TALL SOM FORTELLER

Two-Phase Learning

PREPARED BY TATSIANA PEKARSKAYA AND LI CHUN ZHANG, STATISTICS NORWAY

Statistisk sentralbyrå
Statistics Norway

Agenda

What is Two-Phase Learning and what is it for?

Procedure (algorithm, data split, outcome)

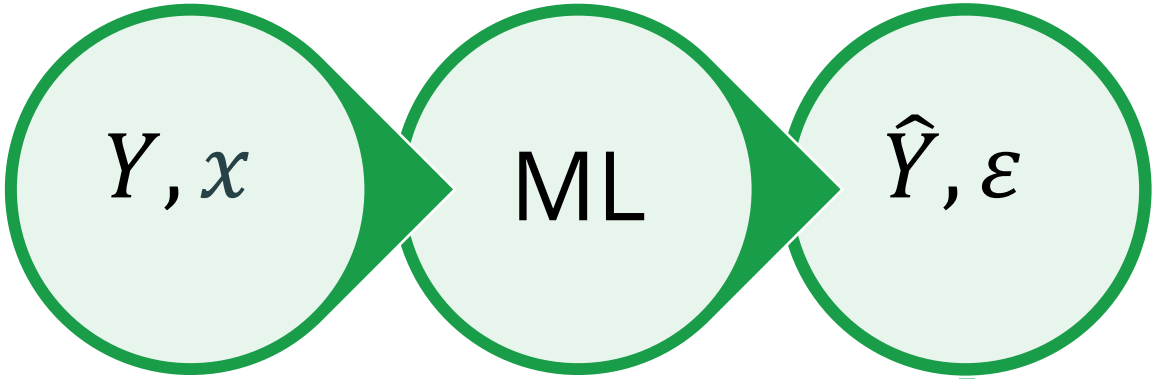
Practical example

Conclusions

What is Two-Phase Learning and what is it for?

