Model calibration and model error

Daniel Wallach

INRA (Toulouse)

Essential to include error estimates with model results

- Project proposal coordinated by RMT
  - INRA-ICTA-CIRAD
  - Financed 2010-2013
  - Coordinator F. Brun

Objectives

- Associate an error estimate to model outputs
  - Usual practice: model evaluation, then use
  - Replace by: every prediction has an associated error estimate

- Identify methods and tools for error estimation applicable to different situations
- Apply to 8 case studies (test methods, classify cases)
- Make methods and tools available to other projects

- Major question:
- What is the effect of calibration on model error?
    - Context: model built up from individual equations
    - Then test full model on field data. Fix most parameters, estimate some to get better agreement to field data.

Answer

- Calibrated parameters become empirical constants that compensate for errors in model.
- Rest of talk is formal statistical statement of that, and conclusions for model error.

1. General statistical theory (not much)
2. Apply to individual equations
3. Apply to crop model
4. Draw conclusions

Statistical theory

- Simplify in 2 ways
  - Randomly chosen fields, one measurement per field (avoid correlated errors)
  - Asymptotic results (avoid randomness added by sample)
- Based on White 1981

True responses (observed values)

- Set of data Y_i
  - E.g. yields

Model

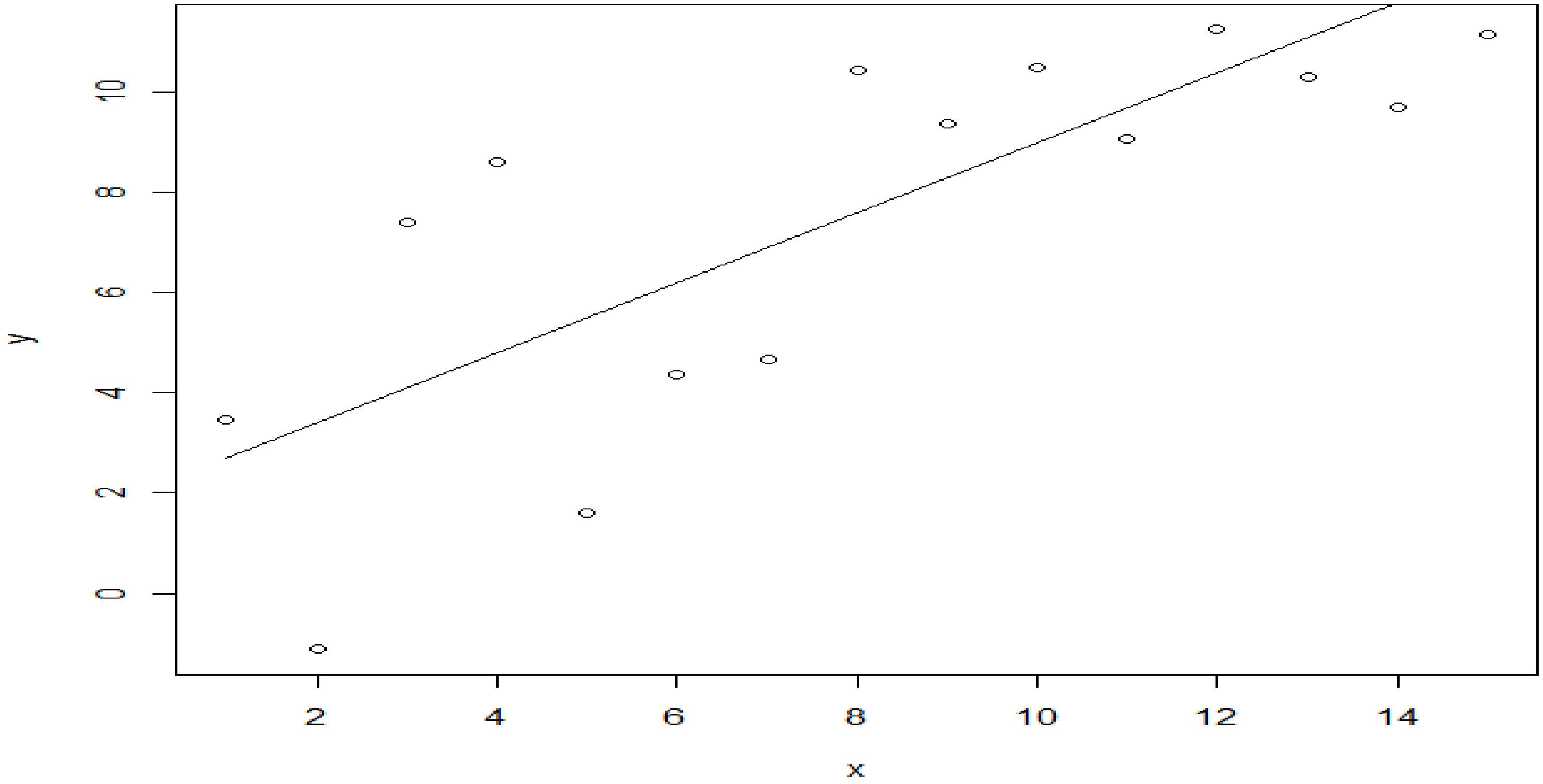
- We have some model $f(X_i;\theta)$
  - Model is well-behaved

