Bayes Lecture 2

- Latent Variables
  - Errors in Variables
  - Missing Data
- Hierarchical models
- State Space (Hidden Markov) models

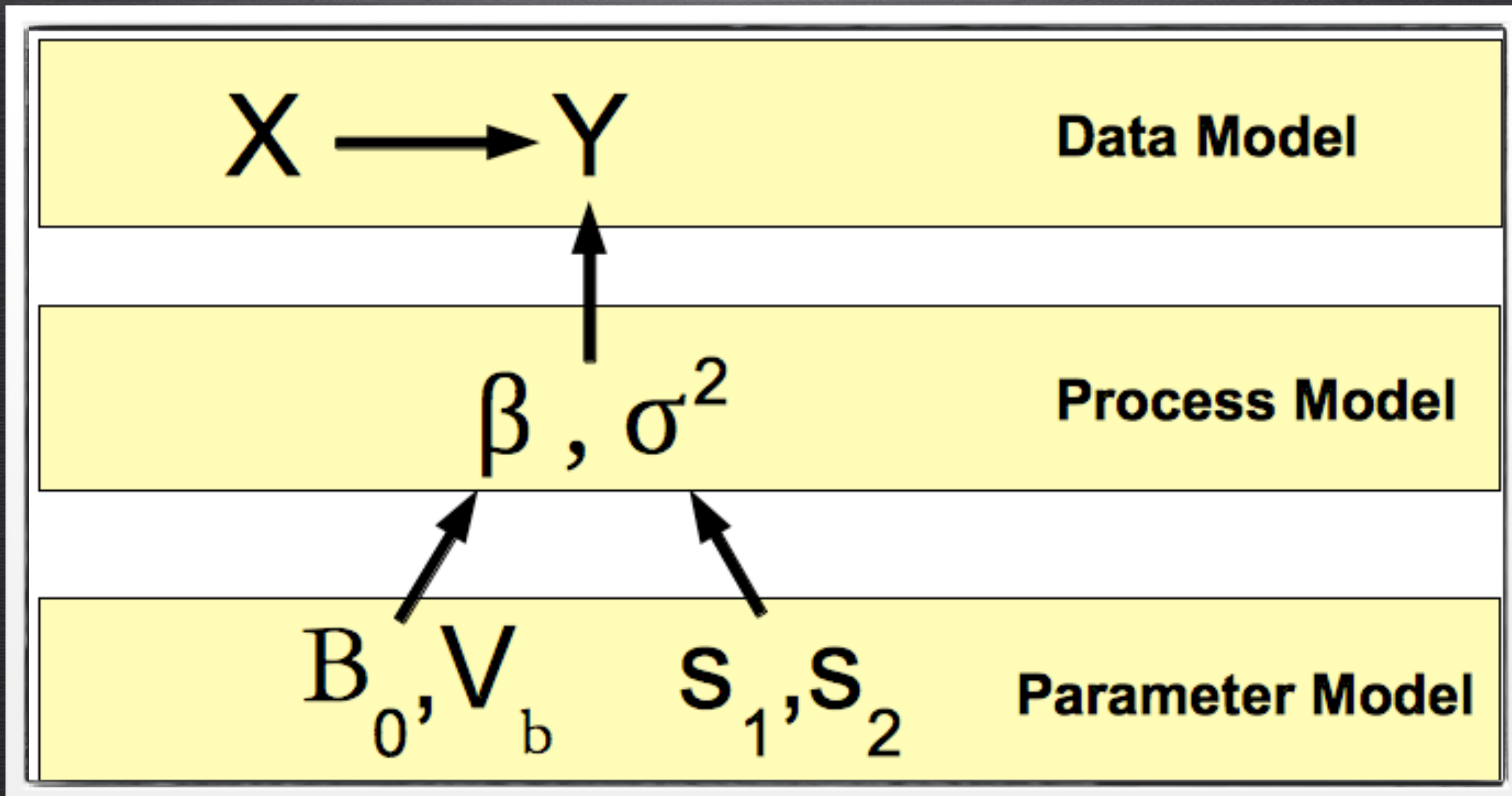
Linear model assumptions

The linear model rests on some important assumptions:

- Errors are additive and normally distributed
- Errors are homoskdastic (don't vary across Xs)
- Observations are independent (conditional on the linear predictor)
- Linear (in covariates) mean function
- All error/randomness is in the value of the response (i.e., the X values are precisely known)
- There is no (systematic) missing data

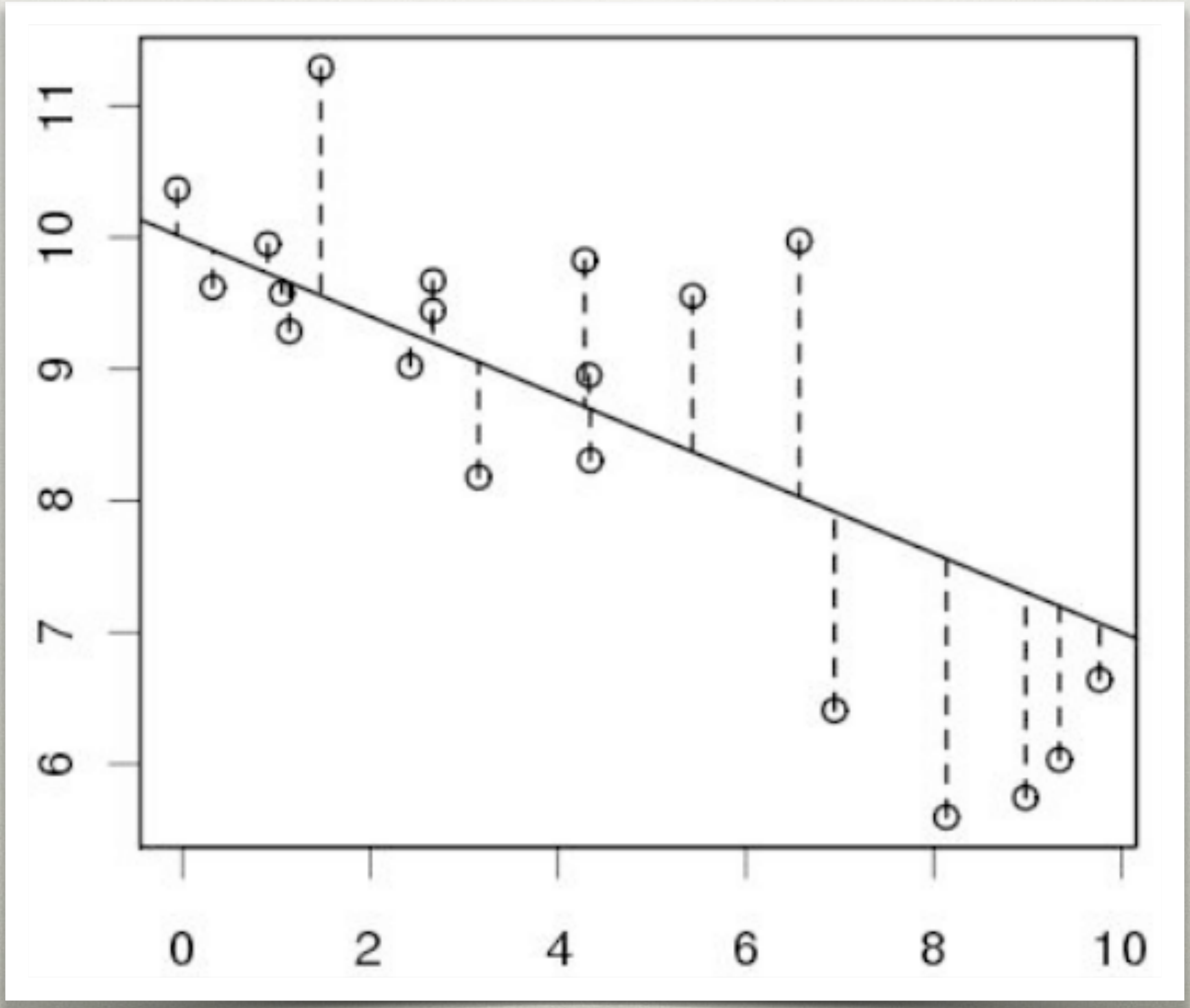
Ecological data rarely conform to these assumptions!

