Model evaluation

What is model evaluation?

- How well does the model fulfill its objectives?
  - Objective
    - Good predictions
    - Good decisions

For a certain range of conditions

  - The result is on a continuum, from very poor to very good
  - We are treating the model as an engineering tool

Why evaluate?

- The modeler needs evaluation
  - Without evaluation, modeling is not a science
  - Think fortune telling
- The user needs evaluation
  - How can we make decisions if we don’t know reliability of information?

Predictive quality

Define prediction error

- $e=Y-f(X;\theta)$
  - Y is observation (for some target population)
  - $f(X;\theta)$ is model (f=equations, X=inputs, θ=parameters)
- We are interested in distribution of e
  - We don't know e for each prediction
  - If we did, we would get perfect predictions

Two viewpoints for prediction error

1. Model equations and parameters are fixed. Inputs are perfectly well known.
   - How well does this specific model predict?
   - e has distribution because of Y
2. Model equations, parameters and inputs are uncertain.
   - How good are predictions, averaged over the distribution of models and parameters?
   - e.g. averaged over climate models for future climate
   - e has distribution because of Y and f(X;θ)

Summary of prediction error

- Mean squared error of prediction (MSEP).
- Squared error, for fixed model, averaged over target population.

$$MSEP = E\left\{\left[Y - f(\hat{X};\hat{\theta})\right]^2\right\}$$

Estimation of MSEP

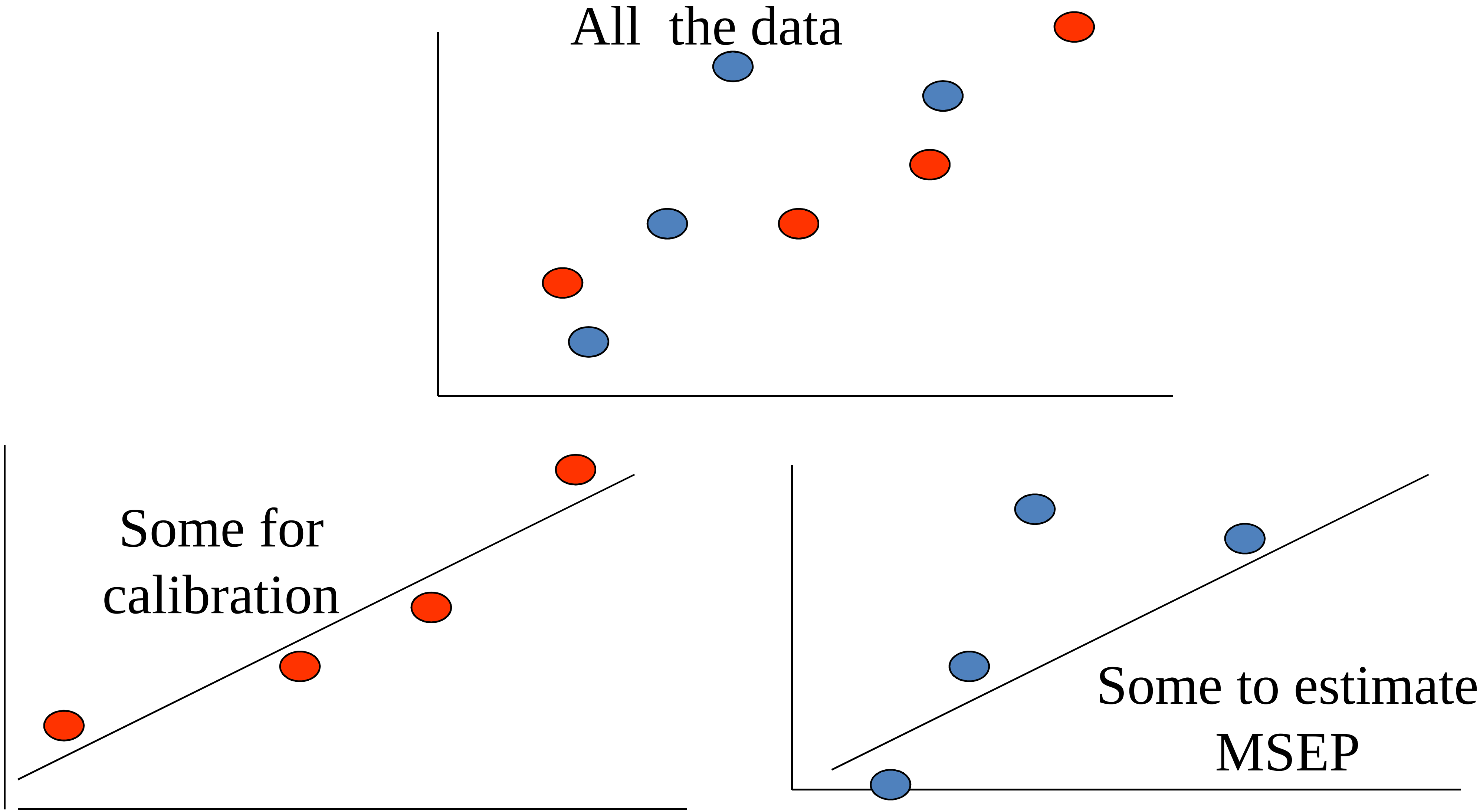
- In general, MSEP can't be measured.
  - concerns all predictions of interest
- Estimate MSEP based on a sample.

