9.S918: Quantitative Inference in Brain and Cognitive Sciences

Last class!

- **Instrumental Variables**
- **Semester recap**
- **Modeling do's & don'ts**

Roger Levy
Dept. of Brain & Cognitive Sciences
Massachusetts Institute of Technology

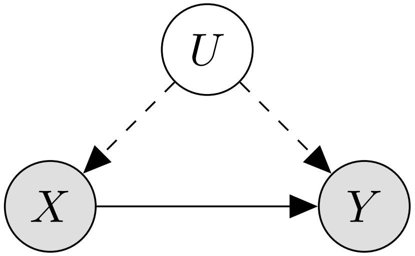
May 12, 2025

Instrumental variables (IVs): motivation

- Simple question of great scientific & policy interest:

  What is the causal effect of smoking on physical health?

- Not feasible to estimate with a randomized experiment:
  - Any such effects clearly accrue over a long time period
  - Assigning people to a "smoking" condition may be unethical
  - Even with randomized assignment: **compliance** issues
- We have **observational** data, but there are **confounds**
  - Socioeconomic status, education, healthcare access
  - Mental health
  - Genetic predispositions
  - ...



- Confound both **presence** and **magnitude** of effect
- How can we estimate the causal effect scientifically?

Some econometric concepts

- Consider the linear model:

$$Y = \alpha + \beta X + \gamma U + \xi$$

  where ξ is the only stochastic component and $\xi \perp U \mid \{\}$

- If U has a causal effect on X, but U is unmeasured and we omit it from the regression, instead fitting:

$$Y = \alpha + \beta X + \epsilon$$

- the $U \rightarrow X$ effect creates a **correlation between X and ϵ that we have no way of identifying from the data**
- This means that β is **unidentifiable** without further assumptions and/or types of data
- This is a case of **endogeneity**, defined more generally as **correlation between an explanatory variable and the error term in a model fitted to data**

Instrumental variable for smoking→health

$$Y = \alpha + \beta X + \gamma U + \xi$$

$$Y = \alpha + \beta X + \epsilon$$

- So, are we sunk? Not necessarily!
- Suppose we had a measured source of variation in X that we knew was **uncorrelated with the residual error** ϵ
- We could use that variation to estimate the causal effect of X on Y!
- Conceptual example for smoking & health:
  - Different U.S. states vary in cigarette tax and thus prices
  - This variation is arguably minimally correlated with U
  - Instead of looking at the relationship between actual smoking and health, we'll look at the relationship between **smoking as predicted by cigarette prices and health!**

2-stage causal effect estimation with IVs

$$Y = \gamma_0 + \gamma_X X + \gamma_W W + \gamma_U U + \xi$$

- X: amount of smoking
- Y: health outcomes
- Z: cigarette prices, the **instrumental variable (IV)**
- W: other relevant covariates that might affect X and/or Y
- Step 1: predict smoking from cigarette prices

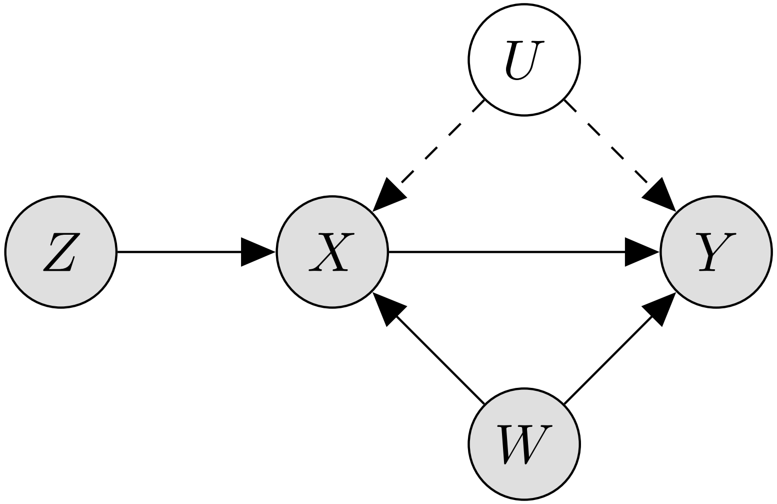
$$X = \pi_0 + \pi_Z Z + \pi_W W + \tau$$