$$\vec{y} \sim N(\boldsymbol{X}\vec{\beta}, \sigma^2)$$

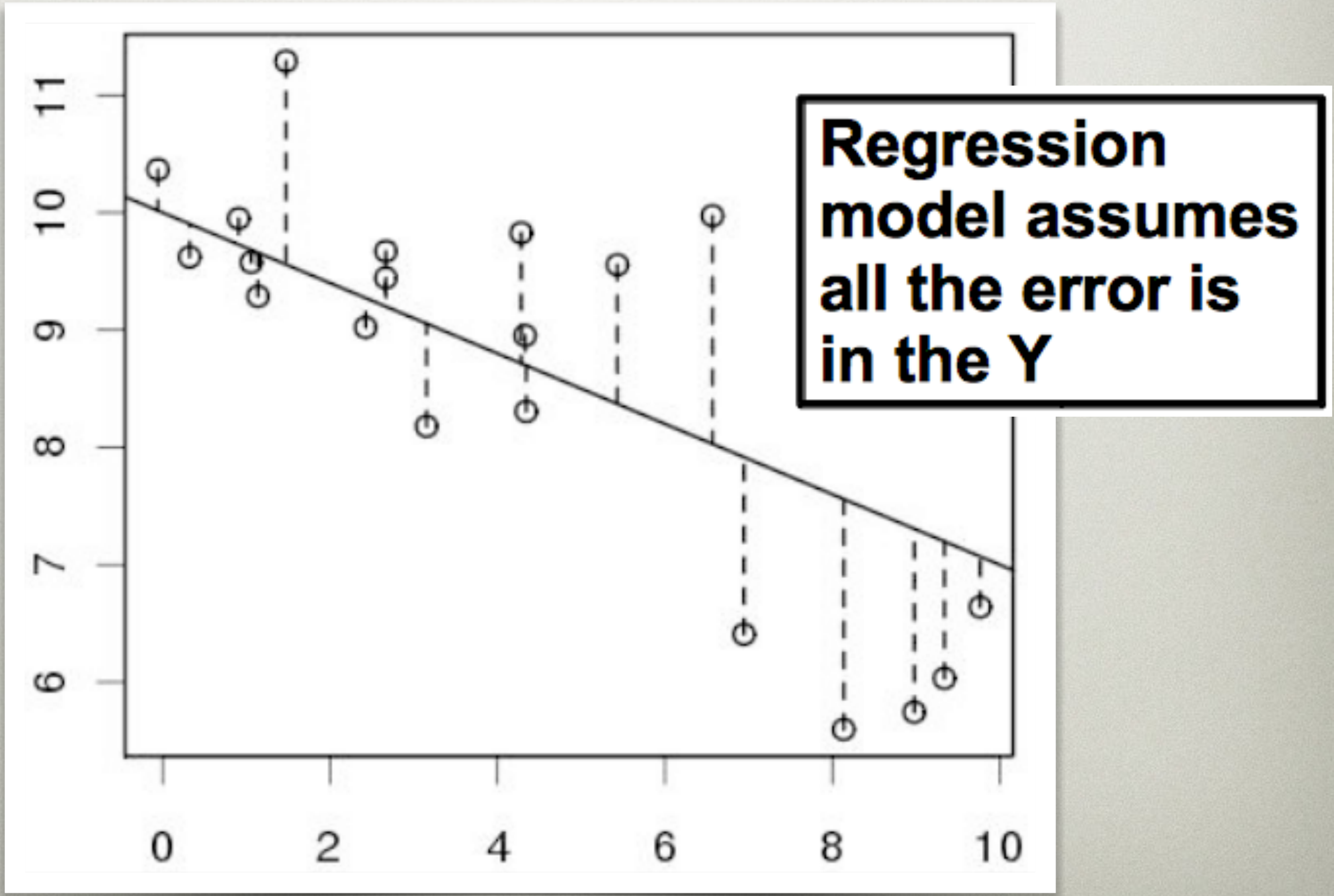


Graph Notation

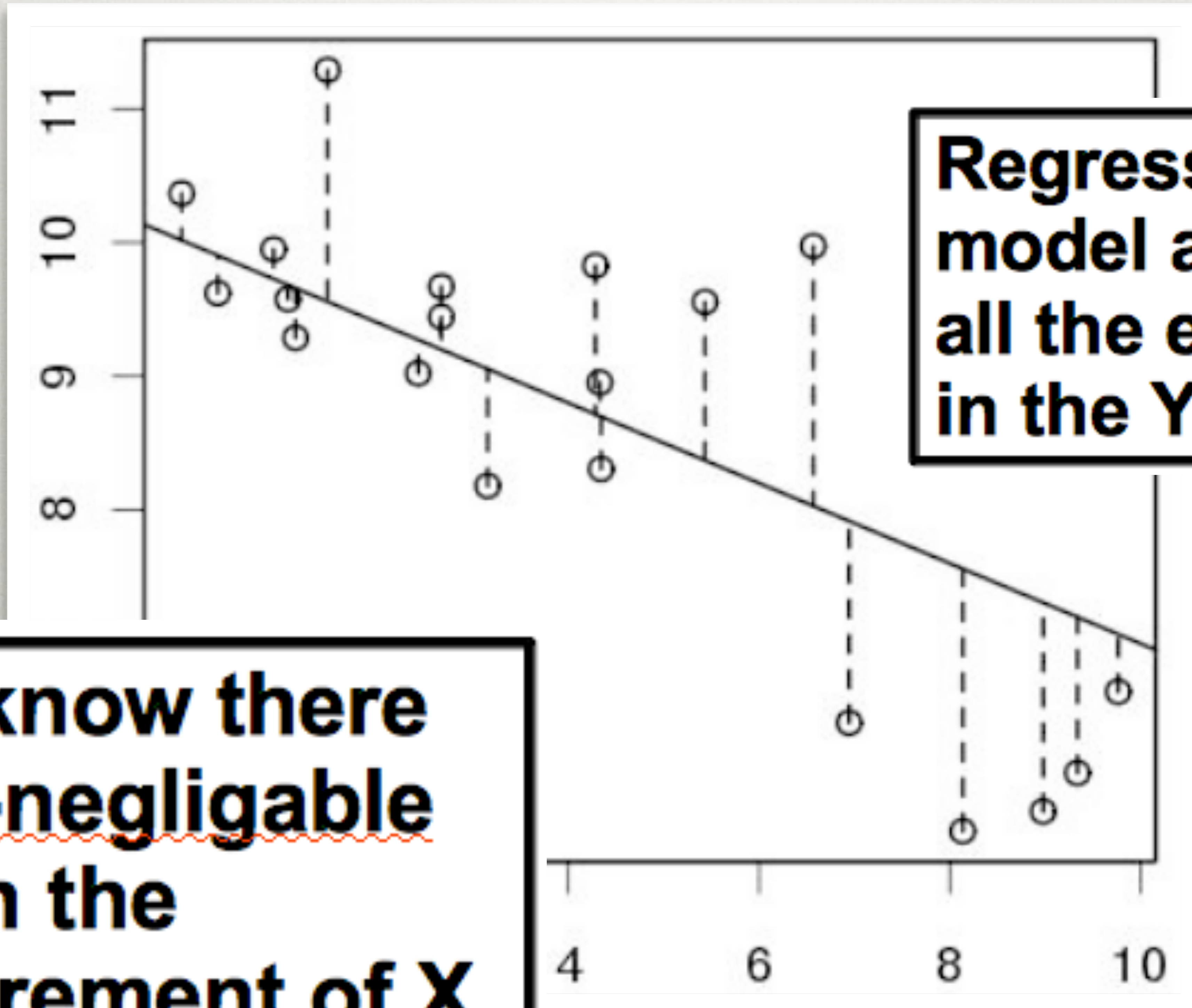
Errors in Variables



Errors in Variables