$$\hat{MSEP} = (1/N)\sum_{i=1}^{N}\left\{\left[Y_i - f(\hat{X};\hat{\theta})\right]^2\right\} = MSE$$

DANGER!

1. Sample must represent target population
   - Of course. If sample is different than target population, then errors for sample aren't necessarily representative of errors of population
     - e.g. Climate change. Are errors for sample representative of errors under climate change?

2. Sample musn't be used for calibration
   – If the model is specifically fit to the data, in general sample error < population error.
   – One solution is data splitting. Use part of data for calibration, separate data for evaluation


All the data
Some for calibration
Some to estimate MSEP

Model complexity and MSEP

- MSEP can be written as the sum of three contributions
  - Helps understand the relation of MSEP to complexity
  - Even though in practice we can't calculate the three contributions

The three sources of error

1. The model explanatory variables X don’t explain all the variability of the system
   – First term measures TRUE-BEST(X)

- The model used doesn’t have the same equations as the best function of X
  – Second term measures BEST(X) – f(X,θ*)
- The estimated parameters are not the best possible
  – Third term measures f(X,θ*)-f(X,θ)

- Illustrate with an artificial, simple case
  - The principle applies to all models

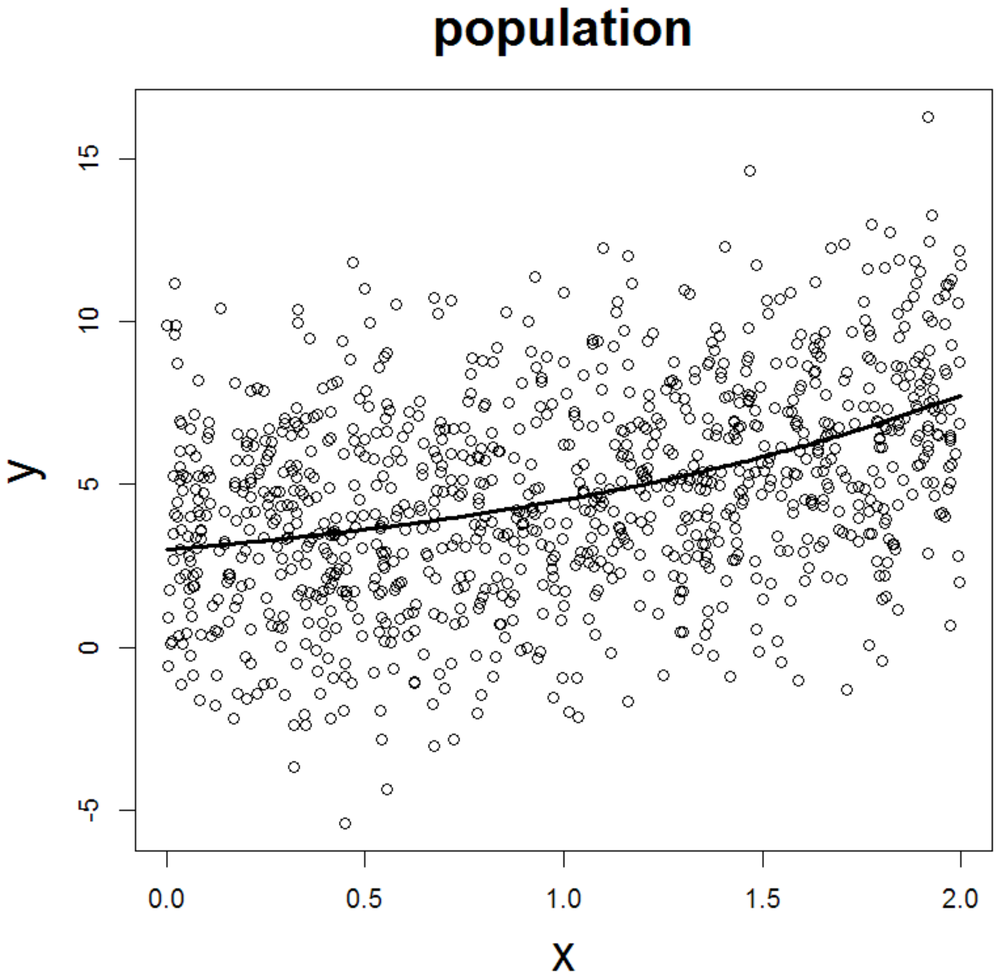
- TRUE behavior of y

  $TRUE=3+x+0.4x^2+0.1x^3+0.02x^4+\varepsilon \quad \varepsilon \sim N(0,3^2)$

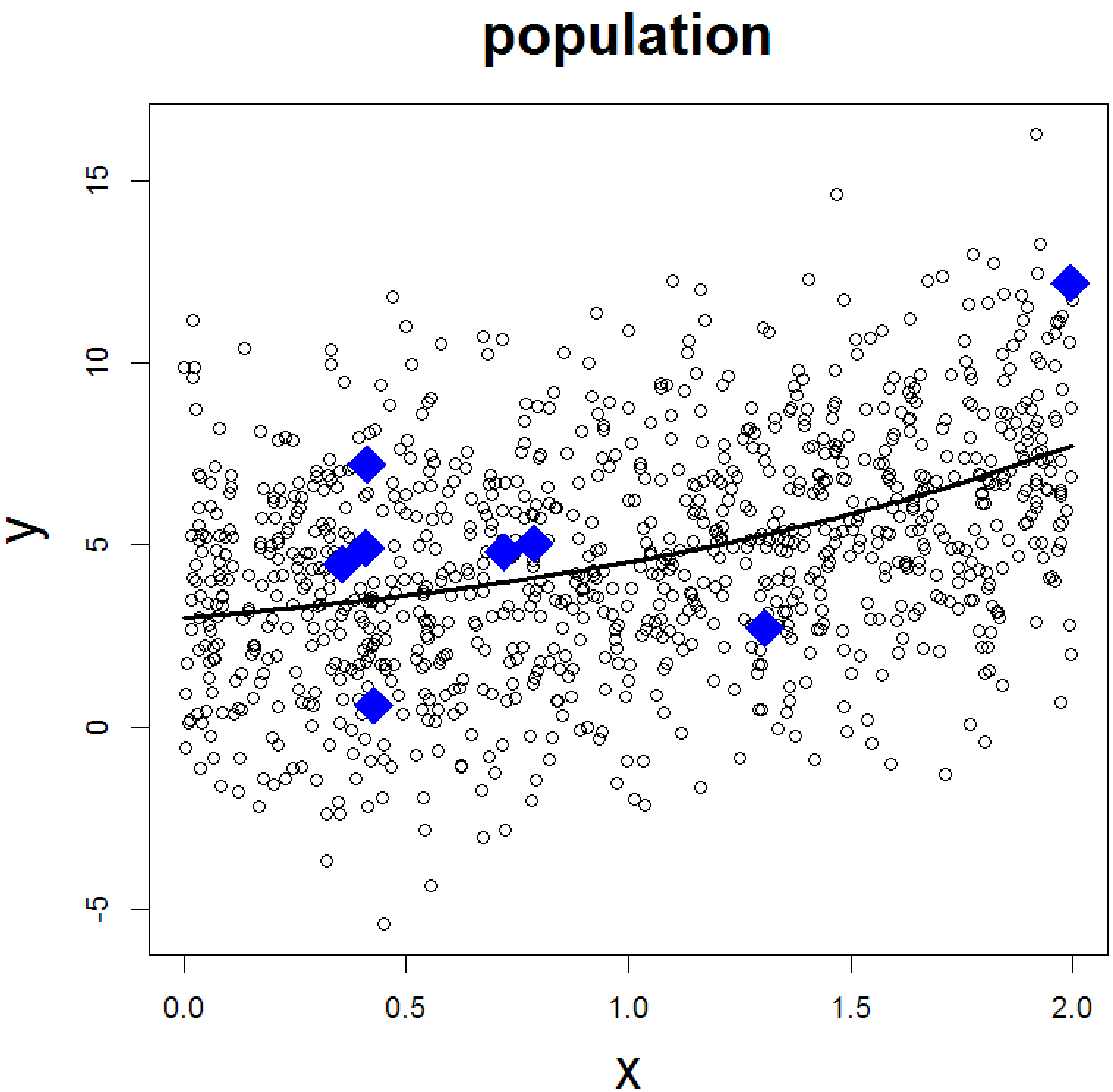
  x is the explanatory variable $x \sim U(0,2)$