Parameter estimation – OLS

$$\hat{\theta}_n = \arg\min_{\theta} \left[n^{-1} \sum_{i=1}^{n} (Y_i - f(X_i;\theta))^2 \right]$$

Case 1 Model is « correctly specified »

- There exists $\theta^{(0)}$ such that $Y_i = f(X_i; \theta^{(0)}) + \varepsilon_i$
- $E(\varepsilon_i)=0$, ε_i independent of X_i
- Then $\hat{\theta}_n \to \theta^{(0)}$
- That is standard text book case
- OLS gives, asymptotically, the true parameter value


y
0
2
4
6
8
10
x
2
4
6
8
10
12
14

Case 2. Model is « misspecified »

- There exists no $\theta^{(0)}$ such that
  - $E(\varepsilon_i)=0$, ε_i independent of X_i
  - Example,

$$Y_i = 0.09\exp(0.4X_i)$$

$$f(X_i;\theta) = \theta_1 X_i + \theta_2 X_i^2$$

$$Y_i = f(X_i;\theta) + \varepsilon_i$$

$$\varepsilon_i = 0.09\exp(0.4X_i) - (\theta_1 X_i + \theta_2 X_i^2)$$

OLS for misspecified model

- Define MSEP

$$MSEP(\theta)=\int(Y-f(x;\theta))^{2}dF(x)$$

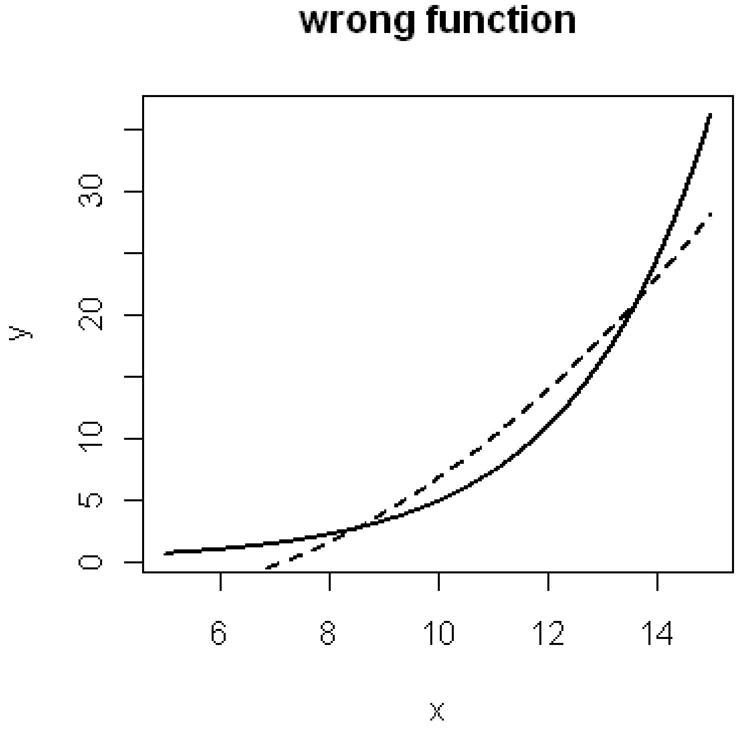
- Define

$$\theta^{*}=\underset{\theta}{\arg\min}\,MSEP(\theta)$$

- Then

$$\hat{\theta}_n \overset{a.s.}{\rightarrow} \theta^{*}$$

- OLS gives parameters that make model best possible approximation to true response.



Consequences of misspecification

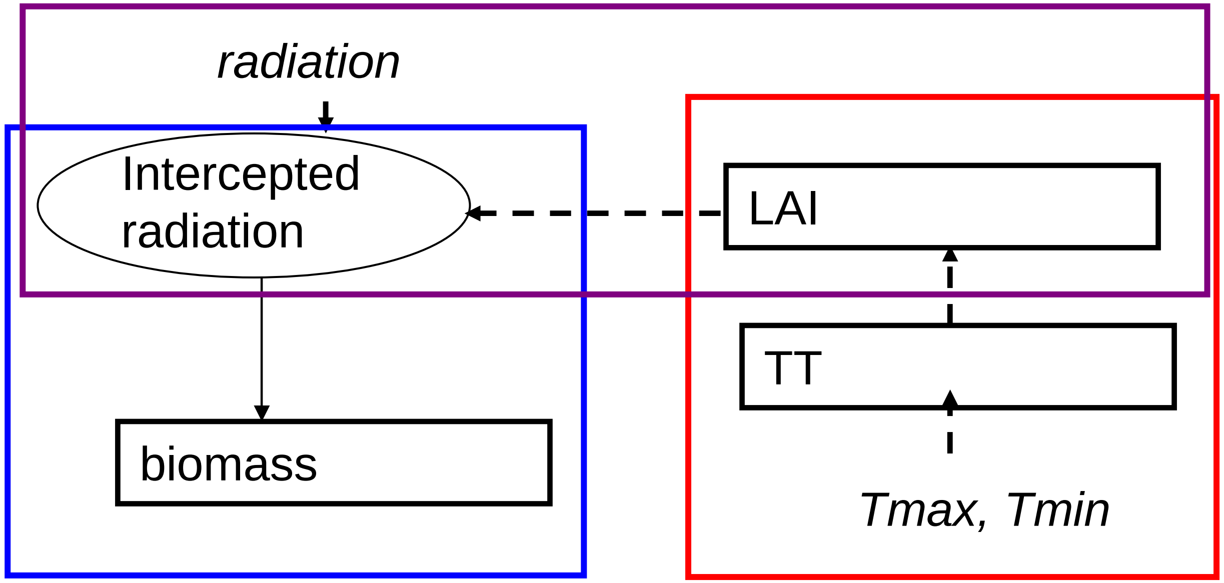
- OLS parameters are just empirical correction factors
  - No relation to parameters in true response
- Different false models give different OLS parameter values
- OLS parameter values depend on population for calibration