Errors in Variables

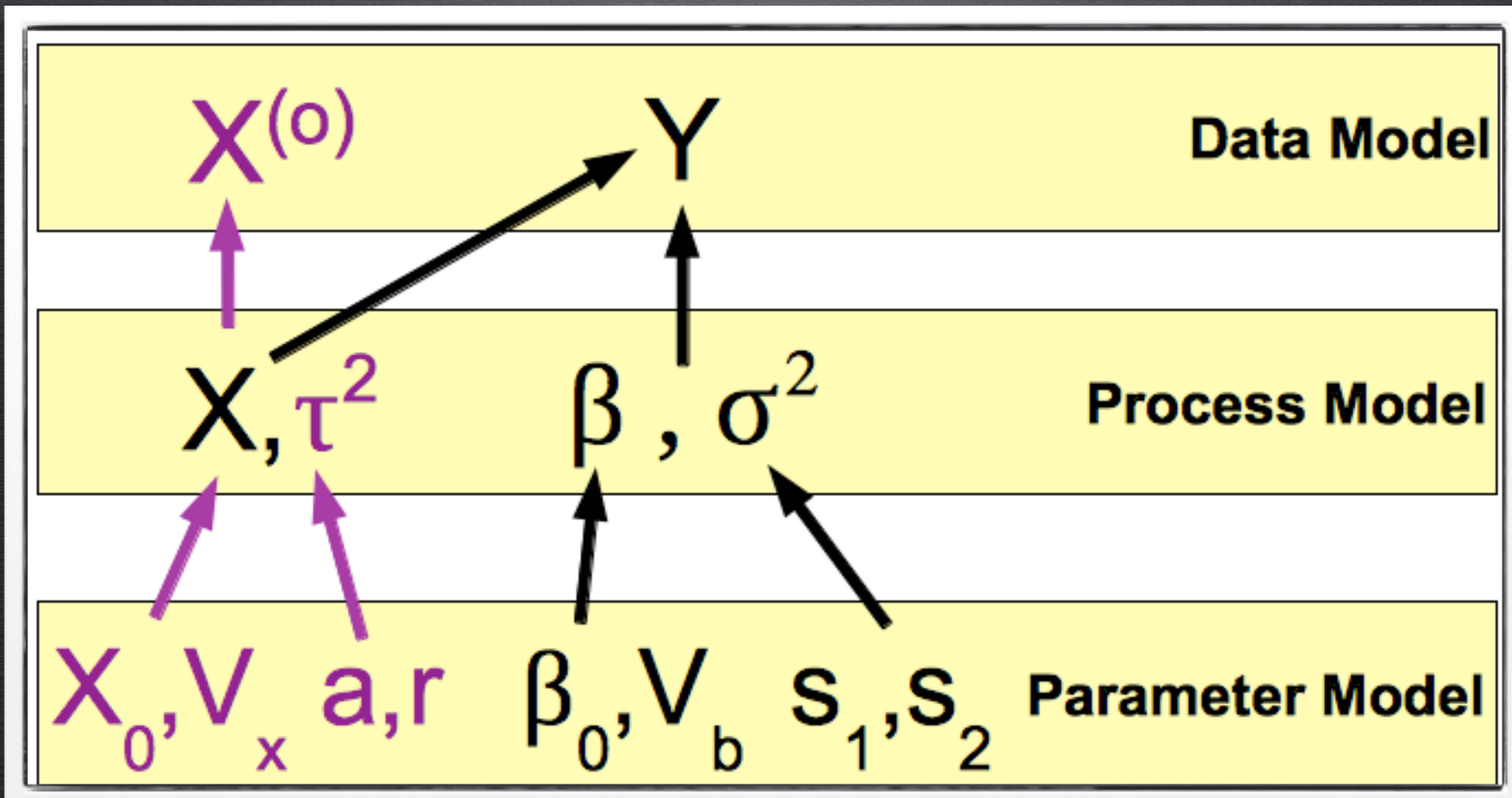


Regression model assumes all the error is in the Y

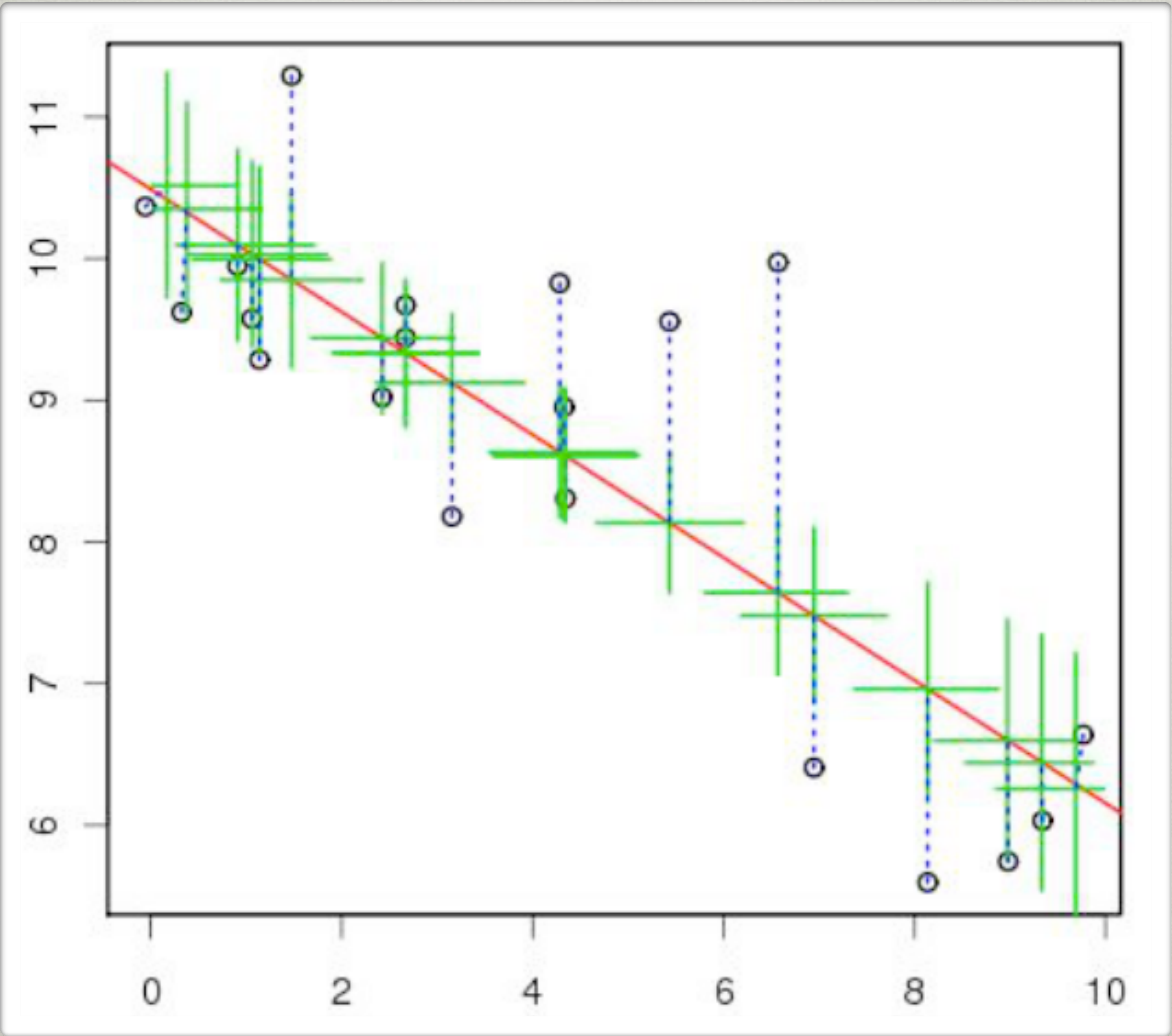
Often know there is non-negligable error in the measurement of X

