INTRODUCTION TO STATISTICAL MODELING

Confidence and collinearity

Understanding a covariate

- Explanatory variable that's not of direct interest...
- ...But is important in the system under study
- Essential to constructing a model that reflects the real world

School outcomes (revisited)

```
# Modeling without covariates
> mod1 <- lm(sat ~ expend, data = SAT)
> effect_size(mod1, ~ expend, bootstrap = TRUE)
   slope stderr_effect expend to:expend
1 -20.89           5.2  5.768      7.13

# With frac as a covariate
> mod2 <- lm(sat ~ expend + frac, data = SAT)
> effect_size(mod2, ~ expend, bootstrap = TRUE)
  slope stderr_effect expend to:expend frac
1 12.29             4  5.768      7.13   28
```

Adding more covariates

```
> mod3 <- lm(sat ~ expend + frac + ratio, data = SAT)
> effect_size(mod3, ~ expend, bootstrap = TRUE)
  slope stderr_effect expend to:expend frac ratio
1 11.01           3.3  5.768      7.13   28  16.6

> mod4 <- lm(sat ~ expend + frac + salary, data = SAT)
> effect_size(mod4, ~ expend, bootstrap = TRUE)
  slope stderr_effect expend to:expend frac salary
1 13.33           6.8  5.768      7.13   28  33.29

> mod5 <- lm(sat ~ expend + frac + salary + ratio, data = SAT)
> effect_size(mod5, ~ expend, bootstrap = TRUE)
  slope stderr_effect expend to:expend frac salary ratio
1 4.463           9.5  5.768      7.13   28  33.29  16.6
```

Multicollinearity and alignment

- *Collinear* refers to two variables being in alignment
- Example: education and poverty
  - May vary at the individual level
  - Go hand-in-hand at the population level

Calculating alignment

```
> auxiliary_mod <- lm(expend ~ frac, data = SAT)
> R2 <- rsquared(auxiliary_mod)
> R2
[1] 0.35

# Variance inflation factor (VIF)
> 1 / (1 - R2)
[1] 1.5

# Standard error Inflation Factor
> sqrt( 1 / (1 - R2))
[1] 1.2
```

Collinearity between expend and covariates

```
# Include student-teacher ratio as a covariate
> R2 <- rsquared(lm(expend ~ frac + ratio, data = SAT))
> acos(sqrt(R2)) * 180 / pi # Angle in degrees
[1] 50
> sqrt(1 / (1 - R2)) # SE inflation
[1] 1.3

# Adding both ratio and salary
> R2 <- rsquared(lm(expend ~ frac + ratio + salary, data = SAT))
> acos(sqrt(R2)) * 180 / pi # Angle in degrees
[1] 19
> sqrt(1 / (1 - R2)) # SE inflation
[1] 3.1
```

Let's practice!

Start modeling!

Some important ideas

- Prediction error
- Effect size
- Covariates
- Quantitative vs. categorical responses
- Bootstrapping and cross validation

Beyond `lm()` and `rpart()`

- Generalized linear models (e.g. logistic regression)
- Machine learning techniques (e.g. random forests)
- Time series methods
- Survival analysis
- Causal inference
    - Donald Rubin: matched sampling
    - Judea Pearl: directed acyclic graphs (DAGs)

Modeling is a creative process

- Try a variety of things (models, explanatory variables, etc.)
- Build many models and compare them
- There is no right or wrong

What's ~~correct~~ useful?

"All models are wrong; some are useful."

- *George Box*

Let's get modeling!