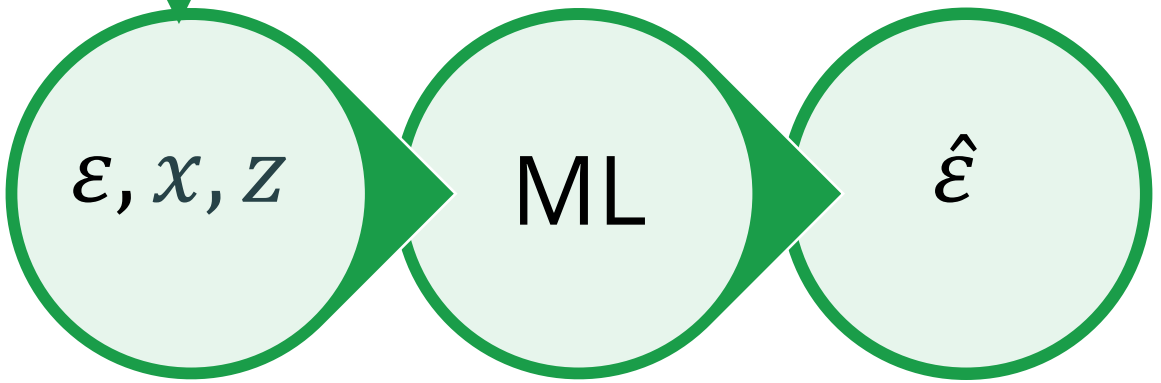
Two-Phase learning

1st-phase



Goal: $E(Y - \hat{Y}|x) = 0$

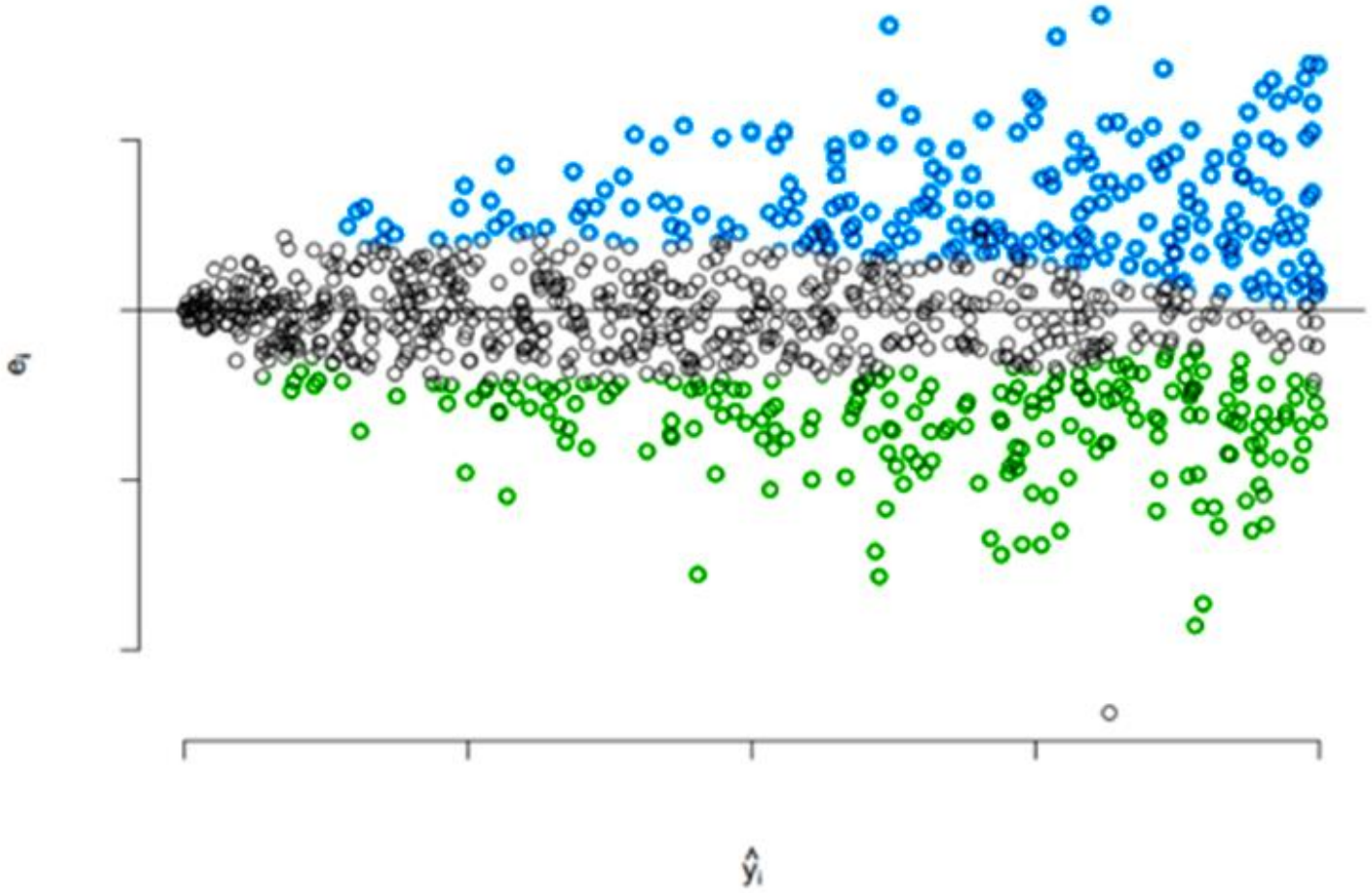
2nd-phase



Goals of 2nd-phase learning

- Capture potential mean and variance heterogeneity
- Provide individual-level description of prediction uncertainty
- Possibly adjust the 1st-phase model prediction on the flight