- Step 2: predict health from **predicted smoking**

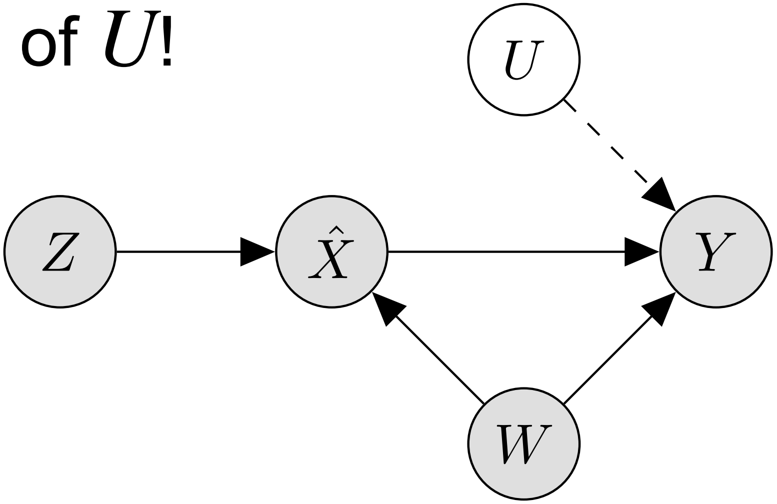
$$Y = \beta_0 + \beta_{\hat{X}} \hat{X} + \beta_W W + \nu$$

- Notice that the error term ν rolls together contributions of $(X - \hat{X})$, U, and ξ, and is **uncorrelated with** $\hat{X}$**!**

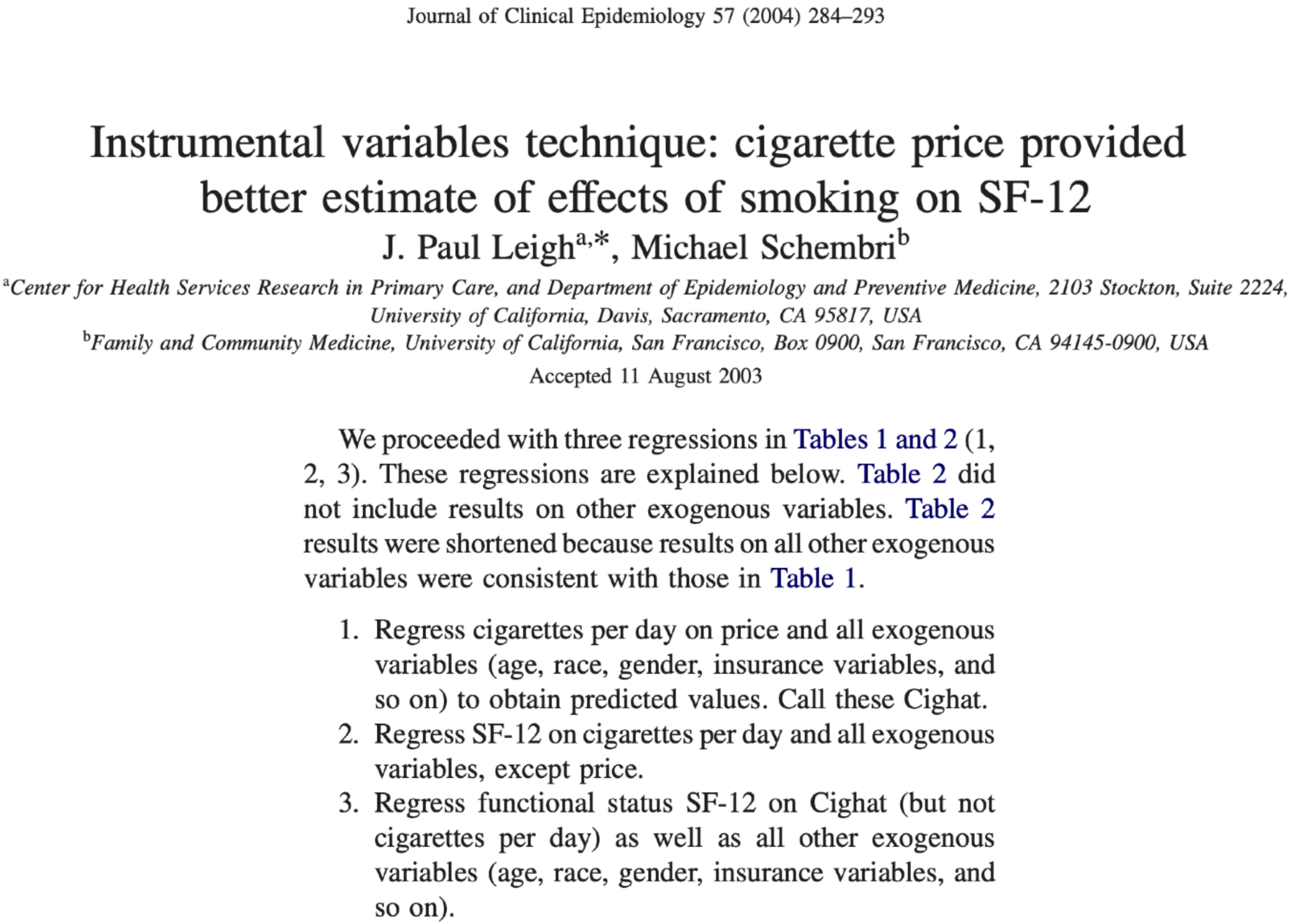
A graphical view of instrumental variables