- BEST(X)

  $BEST(X)=3+x+0.4x^2+0.1x^3+0.02x^4$



$BEST(X)=3+x+0.4x^2+0.1x^3+0.02x^4$



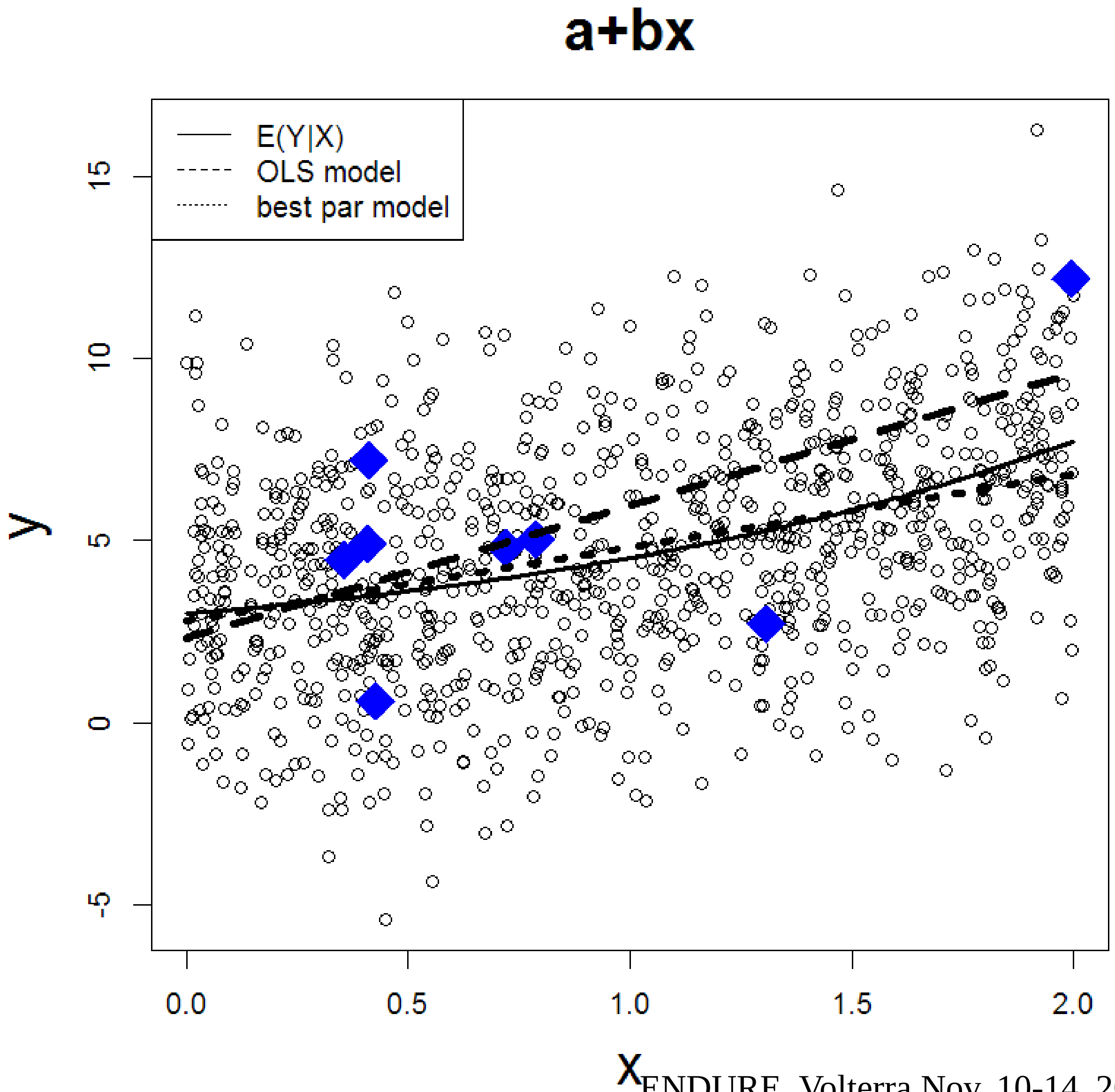
A random sample of size 8 for calibration