Individual equations

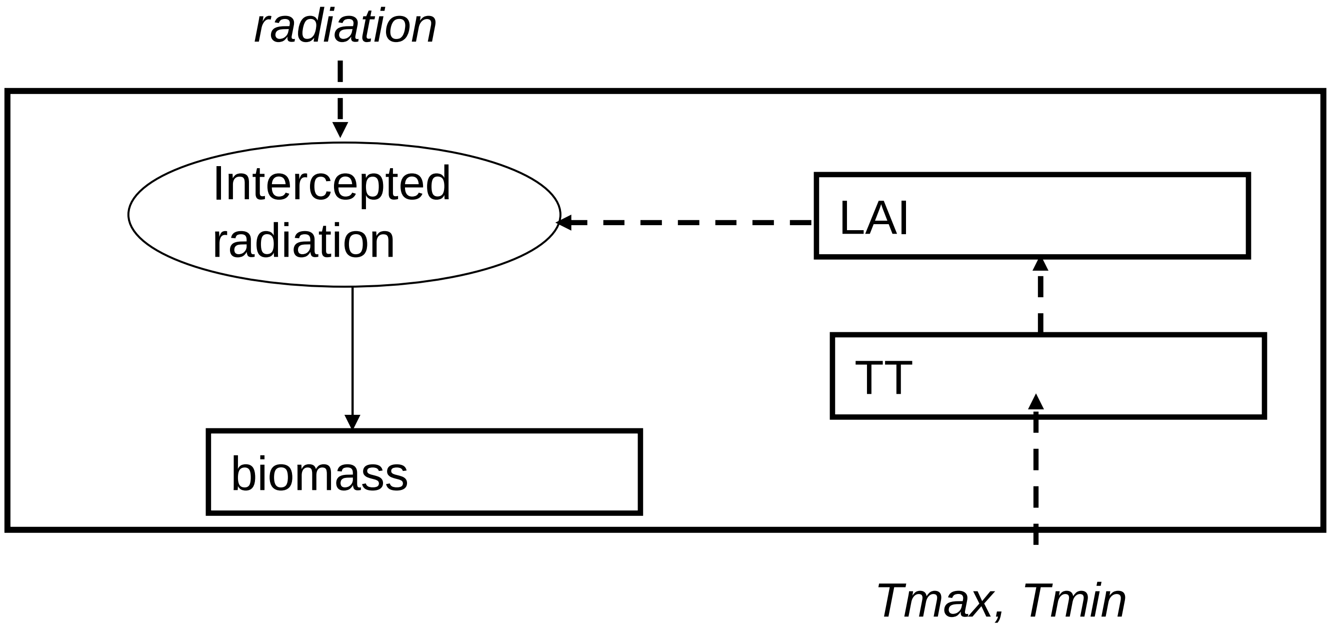
$$B_i^{ind}(d) = \theta_B * \sum IRAD_i(d')$$

$$LAI^{ind}(d) = \left[\frac{\theta_{LAI,1}}{1+(\theta_{LAI,2}/\theta_{LAI,1}-1)\exp(-\theta_{LAI,3} * TT(d-1))} \right]$$

$$IRAD^{ind}(d) = 0.48PAR(d) * (1-\exp(-\theta_{IRAD,1} LAI(d))$$



Crop model

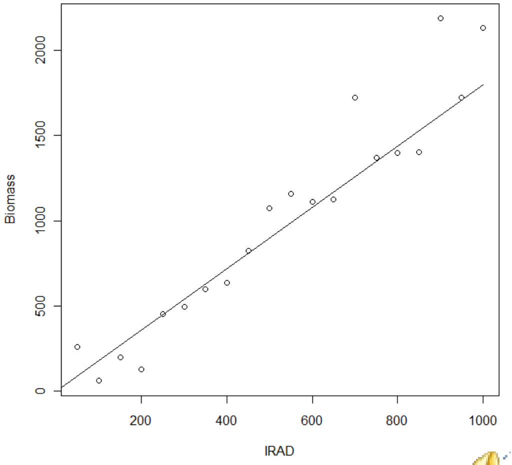


- Biomass model correctly specified if

$$B(d) = \theta^{(0)}{}_{B} * \sum IRAD(d') + \varepsilon_B$$

$$E(\varepsilon_B) = 0$$

$$\varepsilon_B, \quad IRAD \text{ all independent}$$



- OLS asymptotically gives true parameter value
- In general, we just have an approximate value of the parameter