- By replacing X with **only** the part you can predict from Z and W, you have a new variable that is unconditionally independent of U!



Practical ex. of instrumental variable



Journal of Clinical Epidemiology 57 (2004) 284–293

Instrumental variables technique: cigarette price provided better estimate of effects of smoking on SF-12

J. Paul Leigh[a,*], Michael Schembri[b]

[a]*Center for Health Services Research in Primary Care, and Department of Epidemiology and Preventive Medicine, 2103 Stockton, Suite 2224, University of California, Davis, Sacramento, CA 95817, USA*

[b]*Family and Community Medicine, University of California, San Francisco, Box 0900, San Francisco, CA 94145-0900, USA*

Accepted 11 August 2003

We proceeded with three regressions in Tables 1 and 2 (1, 2, 3). These regressions are explained below. Table 2 did not include results on other exogenous variables. Table 2 results were shortened because results on all other exogenous variables were consistent with those in Table 1.

1. Regress cigarettes per day on price and all exogenous variables (age, race, gender, insurance variables, and so on) to obtain predicted values. Call these Cighat.
2. Regress SF-12 on cigarettes per day and all exogenous variables, except price.
3. Regress functional status SF-12 on Cighat (but not cigarettes per day) as well as all other exogenous variables (age, race, gender, insurance variables, and so on).

