

# Mixed Effects Regression

Dr. Rachael Tatman

September 26, 2018





Classification ✓

Regression ✓

Hypothesis testing ✓

Feature selection ✓

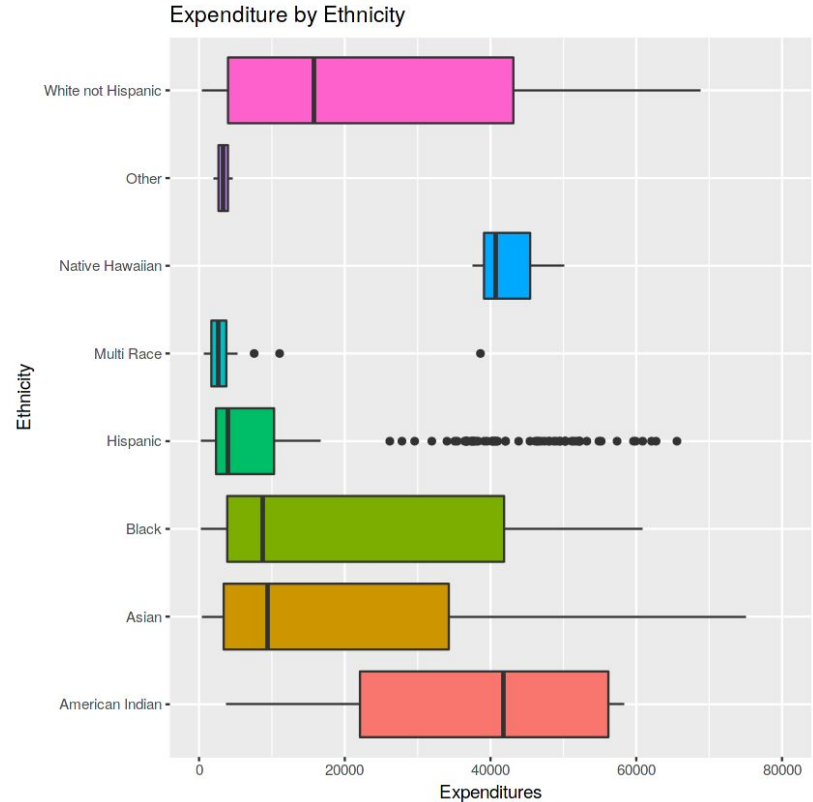
Dimensionality reduction ✓

Unsupervised learning ✗

# Simpson's Paradox

The state of California is spending more per student on white, not Hispanic students than on many other groups of students! 🤖

