Simple and Efficient Bootstrap Validation of Predictive Models Using SAS/STAT® Software

Isaiah Lankham
University of California
Office of the President
Oakland, CA

Matthew Slaughter
Kaiser Permanente
Center for Health Research
Portland, OR

Statistical Programming Section • 1 April 2020 • 11:30 a.m.

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

validate

From rms v5.1-4
by Frank E Harrell Jr

99.99th Percentile

Resampling Validation Of A Fitted Model's Indexes Of Fit

The `validate` function when used on an object created by one of the `rms` series does resampling validation of a regression model, with or without backward step-down variable deletion.

Keywords models, methods, regression, survival

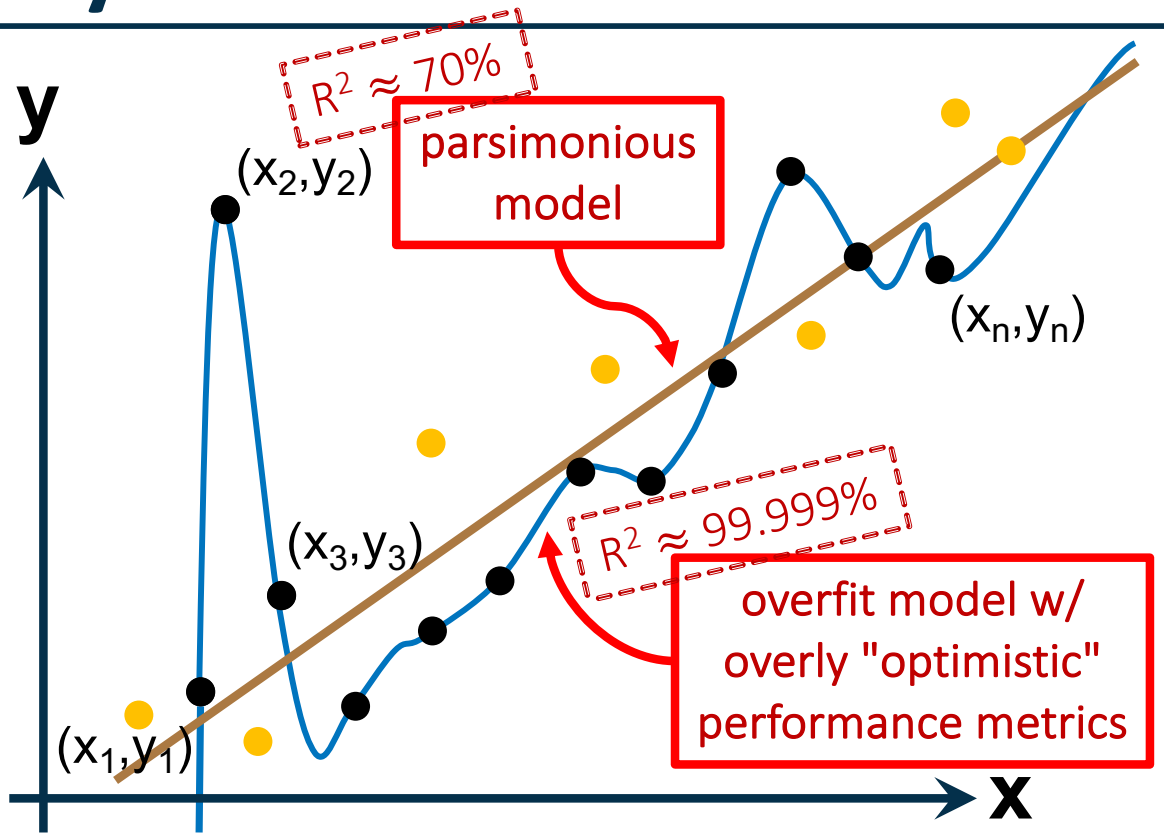
Usage

Motivation: How can this be replicated in SAS?

```
# fit <- fitting.function(formula=response ~ terms, x=TRUE, y=TRUE)
validate(fit, method="boot", B=40,
         bw=FALSE, rule="aic", type="residual", sls=0.05, aics=0,
         force=NULL, estimates=TRUE, pr=FALSE, …)
```