Practical ex. of instrumental variable

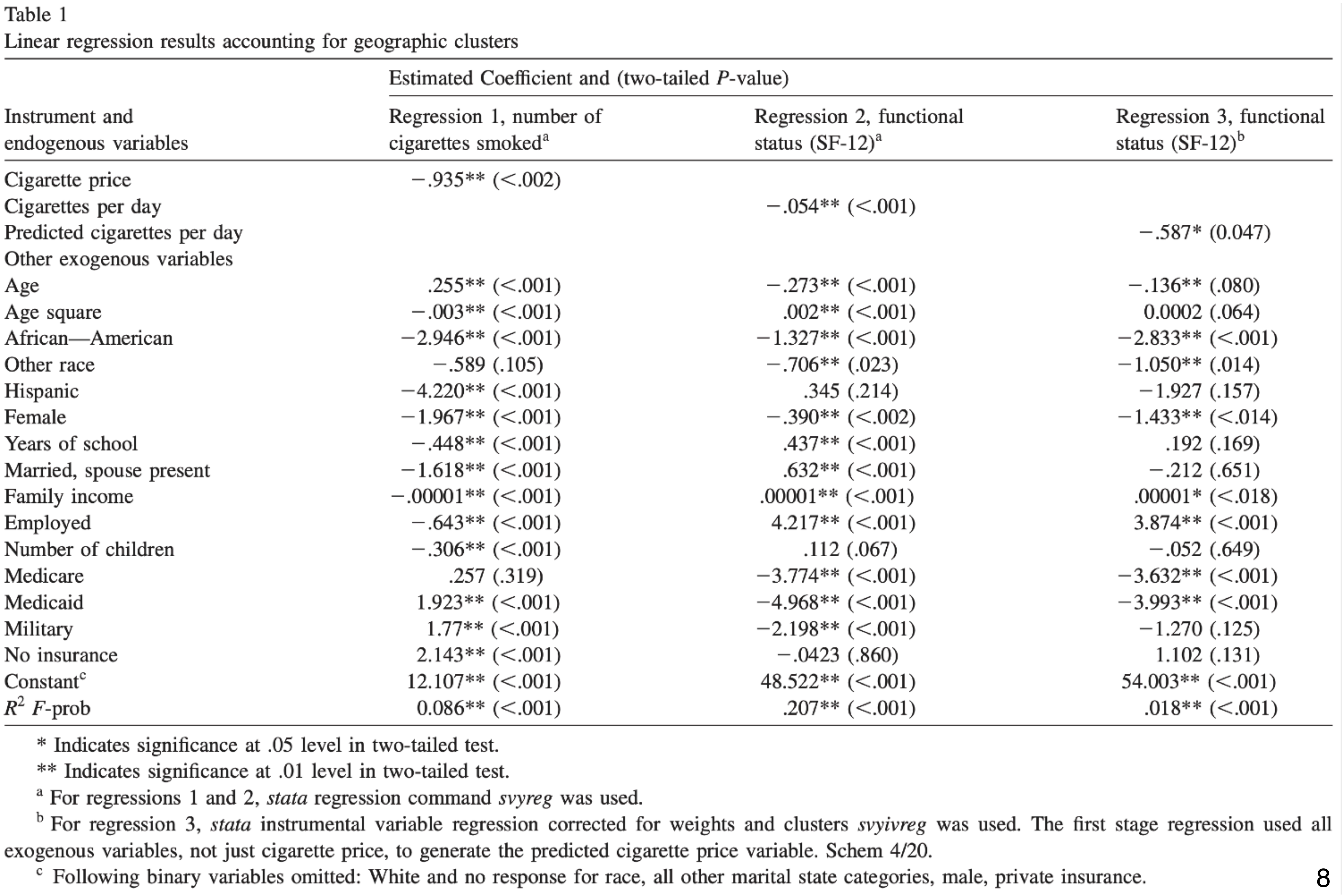


Table 1

Linear regression results accounting for geographic clusters

	Estimated Coefficient and (two-tailed *P*-value)		
Instrument and endogenous variables	Regression 1, number of cigarettes smoked[a]	Regression 2, functional status (SF-12)[a]	Regression 3, functional status (SF-12)[b]
Cigarette price	−.935** (<.002)		
Cigarettes per day		−.054** (<.001)	
Predicted cigarettes per day			−.587* (0.047)
Other exogenous variables			
Age	.255** (<.001)	−.273** (<.001)	−.136** (.080)
Age square	−.003** (<.001)	.002** (<.001)	0.0002 (.064)
African—American	−2.946** (<.001)	−1.327** (<.001)	−2.833** (<.001)
Other race	−.589 (.105)	−.706** (.023)	−1.050** (.014)
Hispanic	−4.220** (<.001)	.345 (.214)	−1.927 (.157)
Female	−1.967** (<.001)	−.390** (<.002)	−1.433** (<.014)
Years of school	−.448** (<.001)	.437** (<.001)	.192 (.169)
Married, spouse present	−1.618** (<.001)	.632** (<.001)	−.212 (.651)
Family income	−.00001** (<.001)	.00001** (<.001)	.00001* (<.018)
Employed	−.643** (<.001)	4.217** (<.001)	3.874** (<.001)
Number of children	−.306** (<.001)	.112 (.067)	−.052 (.649)
Medicare	.257 (.319)	−3.774** (<.001)	−3.632** (<.001)
Medicaid	1.923** (<.001)	−4.968** (<.001)	−3.993** (<.001)
Military	1.77** (<.001)	−2.198** (<.001)	−1.270 (.125)
No insurance	2.143** (<.001)	−.0423 (.860)	1.102 (.131)
Constant[c]	12.107** (<.001)	48.522** (<.001)	54.003** (<.001)
R^2 *F*-prob	0.086** (<.001)	.207** (<.001)	.018** (<.001)

* Indicates significance at .05 level in two-tailed test.