Look at a series of $f(X;\theta)$

- $f_2(X;\theta)=a+b1*x$ 2 parameters
- $f_3(X;\theta)=a+b1*x+b2*x^2$ 3 parameters
- $f_4(X;\theta)=a+b1*x+b2*x^2+b3*x^3$ 4 parameters
- $f_5(X;\theta)=a+b1*x+b2*x^2+b3*x^3+b4*x^4$ 5 parameters

  This is the correct model

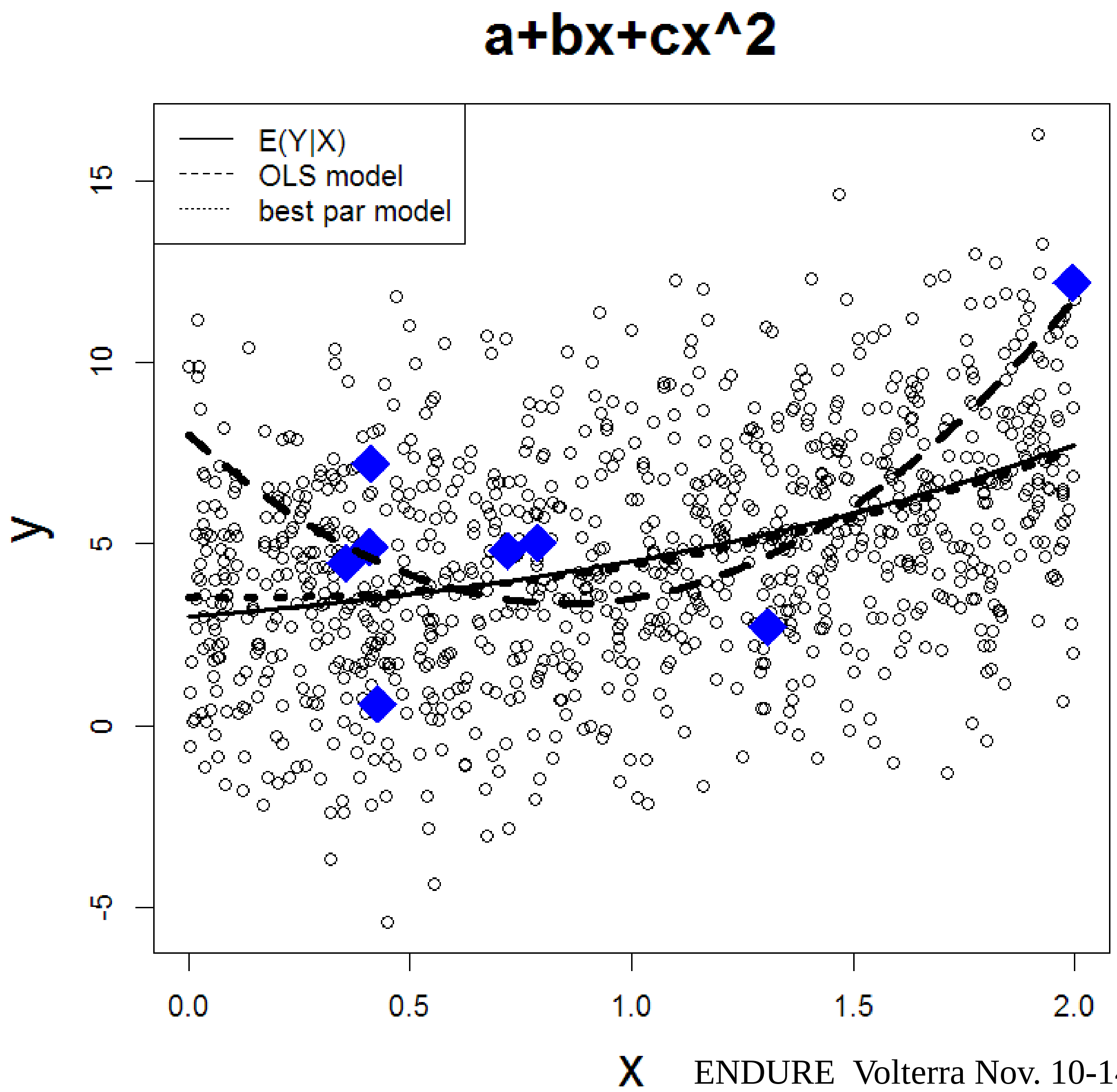
- For each model:
  - Calculate θ^* (use 1000 data points) and θ (OLS using 8 data points)
  - Calculate $MSE=(1/8)\sum(y_i-f(X_i,\theta_{OLS}))^2$
    - MSE measures fit to data
  - Calculate $MSEP_{\theta^*}=(1/1000)\sum(y_i-f(X_i,\theta^*))^2$
  - Calculate $MSEP_{\theta}=(1/1000)\sum(y_i-f(X_i,\theta))^2$



$f(X;\theta)$

Best(X)