$$\vec{y} \sim N(\boldsymbol{X}\vec{\beta}, \sigma^2)$$

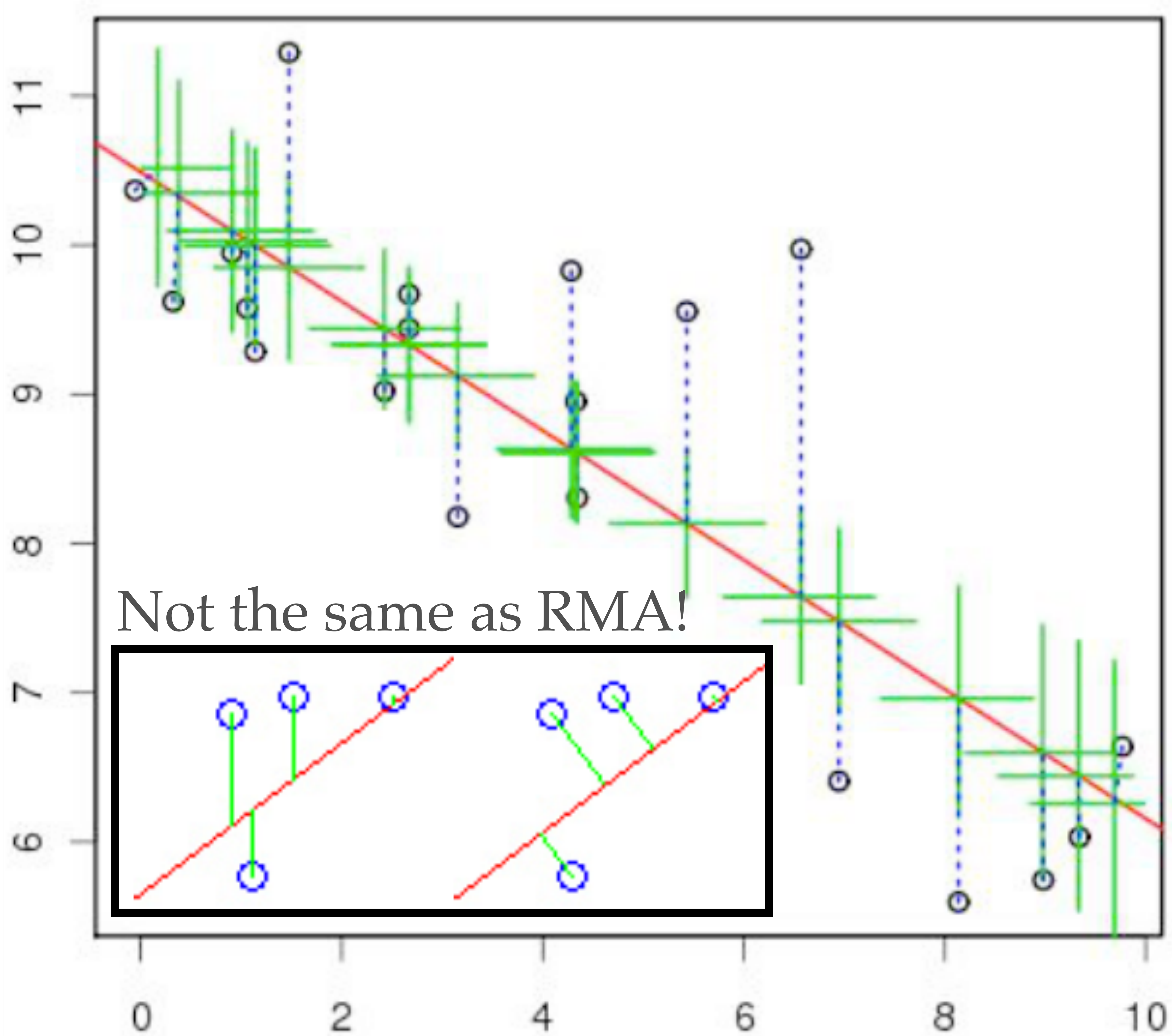
$$x^{(o)} \sim N(x, \tau^2)$$



```
model {
  ## priors
  for(i in 1:2) { beta[i] ~ dnorm(0,0.001)}
  sigma ~ dgamma(0.1,0.1)
  tau ~ dgamma(0.1,0.1)
  for(i in 1:n) { x[i] ~ dunif(0,10)}

  for(i in 1:n){
    xo[i] ~ dnorm(x[i],tau)
    mu[i] <- beta[1]+beta[2]*x[i]
    y[i] ~ dnorm(mu[i],sigma)
  }
}
```



11
10
9
8
7
6
0
2
4
6
8
10


Not the same as RMA!
11
10
9
8
7
6
0
2
4
6
8
10

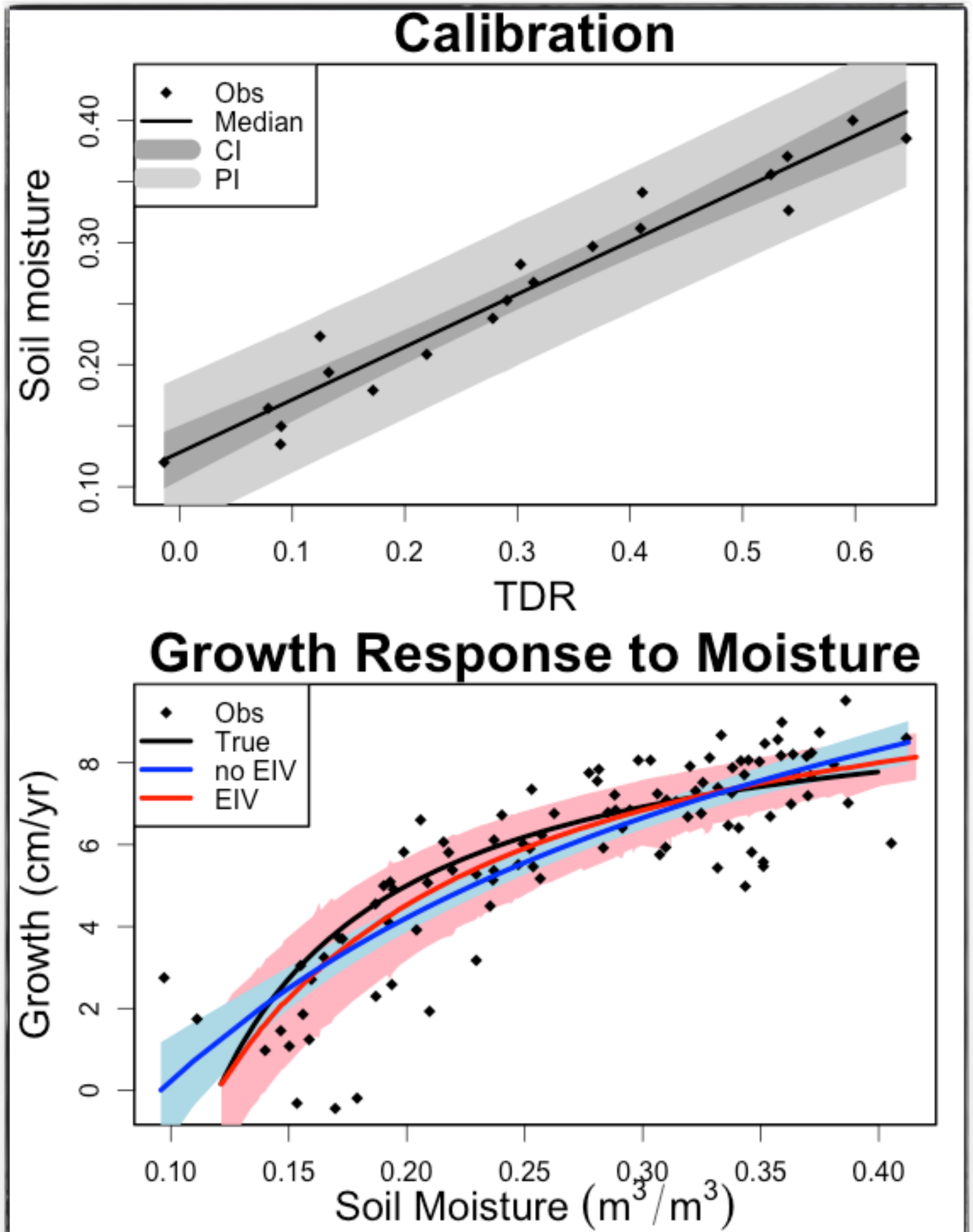
Additional Thoughts on EIV

$$x^{(o)} \sim g(x|\theta)$$

- Errors in X's need not be Normal
- Errors need not be additive
- Can account for known biases

$$x^{(o)} \sim N(\alpha_0 + \alpha_1 x, \tau^2)$$

- Observed data can be a different type (proxy)
  - e.g. calibration curves
- Very useful to have informative priors