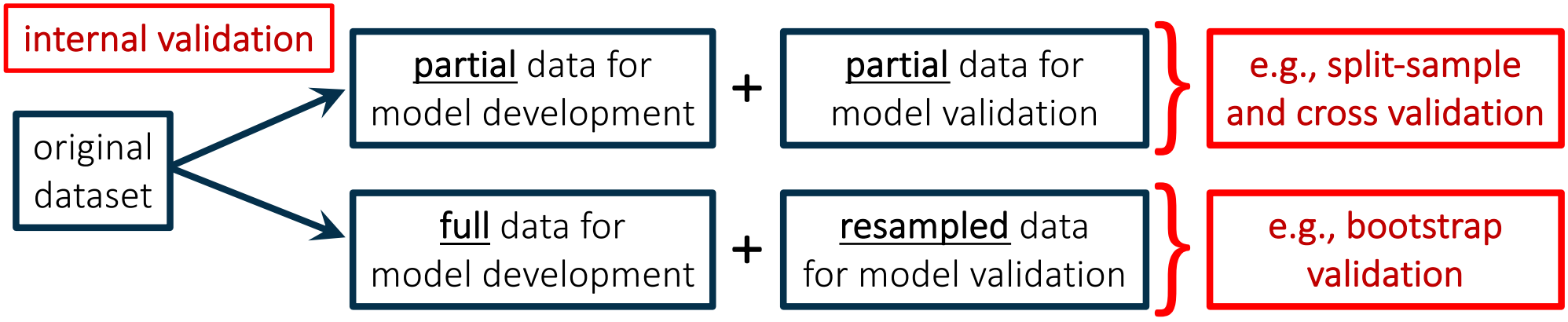
source: https://www.rdocumentation.org/packages/rms/versions/5.1-4/topics/validate