Mean & variance heterogeneity



$$E\left(Y - \hat{Y}\middle|x\right) = 0$$

On-the-flight correction

Adjusted prediction=

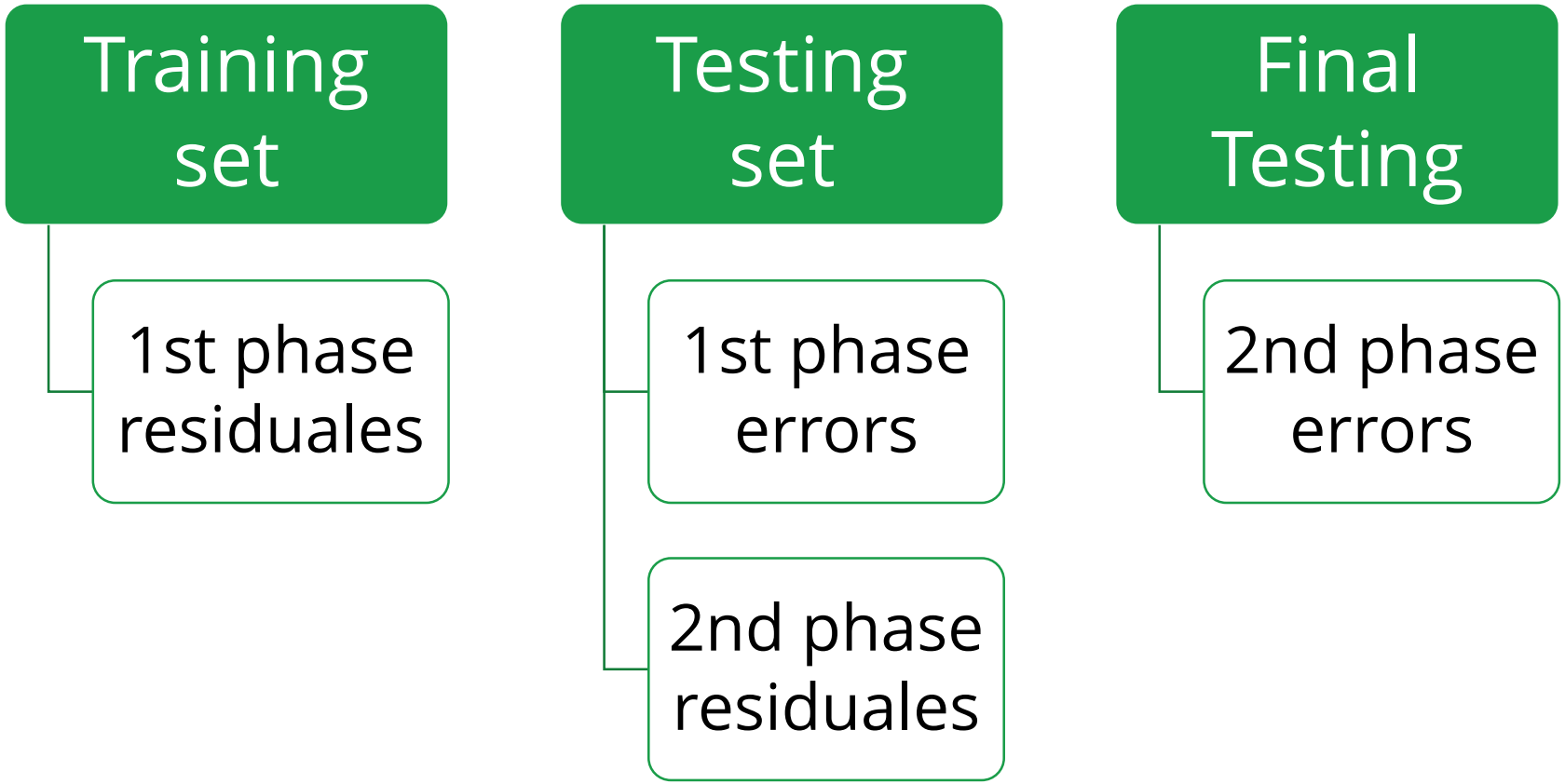
1st-phase model prediction +

2nd-phase error prediction

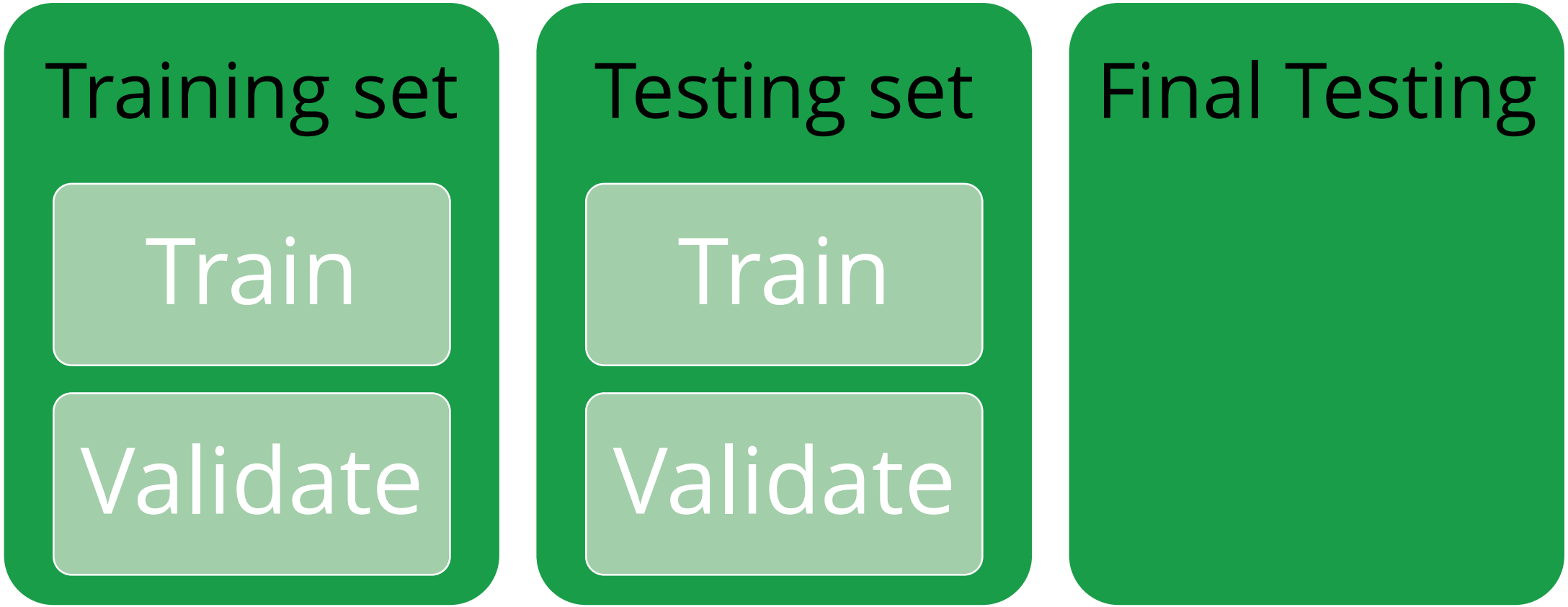
TALL | SOM FORTELLER

Procedure of Two-Phase Learning

1. Data split, Error Vs Residual



Data split with hold-on sets



2. 1st-phase learning

$Y = \hat{Y} + \varepsilon$, where $\hat{Y} = f(y|x)$

- f – 1st-phase model
- x – explanatory variables/features
- ε – 1st-phase error

3. 2nd-phase learning

$h(\varepsilon) = \widehat{h(\varepsilon)} + \xi$, where $\widehat{h(\varepsilon)} = g(h(\varepsilon)|x, z)$

- g – 2nd-phase model
- z – additional explanatory variables
- ξ – 2nd-phase error
- $h(\varepsilon)$ – function of ε, e. g. $(\varepsilon^2, sign(\varepsilon))$, incl. $h(\varepsilon) = \varepsilon$

Adding covariates