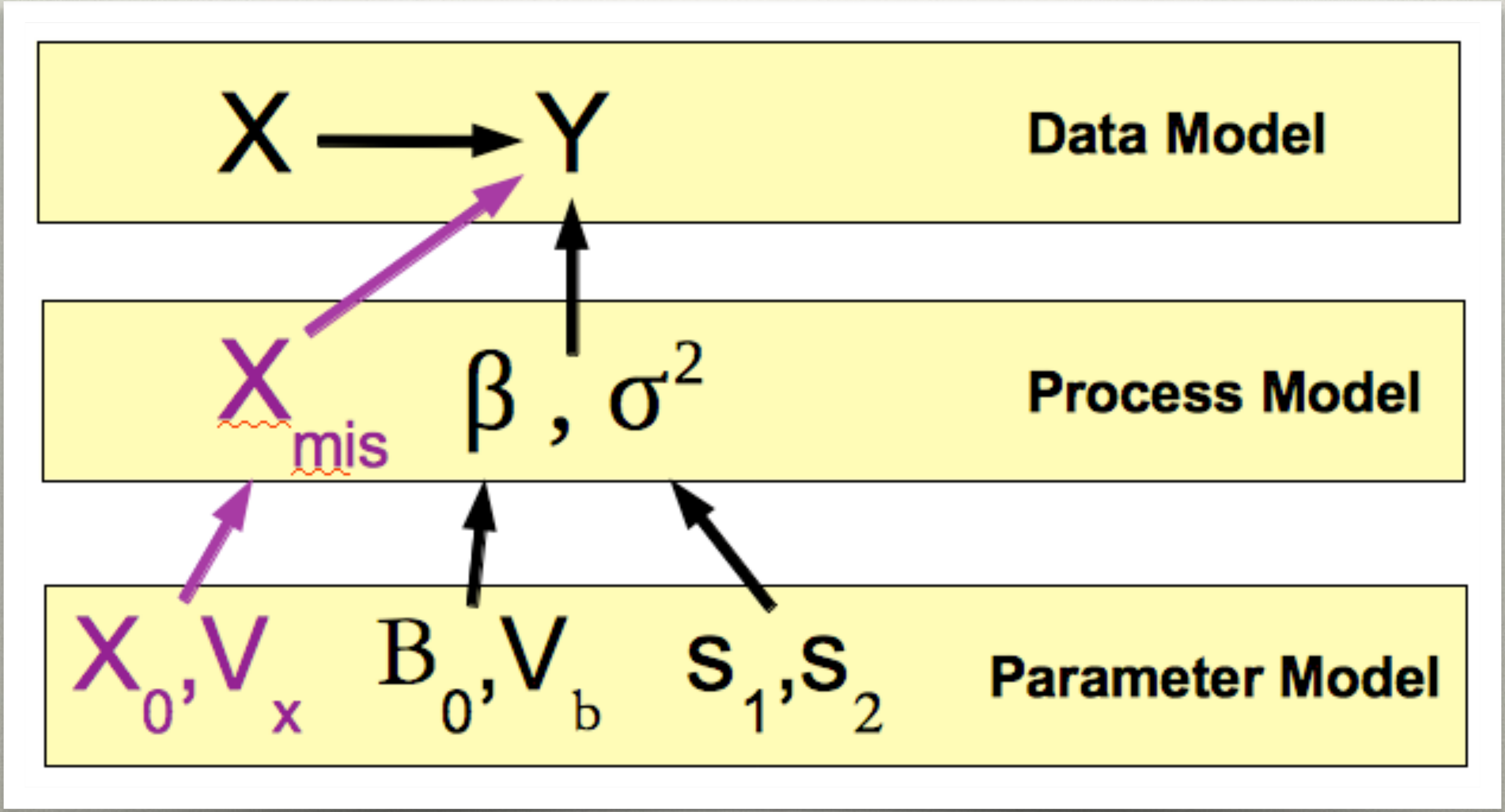

Calibration
Obs
Median
CI
PI
Soil moisture
0.10
0.20
0.30
0.40
0.0
0.1
0.2
0.3
0.4
0.5
0.6
TDR
Growth Response to Moisture
Obs
True
no EIV
EIV
Growth (cm/yr)
0
2
4
6
8
0.10
0.15
0.20
0.25
0.30
0.35
0.40
Soil Moisture (m3/m3)

Latent Variables

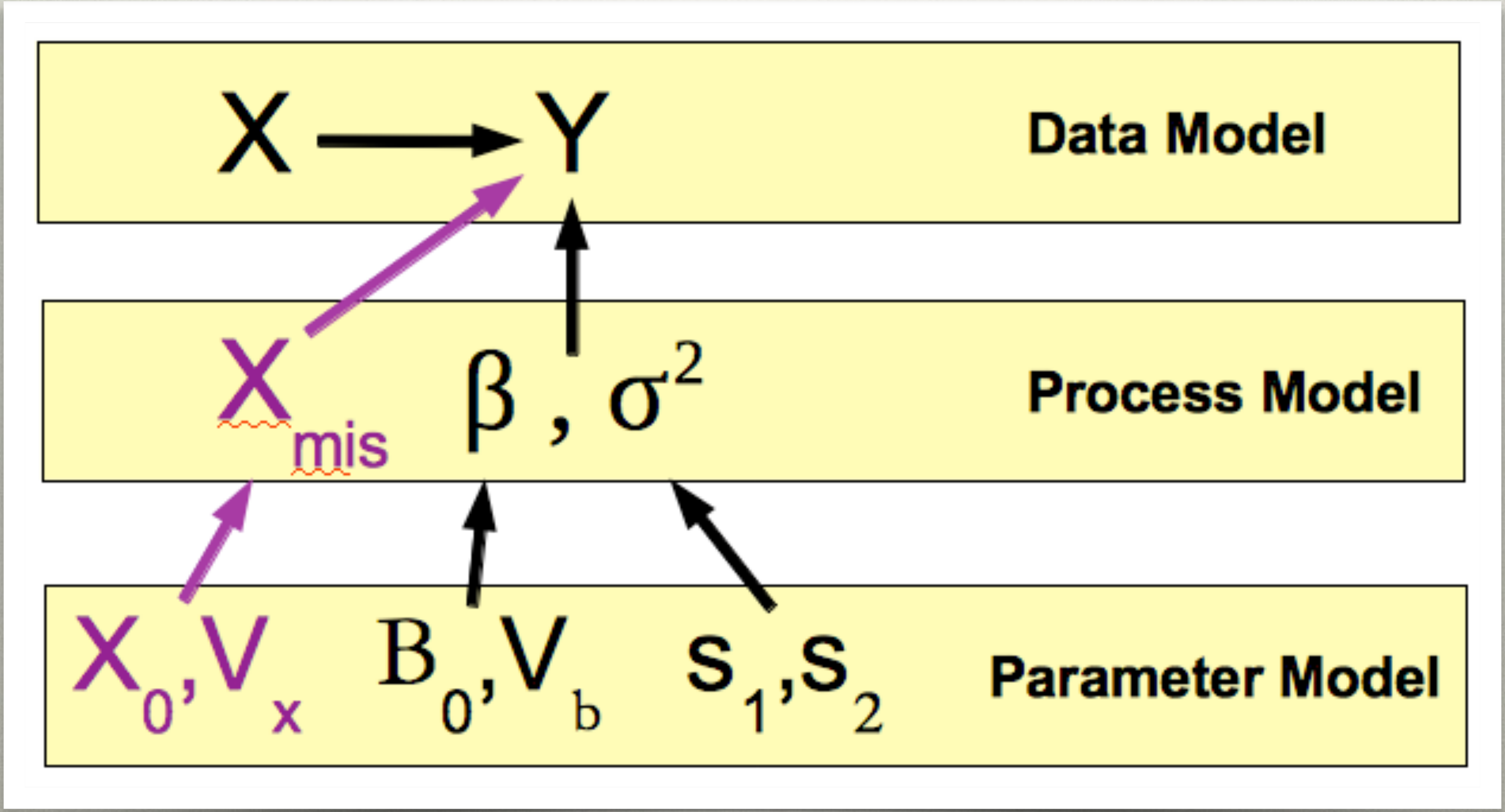
- Variables that are not directly observed
- Values are inferred from model
  - Parameter model: prior on value
  - Data and Process models provide constraint