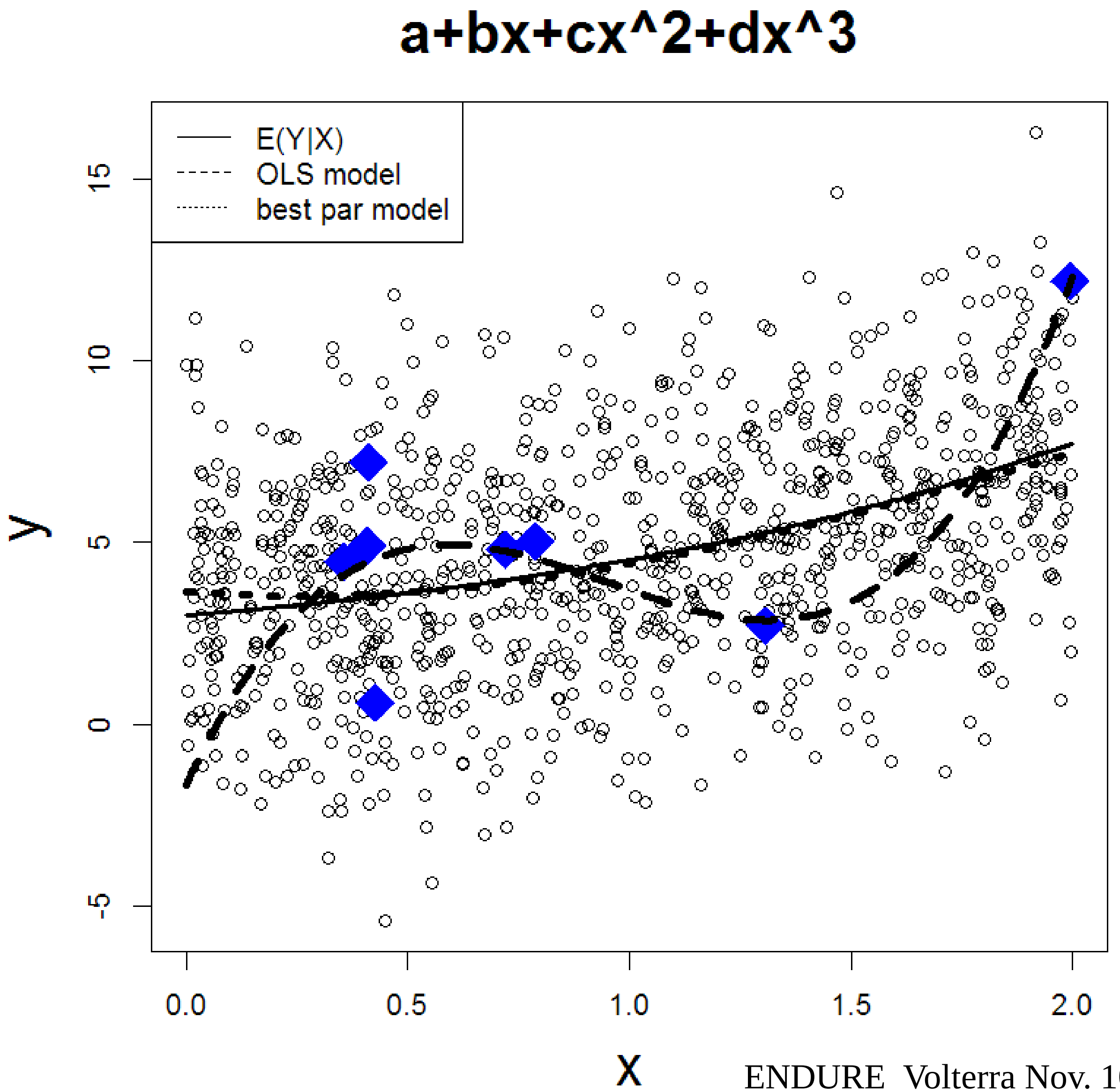
$f(X;\theta^*)$



f(X;θ)

Best(X)

