A Framework for Simple and Efficient Bootstrap Validation in SAS®, with Examples
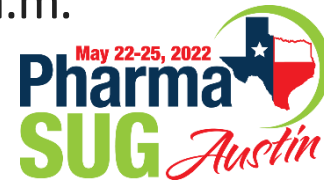
Isaiah Lankham
University of California
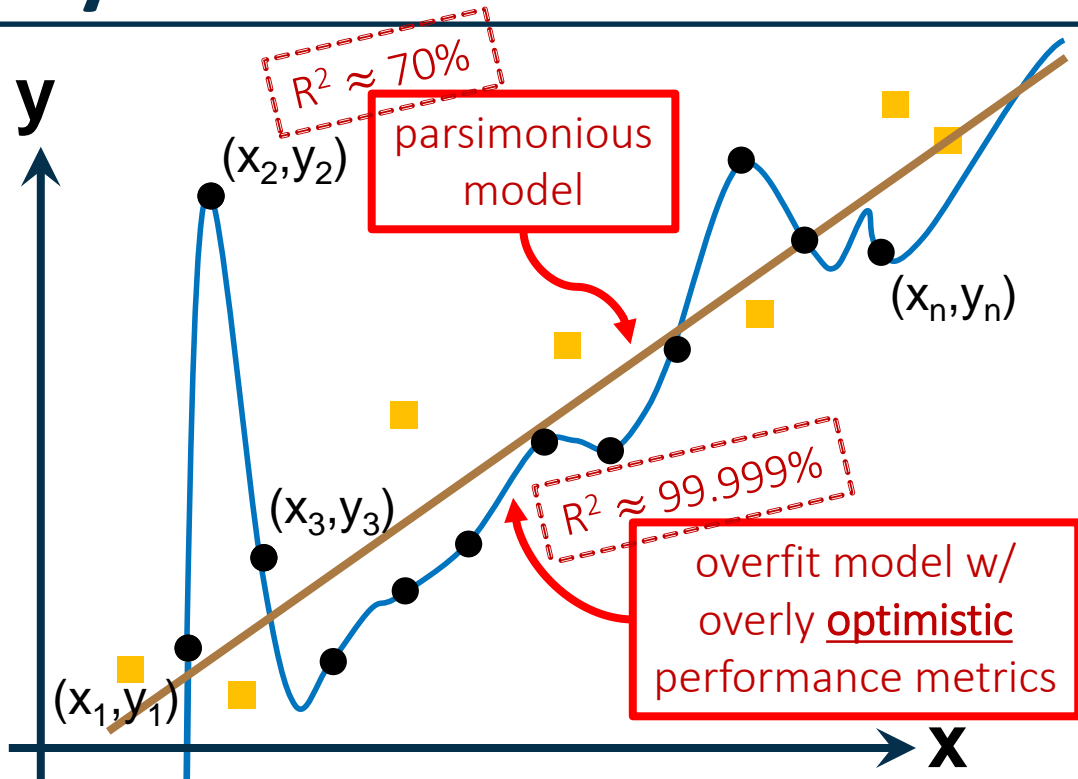Office of the President
Oakland, CA

Matthew T. Slaughter
Kaiser Permanente
Center for Health Research
Portland, OR

Statistics and Analytics Paper SA-149 • May 24, 2022 • 8:30 a.m.
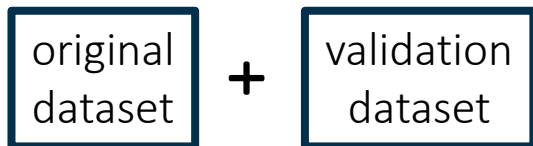
Presentation files: https://github.com/saspy-bffs

Why We Validate Predictive Models

y

$R^2 \approx 70\%$

(x_2, y_2)

parsimonious model

(x_n, y_n)

(x_3, y_3)

$R^2 \approx 99.999\%$

overfit model w/ overly **<u>optimistic</u>** performance metrics

(x_1, y_1)

x

x	y
x_1	y_1
x_2	y_2
$\vdots$	$\vdots$
x_n	y_n

Presentation files: https://github.com/saspy-bffs

How We Validate Predictive Models

external validation

original dataset **+** validation dataset

full data means no loss of statistical power

internal validation

original dataset

→ **partial** data for model development **+** **partial** data for model validation } e.g., split-sample and cross validation

→ **full** data for model development **+** **resampled** data for model validation } e.g., bootstrap validation

Presentation files: https://github.com/saspy-bffs
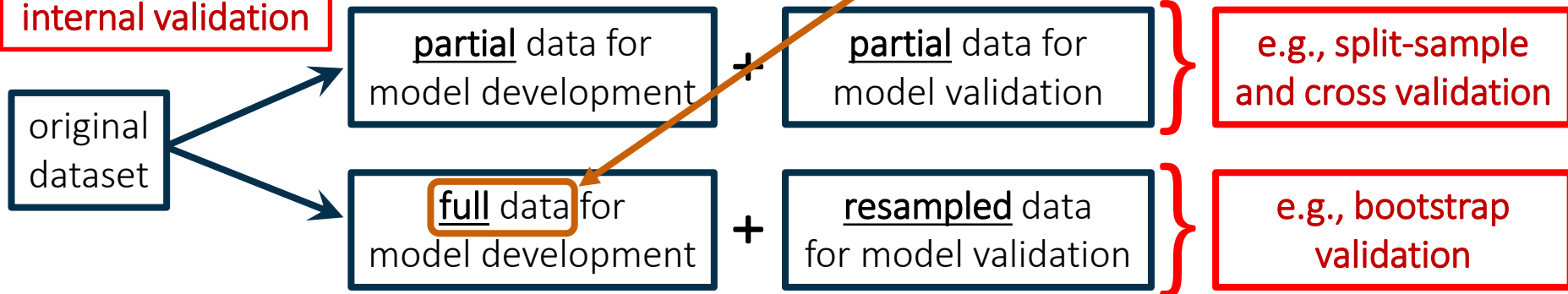
Call to Action!

- **Takeaway**: Use internal validation for predictive modeling. Bootstrap validation allows us to internally validate models without loss of statistical power.

- **Tradeoff**: Computing resources.

- **Read**: PharmaSUG 2022 Paper SA-149

- **Replicate**: Examples available on GitHub

Presentation files: https://github.com/saspy-bffs