- MCMC integrates over (by sampling) the values the unobserved variable could take on
- Contribute to uncertainty in parameters, model
- Ignoring this variability can lead to falsely <u>overconfident</u> conclusions

Missing Data



Missing Data



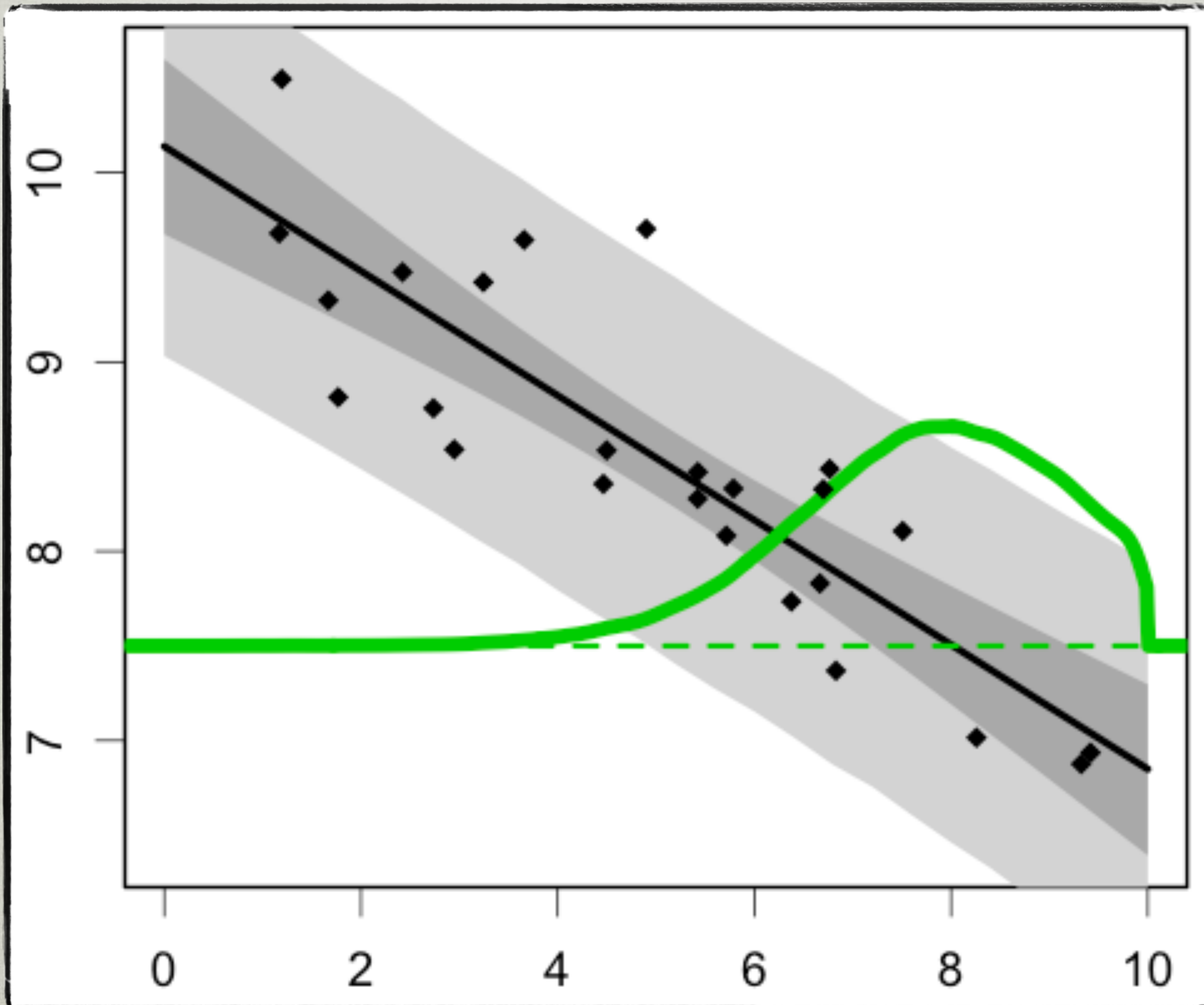
Data needs to be **Missing At Random!!**

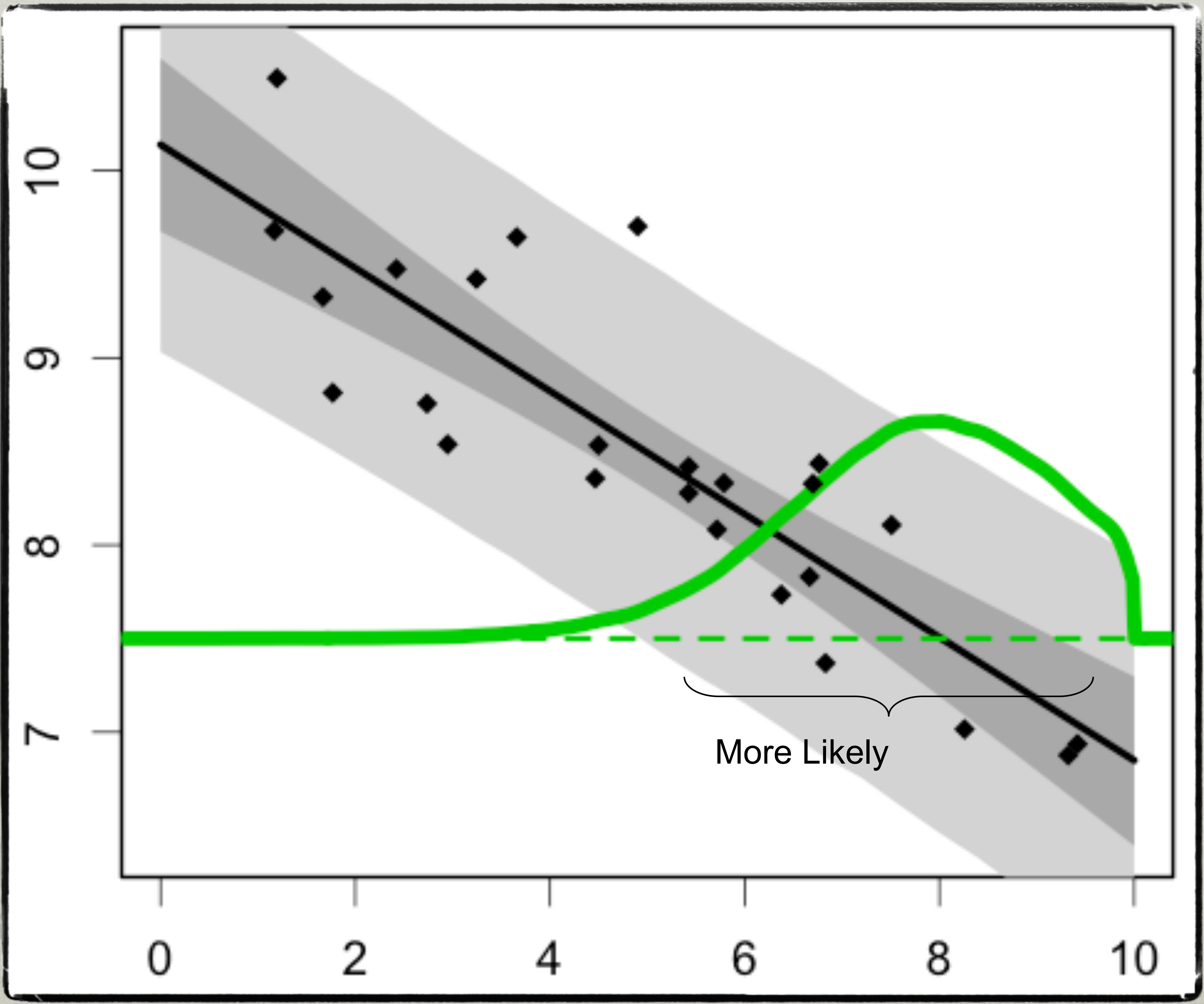
JAGS example: Simple Regression

```
model{
  ## priors
  for(i in 1:2) { beta[i] ~ dnorm(0,0.001)}
  sigma ~ dgamma(0.1,0.1)
  for(i in mis) { x[i] ~ dunif(0,10)}

  for(i in 1:n){
    mu[i] <- beta[1]+beta[2]*x[i]
    y[i] ~ dnorm(mu[i],sigma)
  }
}
```