** Indicates significance at .01 level in two-tailed test.

[a] For regressions 1 and 2, *stata* regression command *svyreg* was used.

[b] For regression 3, *stata* instrumental variable regression corrected for weights and clusters *svyivreg* was used. The first stage regression used all exogenous variables, not just cigarette price, to generate the predicted cigarette price variable. Schem 4/20.

[c] Following binary variables omitted: White and no response for race, all other marital state categories, male, private insurance.

High level summary of the semester

- Fundamentals of probability theory
- Fundamentals of causal inference
  - Potential-outcomes framework
  - Causal graphical models
- Fundamentals of statistical inference
  - Parameter estimation, confidence intervals, hypothesis testing
  - Frequentist and Bayesian approaches
- Regression modeling
  - The beauty of ordinary least squares
  - Generalized linear models
  - Hierarchical/multi-level/mixed-effects models
- Causal effect estimation: back door, front door, instr. vars
- Nonparametric methods

Statistical analysis **do's** and **don'ts**

Do:

- Translate each scientific question into a mathematical form, so that statistical inference on model parameter(s), and/or model comparison, answers the question
- Keep in mind the dimensions & units for all continuous variables in your data—both predictors and responses
- Consider potential confounding factors that would interfere with desired causal interpretations of parameter estimates, and include the right set of controls
- Keep in mind the potential limitations of any parametric assumptions you're making (e.g., that $y \sim x$ is linear)
- Carefully consider the grouping structure of your data; to account for it, use random effects that are maximal with respect to the analysis design for your scientific question

Statistical analysis **do's** and **don'ts**

Do:

- Ensure that your parameters of interest are identifiable
- Focus on the *precision* (CI size) of parameter estimates
- Keep in mind that, at least in the most popular software package(s) such as `lme4`, confidence intervals reported for a fitted mixed effects model are conditional on a point estimate of the random effects covariance matrix, $\hat{\Sigma}$
  - There may be considerable uncertainty regarding $\hat{\Sigma}$, but `lme4` doesn't tell you anything about that!
  - Because of this, **do** prefer likelihood ratio tests to t or z based confidence intervals
  - **Do** also consider Bayesian fitting of mixed models, e.g. with `brms`, to marginalize out uncertainty about Σ

Statistical analysis **do's** and **don'ts**

Do:

- Think about whether, when, and how you might convert your scientific question to a t test, with corresponding effect size estimate
- Use Monte Carlo simulation!
  - To improve your understanding of the models and algorithms you're using
  - For power analysis
  - ...for surely many other reasons that I haven't thought of!

Statistical analysis **do's** and **don'ts**

Don't:

- Transform your dependent variable just because its distribution doesn't perfectly match the residual noise assumptions of your data; consider the impact on:
  - the family of distributions your model can express; and
  - the interpretation of your model coefficients
- Include as "controls" predictors that are (plausibly) causally downstream of your response variable – it can bias the parameter estimates in your model relative to the causal effect(s) you want to estimate
- Be afraid of "singular fits" in point estimation of mixed effects models
- Drop design-critical random effects because of model convergence failure

Statistical analysis **do's** and **don'ts**

Don't:

- Mistake a large p-value (e.g., p>0.2) for strong evidence for the null hypothesis
- Do Bayesian hypothesis testing without principled and carefully justified choice of priors
  - Instead, **do** use domain knowledge to determine what might constitute a "practically equivalent to null" effect size, and estimate your model parameters of scientific interest precisely enough to determine how large your effect size is relative to that (this is basically Kruschke's *region of practical equivalence, or* "ROPE")