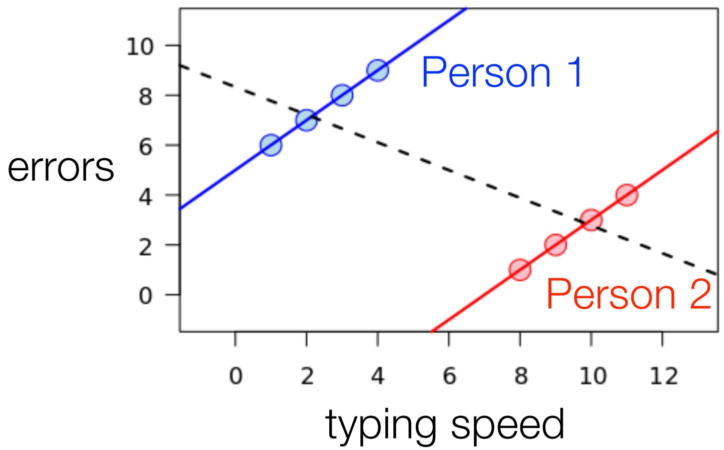



**FIGURE 19.3** An example of a three-level hierarchical data structure

Field, A., Miles, J., & Field, Z (2012). *Discovering statistics using R*. SAGE Publications.

# Reasons for using mixed-effects: Simpsons' paradox

Paradox in which a trend that appears in groups of data disappears when these groups are combined and the reverse trend appears for the aggregate data.




**speed-accuracy trade-off**

people who are faster are also more accurate (between)
when people are faster they become less accurate (within)

**exercise and fatigue**

people who exercise more are less fatigued (between)
when people exercise more they are more fatigued (within)

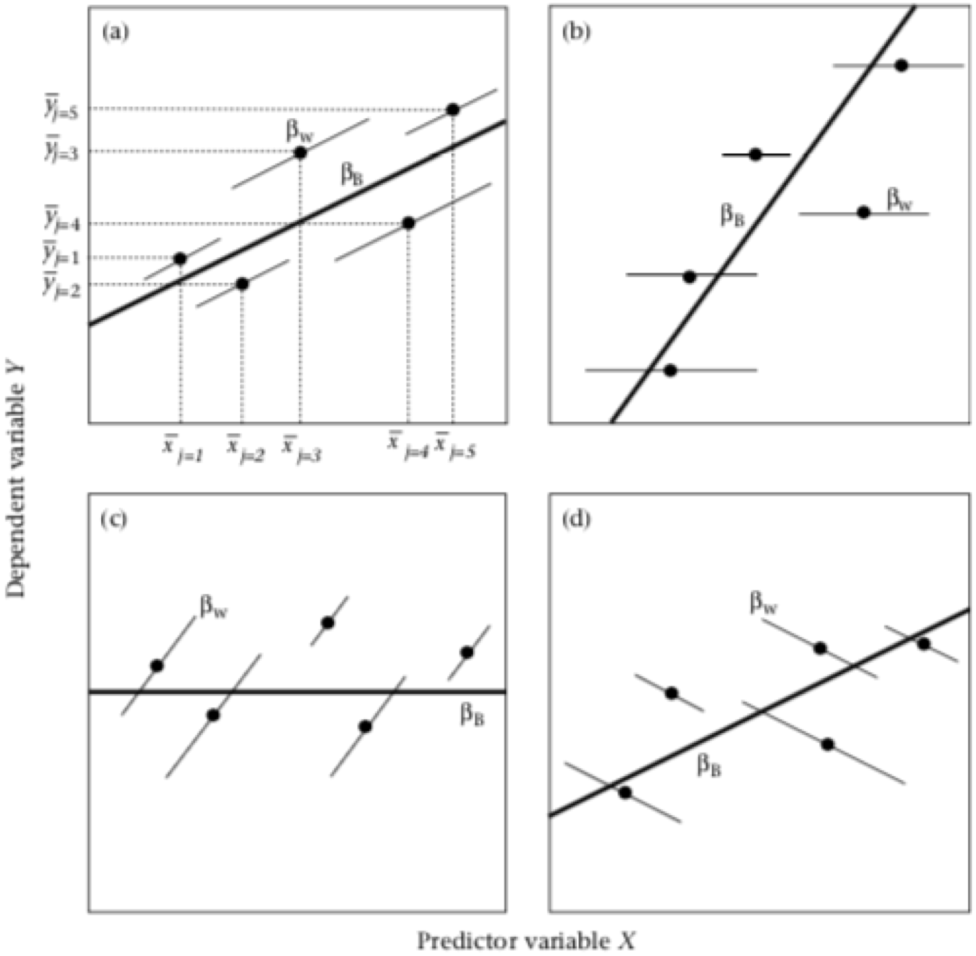
# Fitting Linear Mixed-Effects Models Using **lme4**

model <- lmer(Y ~ x + (x | g), data)