- Visualizing Simpson's Paradox by Kaggle user WesDuckett

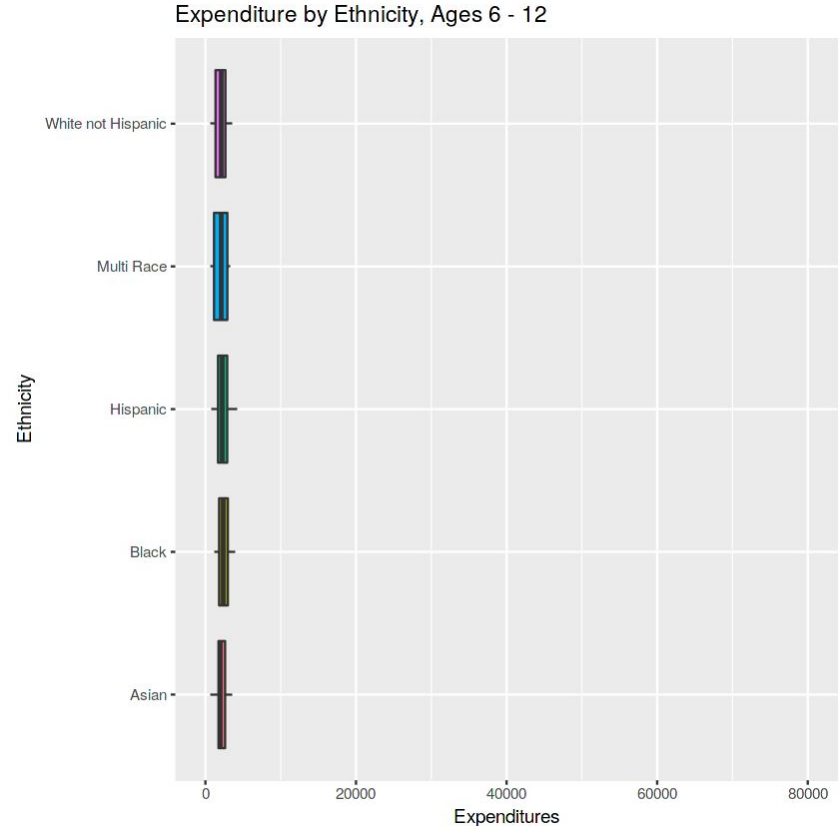


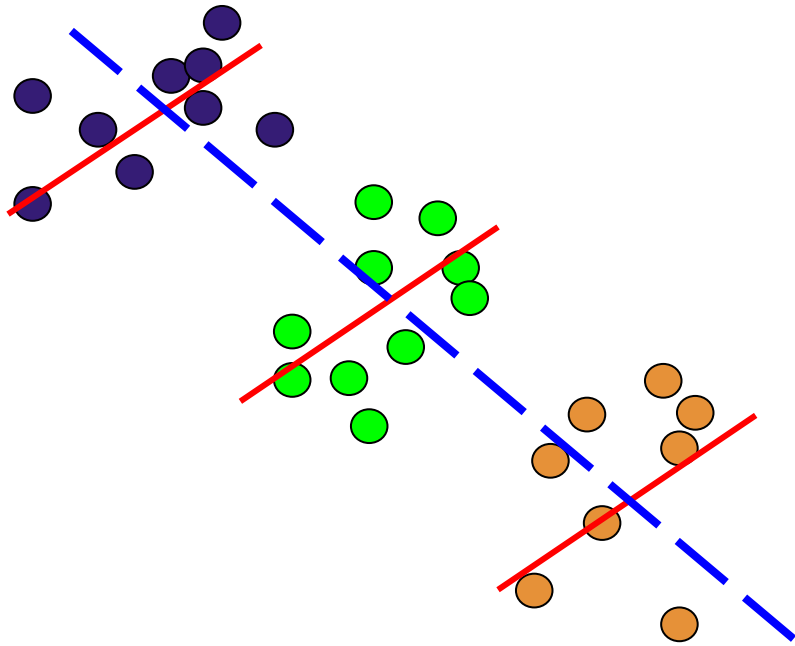
# Simpson's Paradox

...except that effect disappears when you look within age groups.

(There are a higher number of older special needs student who are white, not Hispanic & skewing the spending by pupil.)

- Visualizing Simpson's Paradox by Kaggle user WesDuckett





**It's important to control for grouping variables if you're looking for explanatory power!**

# General Workflow

1. Select **dependent** variable
2. Select **fixed** variables
3. Select **random** variables
4. Fit model
5. Check for **intraclass correlation** of random variables
6. Analyze model output

# General Workflow

1. Select **dependent** variable
  - a. What are you interested in measuring/affecting?
  - b. Examples: Click through rate in an A/B test
2. Select **fixed** variables
3. Select **random** variables
4. Fit model
5. Check for **intraclass correlation** of random variables
6. Analyze model output