Why We Validate Predictive Models



x	y
x_1	y_1
x_2	y_2
⋮	⋮
x_n	y_n

How We Validate Predictive Models

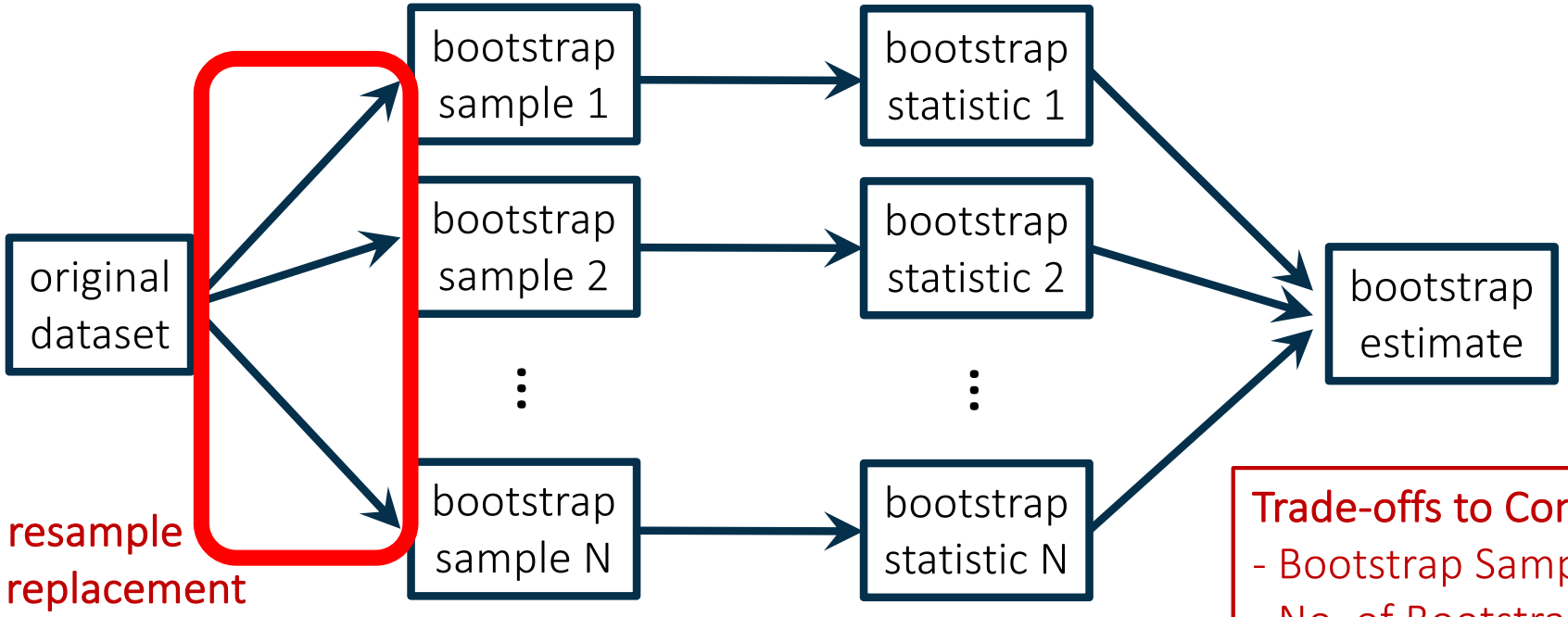


external validation

original dataset + validation dataset

Our Recommendation: Always use internal validation, and try bootstrap validation when computational feasible.