<table>
<tr><th>Formula</th><th>Alternative</th><th>Meaning</th></tr>
<tr><td><code>(1 | g)</code></td><td><code>1 + (1 | g)</code></td><td>Random intercept with fixed mean.</td></tr>
<tr><td><code>0 + offset(o) + (1 | g)</code></td><td><code>-1 + offset(o) + (1 | g)</code></td><td>Random intercept with a priori means.</td></tr>
<tr><td><code>(1 | g1/g2)</code></td><td><code>(1 | g1)+(1 | g1:g2)</code></td><td>Intercept varying among <code>g1</code> and <code>g2</code> within <code>g1</code>.</td></tr>
<tr><td><code>(1 | g1) + (1 | g2)</code></td><td><code>1 + (1 | g1) + (1 | g2).</code></td><td>Intercept varying among <code>g1</code> and <code>g2</code>.</td></tr>
<tr><td><code>x + (x | g)</code></td><td><code>1 + x + (1 + x | g)</code></td><td>Correlated random intercept and slope.</td></tr>
<tr><td><code>x + (x || g)</code></td><td><code>1 + x + (1 | g) + (0 + x | g)</code></td><td>Uncorrelated random intercept and slope.</td></tr>
</table>

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1–51. http://doi.org/10.18637/jss.v067.i01

# Reasons for using mixed-effects: Simpsons' paradox

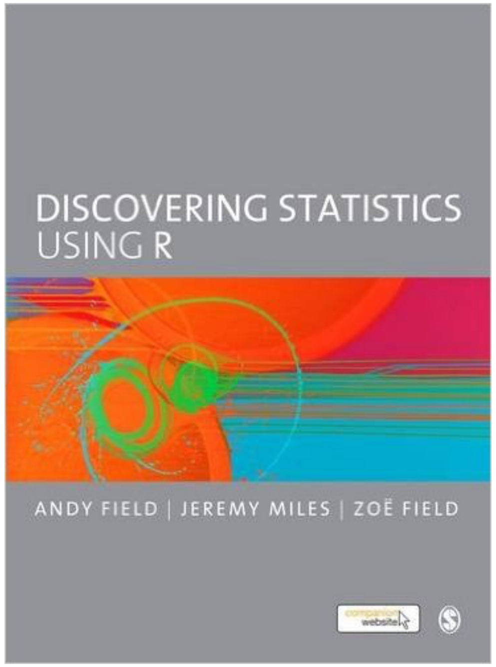


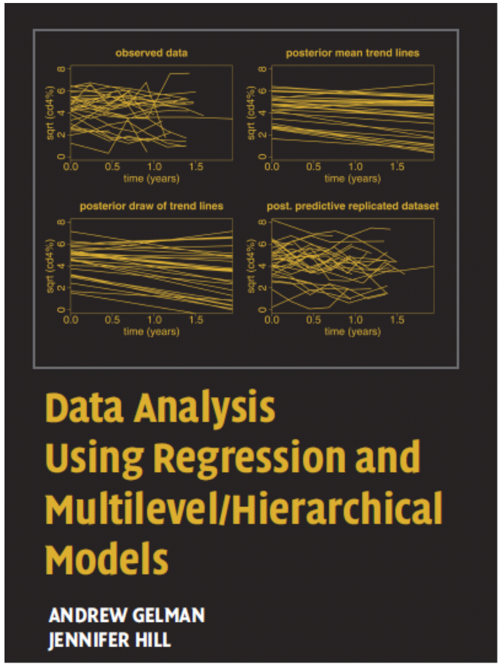

"The procedure for distinguishing within- versus between- subject effects is technically simple and has been around for some time (Davis et al. 1961; see also Raudenbush 1989; Kreft et al. 1995; Snijders & Bosker 1999). The technique is usually called 'within- group centering'"

**Figure 1.** Four different scenarios for how within- and between-subject effects can differ within a data set. We schematically depict the within-subject slopes (thin solid lines; $\beta_W$) of five subjects ($j = 1$ to $j = 5$) with the corresponding between-subject slope (thick solid lines; $\beta_B$) resulting from the association between $\bar{x}_j$ and $\bar{y}_j$ (●).

van de Pol, M., & Wright, J. (2009). A simple method for distinguishing within- versus between-subject effects using mixed models. *Animal Behaviour, 77*(3), 753–758. http://doi.org/10.1016/j.anbehav.2008.11.006

# further reading




Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*(4), 390–412. http://doi.org/10.1016/j.jml.2007.12.005

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology, 103*(1), 54–69. http://doi.org/10.1037/a0028347