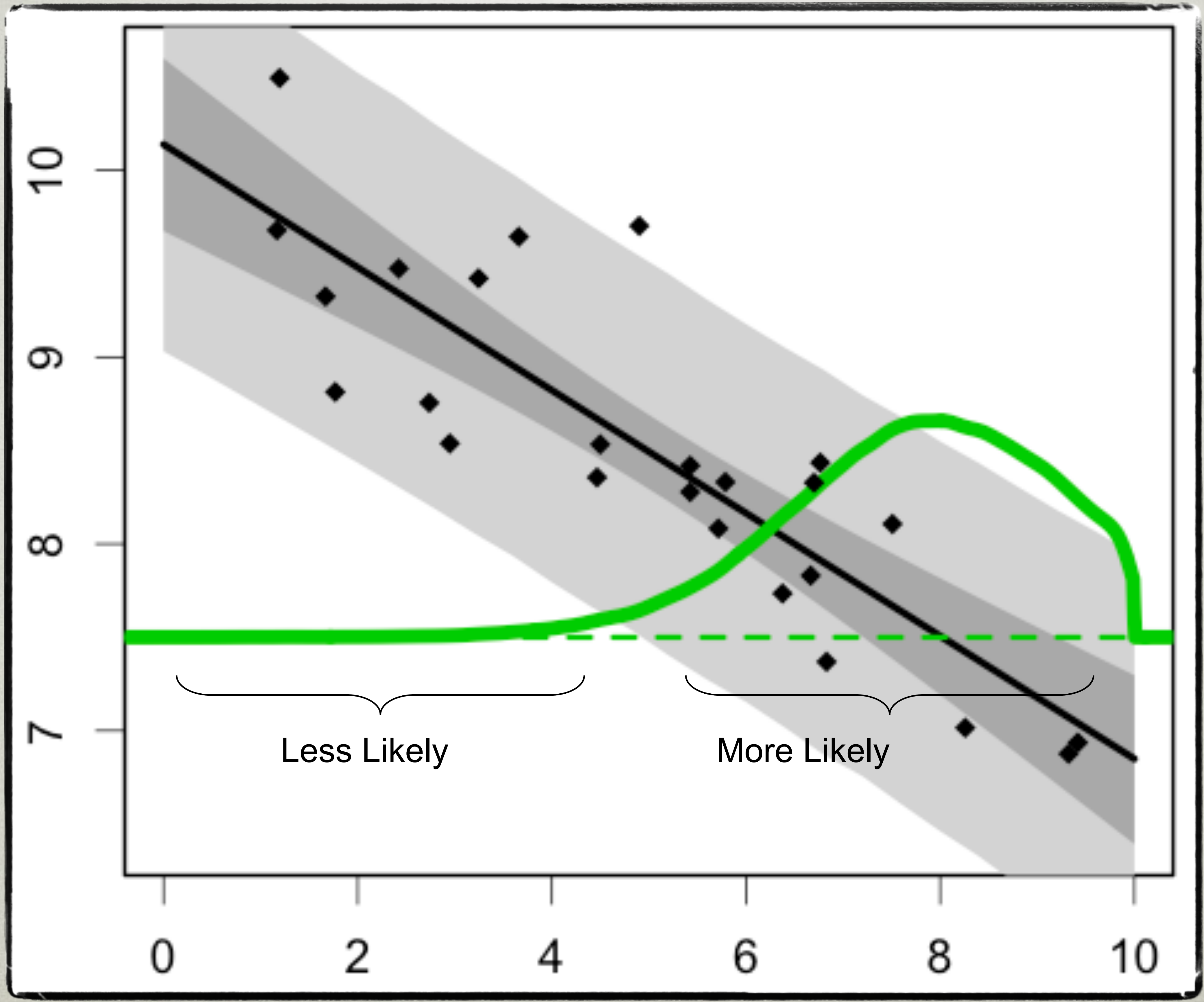
Vector giving indices of missing values

	X	Y
	4.68	8.46
	2.95	8.55
	9.09	7.01
	8.15	9.06
	1.76	11.38
	4.23	9.12
	7.73	7.3
	2.43	8.02
	6.46	8.45
	4.06	8.95
	2.42	9.62
	0.6	9.15
	8.17	7.51
	0.22	10.8
	4.93	9.78
	2.99	10.71
	8.36	8.89
	6.4	8.21
	8.17	6.22
	6.46	5.4
	1.82	10.05
	9.52	7.96
	2.44	9.63
	6.84	7.05
	7.42	8.73
$\text{mis} = 26$	NA	7.5