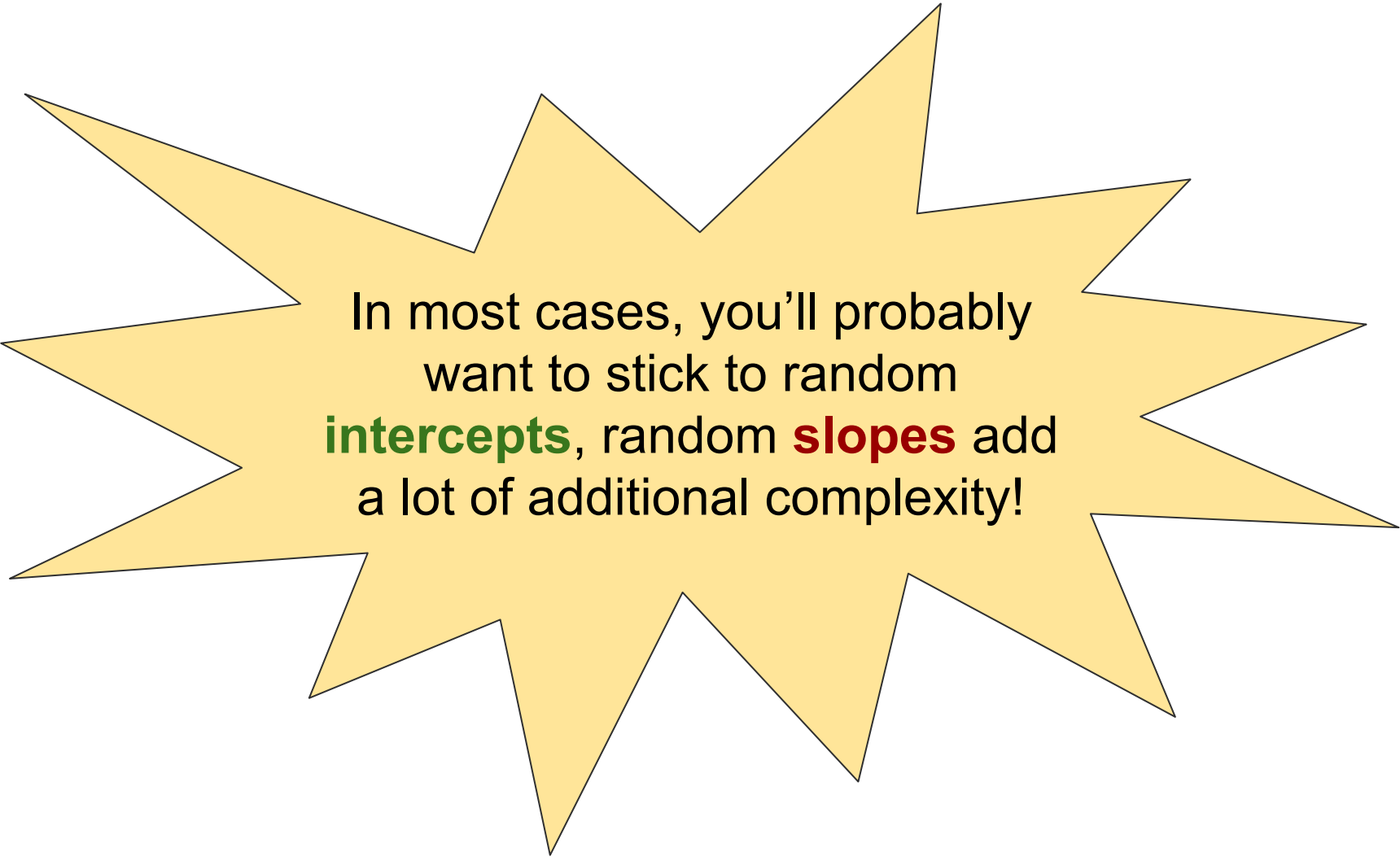


# General Workflow

1. Select **dependent** variable
2. Select **fixed** variables
  - a. What are you interested in modelling the effect of?
  - b. Example: Whether someone was in Group A or Group B
3. Select **random** variables
4. Fit model
5. Check for **intraclass correlation** of random variables
6. Analyze model output

# General Workflow

1. Select **dependent** variable
2. Select **fixed** variables
3. Select **random** variables
  - a. What groups might behave differently that you want to control for?
  - b. Example: Users who are using different browsers
4. Fit model
5. Check for **intra**class **correlation** of random variables
6. Analyze model output



In most cases, you'll probably  
want to stick to random  
**intercepts**, random **slopes** add  
a lot of additional complexity!

# General Workflow

1. Select **dependent** variable
2. Select **fixed** variables
3. Select **random** variables
4. Fit model
  - a. In R: lmer from the lme4 package
  - b. In Python: mixedlm from statsmodels
  - c. Validate as you would with other models of the same class
5. Check for **intra-class correlation** of random variables
6. Analyze model output

# General Workflow

1. Select **dependent** variable
2. Select **fixed** variables
3. Select **random** variables
4. Fit model
5. Check for **intraclass correlation** of random variables
  - a. Check that members of the classes you're treating as groups actually do behave like each other
6. Analyze model output

# General Workflow

1. Select **dependent** variable
2. Select **fixed** variables
3. Select **random** variables
4. Fit model
5. Check for **intraclass correlation** of random variables
6. Analyze model output
  - a. Visualizations (I like the sjPlot package in R)

# Example time!

Very nice Python example:

- <https://www.kaggle.com/ojwatson/mixed-models>

R walkthrough:

- <https://www.kaggle.com/rtatman/mixed-models-logistic-regression>

**Thanks!**

**Questions?**

Twitter: @rctatman



The image features the Kaggle logo, which consists of the word "kaggle" in a lowercase, sans-serif font. The letters are a light blue color. A small trademark symbol (TM) is located at the top right of the letter "e". The logo is centered on a dark gray background that has a repeating pattern of faint, light gray isometric cubes.

kaggle™