- LAI model correctly specified if

$$LAI(d) = \left[\frac{\theta^{(0)}{}_{LAI,1}}{1 + (\theta^{(0)}{}_{LAI,2} / \theta^{(0)}{}_{LAI,1} - 1)\exp(-\theta^{(0)}{}_{LAI,3} * TT(d-1))} \right] + \varepsilon_{LAI}$$

$$E(\varepsilon_{LAI}) = 0$$

$$\varepsilon_{LAI}, \quad Tmin(d), Tmax(d) \ \ \text{all independent}$$

- Intercepted radiation model correctly specified if

$$IRAD(d) = 0.48PAR(d)*(1-\exp(-\theta^{(0)}{}_{IRAD,1}LAI(d)) + \varepsilon_{IPAR}$$

$$E(\varepsilon_{IPAR}) = 0$$

$$\varepsilon_{IPAR}, \quad PAR, LAI \quad \text{all independent}$$

Crop model

- Replace state variables by individual model expressions

$$B_i^{crop}(d) = \theta_B * \sum IRAD^{(crop)}{}_i(d')$$

$$IRAD^{crop}(d) = 0.48PAR(d) * (1 - \exp(-\theta_{IRAD,1} LAI^{crop}(d))$$

$$LAI^{crop}(d) = LAI^{ind}(d) = \left[\frac{\theta_{LAI,1}}{1 + (\theta_{LAI,2} / \theta_{LAI,1} - 1)\exp(-\theta_{LAI,3} * TT(d-1))} \right]$$

- Crop model mixes errors with explanatory variables

$$IRAD(d) = IRAD^{crop}(d) + \tau$$

$$\begin{aligned}
\tau &= IRAD(d) - IRAD^{crop}(d) \\
&= IRAD^{ind}(d) + \varepsilon_{IRAD} - IRAD^{crop}(d) \\
&= 0.48PAR(d)*(1-\exp(-\theta_{IRAD,1}LAI(d)) + \varepsilon_{IRAD} - \\
&\quad 0.48PAR(d)*(1-\exp(-\theta_{IRAD,1}LAI^{crop}(d)
\end{aligned}$$

Crop model is misspecified

- Even if all individual models correctly specified
- Because error not independent of explanatory variables
- If we fix some parameters in crop model at incorrect values, that also contributes to misspecification

Consequences of misspecification (true of all crop models)

- OLS tend to values that minimize MSEP
- OLS is a powerful way of correcting for errors in model

BUT

- OLS doesn’t give “true” parameter values of individual equations
- Different output variables (LAI, biomass, yield) will lead to different OLS parameters
  - There are no optimal parameters that minimize MSEP for all outputs
- Different target populations will lead to different OLS parameter values
  - Even if true response is the same
  - That’s why calibration is useful
  - But don’t have optimal parameters for a different target population

Extent of misspecification

- If slight, ignore
- Simulation studies indicate it’s relatively important

Simulation results

Estimated values for the parameter rue1. Calibration uses yield (Y) or biomass (BM) or a combination. Simulated data.

	true value	fixed 1	fixed 2	fixed 3	fixed 4	fixed 5
calibrate using Y	2.8	2.99	3.48	3.08	3.71	2.62
calibrate using BM	2.8	2.45	3.14	3.25	3.16	3.16
calibrate using Y and BM	2.8	2.54	3.20	3.22	3.25	3.02

Real data and model STICS

- With real data
  - should see different estimated parameters for different variables
  - reducing MSE for one variable can increase it for another

Real data STICS

	number of measurements	initial MSE	MSE after calibration using biomass	MSE after calibration using yield
yield	63	3.88	3.35	1.80
biomass	211	3.76	3.50	11.56
grain N	62	0.18	0.14	0.23
grain number	59	$2.98*10^7$	$2.63*10^7$	$1.13*10^7$
plant N	203	898	915	977

Conclusions

- Calibration can be very useful for improving prediction
  - For the variable, and target population, used in calibration
- But crop models don’t follow std statistical assumptions
  - Can’t assume that calibration improves prediction for other variables or target populations