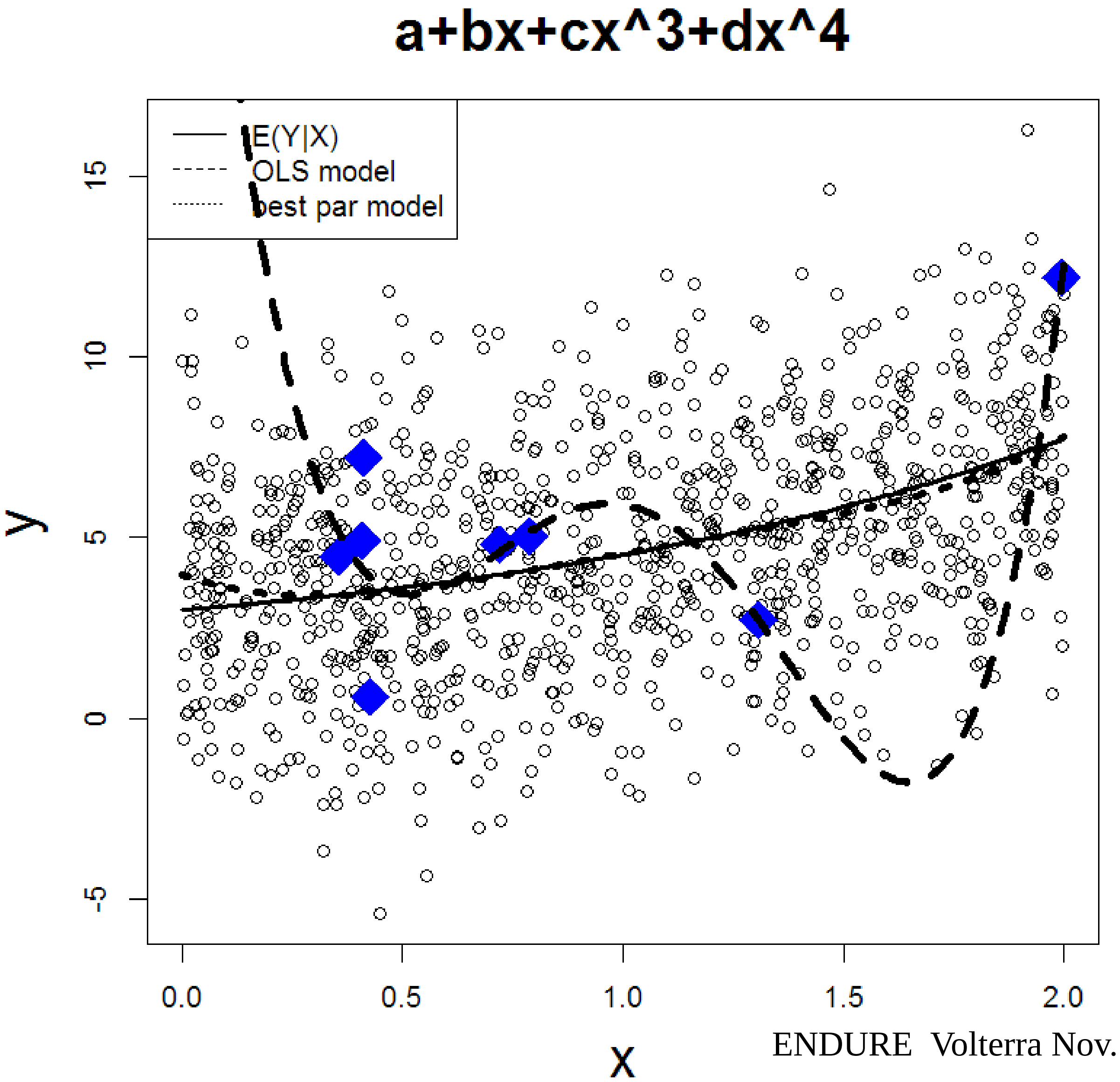
f(X;θ*)



f(X;θ)

Best(X)

f(X;θ*)



Correct model

f(X;θ)

Best(X)

f(X;θ*)

Number of parameters	$MSEP_{BEST}$	$MSEP_{\theta*}$	$MSEP_{\theta}$	MSE
2	9	9	10.9	6.3
3	9	9	12.0	3.9
4	9	9	12.3	3.0
5 (correct)	9	9	61.5	2.7

$MSEP_{BEST}$ same for all models (all use same x)

$MSEP_{\theta*}$ very close to $MSEP_{BEST}$ parameters

$MSEP_{\theta}$ increases with extra complexity (more parameters)

Best model is simpler than correct model

MSE can be very different than $MSEP_{\theta}$

- Best level of complexity?
  - Adding explanatory variables decreases $MSEP_{TRUE}$-$MSEP_{BEST}$
  - But in general increases $MSEP_{\theta*}$-$MSEP_{BEST}$ and $MSEP_{\theta*}$-$MSEP_{\theta*}$ (more functions, more parameters)
  - So include important X, not all X
  - Depends on amount of data

Conclusions

1. To evaluate model, define objectives
   – Including target population
2. Define criteria of evaluation
   – MSEP (or other)
   – Model fixed or uncertain

3. Estimate criterion (for fixed model)
   – Using data from target population
   – Using data that weren't used for calibration
4. Best model will have some intermediate level of complexity
   – More data allows more complexity

The end