Bootstrap Framework

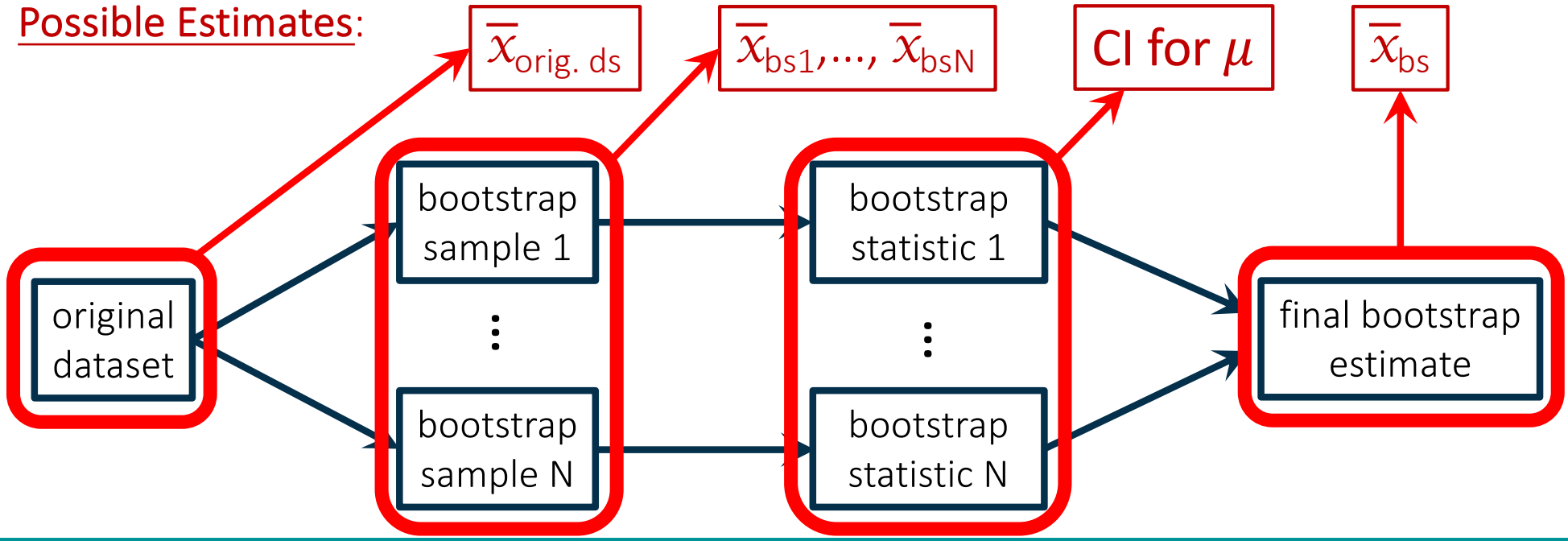


Trade-offs to Consider:
- Bootstrap Sample Size
- No. of Bootstrap Samples

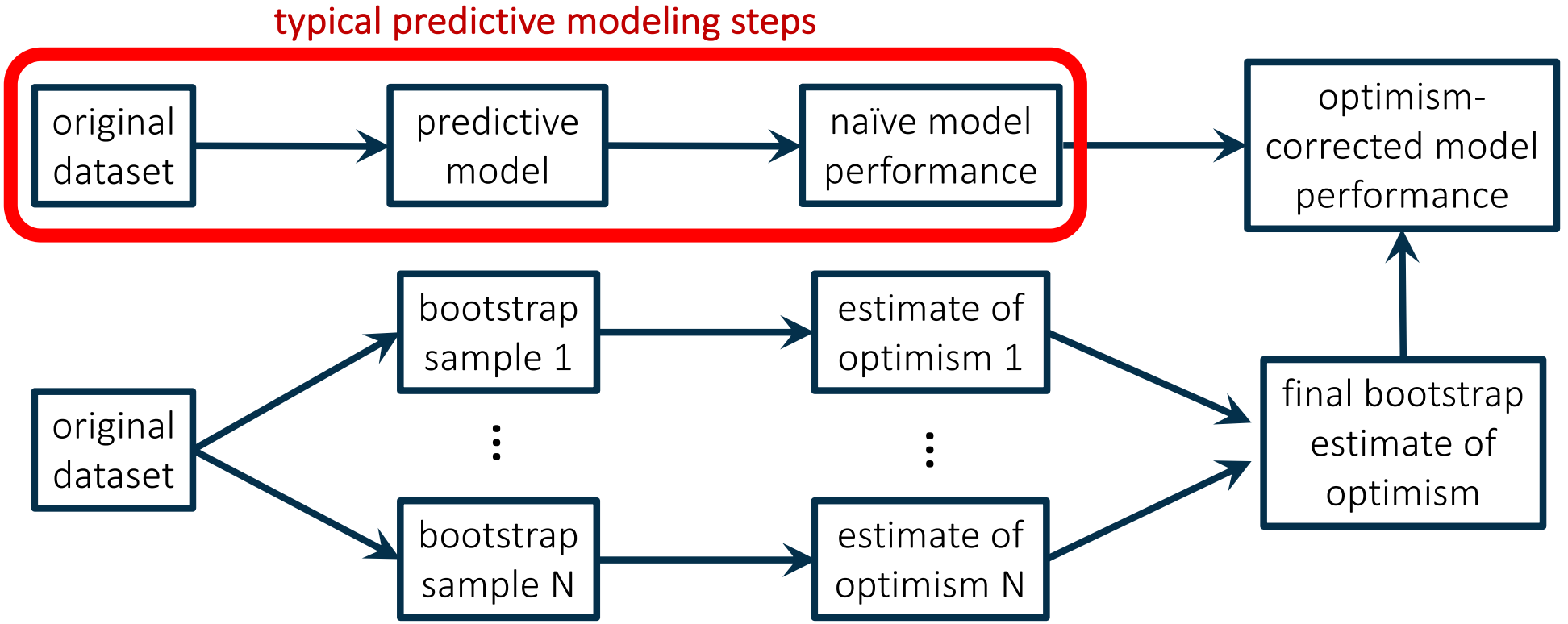
Bootstrap Example

Problem: Estimate an unknown population mean μ

Possible Estimates:



Bootstrap Validation Framework



Bootstrap Validation Example

- National Health and Nutrition Examination Study (NHANES) public dataset
- 21,004 observations (178 rows with usable data)
- 152 variables
- Binary Response Variable: mean systolic BP > 140
- 9 Predictors: age, BMI, cholesterol/triglyceride levels, …

source: https://wwwn.cdc.gov/nchs/nhanes/tutorials/samplecode.aspx

Call to Action!

- **Takeaway 1**: Use internal validation … always!
- **Takeaway 2**: It's okay to use the **full** sample for both model development and Bootstrap Validation.
- **Tradeoff**: Computing resources vs. better estimates of model performance than other validation techniques.
- **Read**: https://tinyurl.com/SGF2020Paper
- **Replicate**: Examples available on GitHub