$g(h(\varepsilon)|x)$ vs $g(h(\varepsilon)|x,z)$

Errors' transformations

$h(\varepsilon)$	Mean heterogeneity	Var heterogeneity
ε	+	
ε^2		+
$sign(\varepsilon)$	+	+

A practical example

Imputing working time for agriculture survey

1st-Phase errors

True errors

	Testing set				*Rest of Population*			
	Overall	DA	ENK	OTHER	Overall	DA	ENK	OTHER
Mean	-2.073	15.194	-25.805	717.738	-157.587	-125.796	-173.411	701.6334
RMSE	1786.287	1667.939	1563.843	5406.214	1309.380	1256.090	1188.181	4407.440
PPS	0.400	0.406	0.399	0.424	0.326	0.390	0.321	0.465

2nd-Phase predicted errors

Feature	Mean($\hat{\epsilon}$)				Model Diagnostics		
Setting	Overall	DA	ENK	OTHER	R^2	RMSR	RMSE
x	-115.860	23.944	-118.323	-212.950	0.096	656.419	1255.236
x,z	-139.147	-71.477	-148.551	288.031	0.134	1300.835	1282.672

TALL

SOM FORTELLER

Conclusions

TALL SOM FORTELLER

Thanks for attention.
Questions?

Statistisk sentralbyrå
Statistics Norway