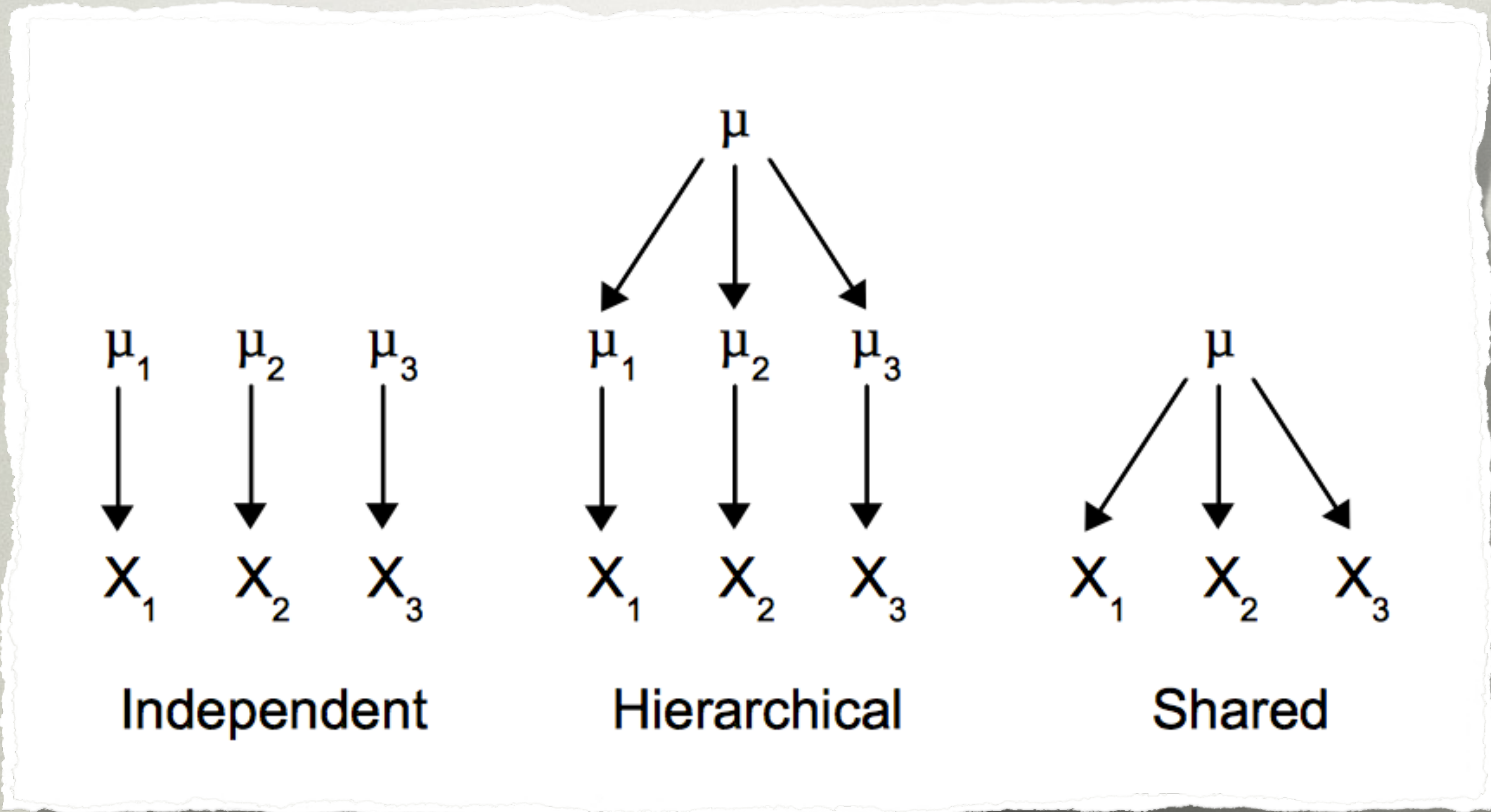

10
9
8
7
0
2
4
6
8
10

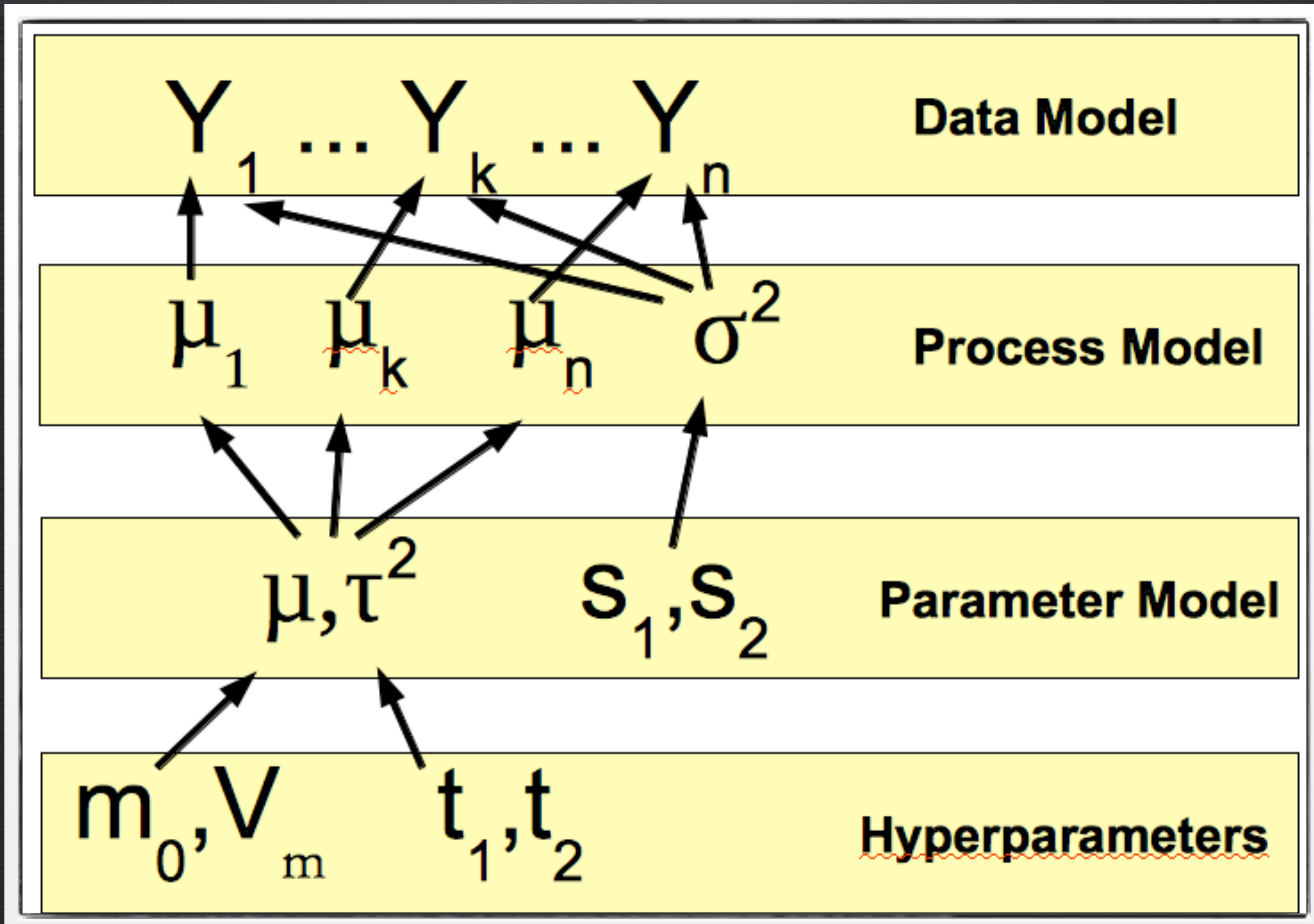


10
9
8
7
0
2
4
6
8
10
More Likely

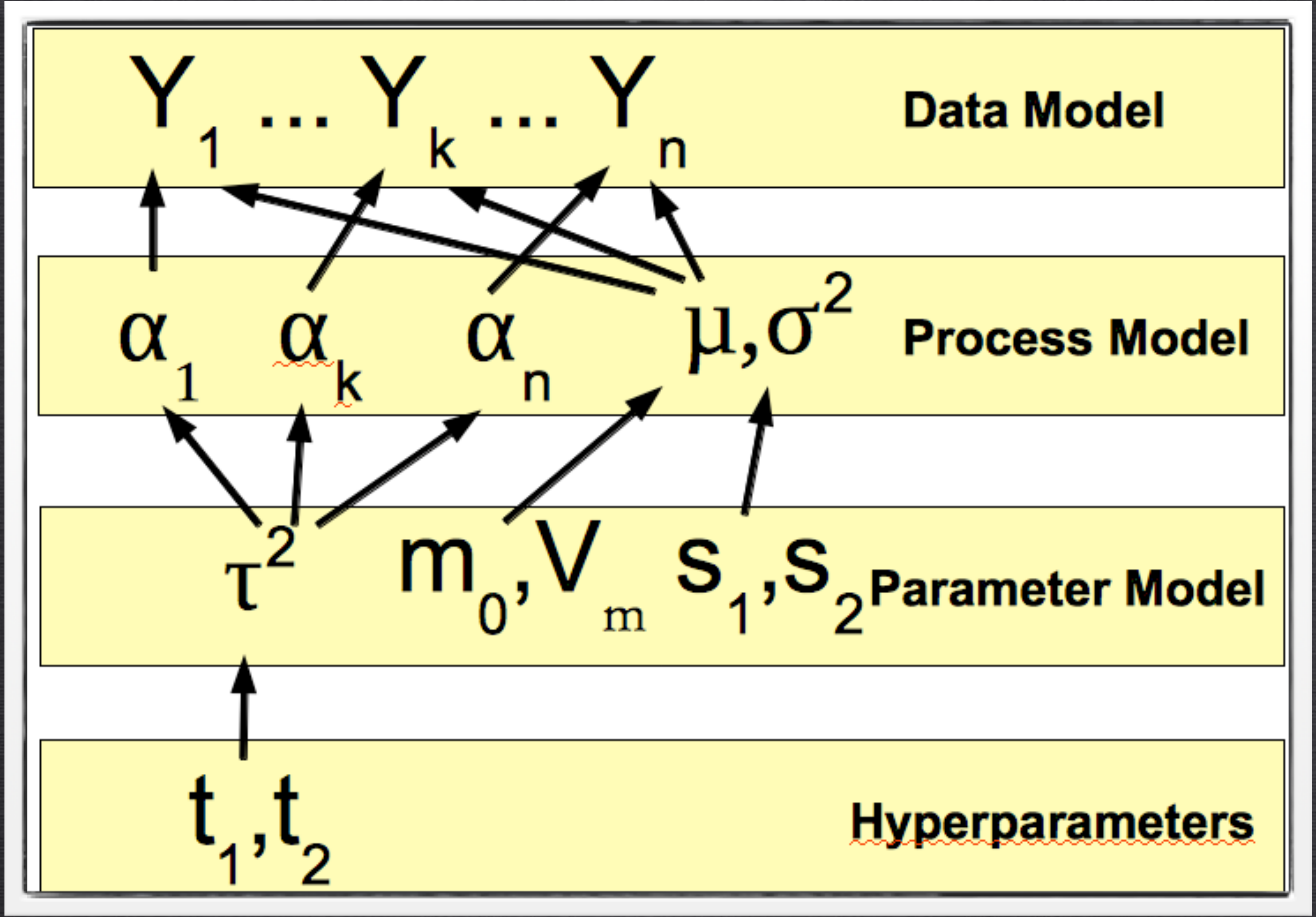

10
9
8
7
Less Likely
More Likely
0
2
4
6
8
10

Hierarchical Models



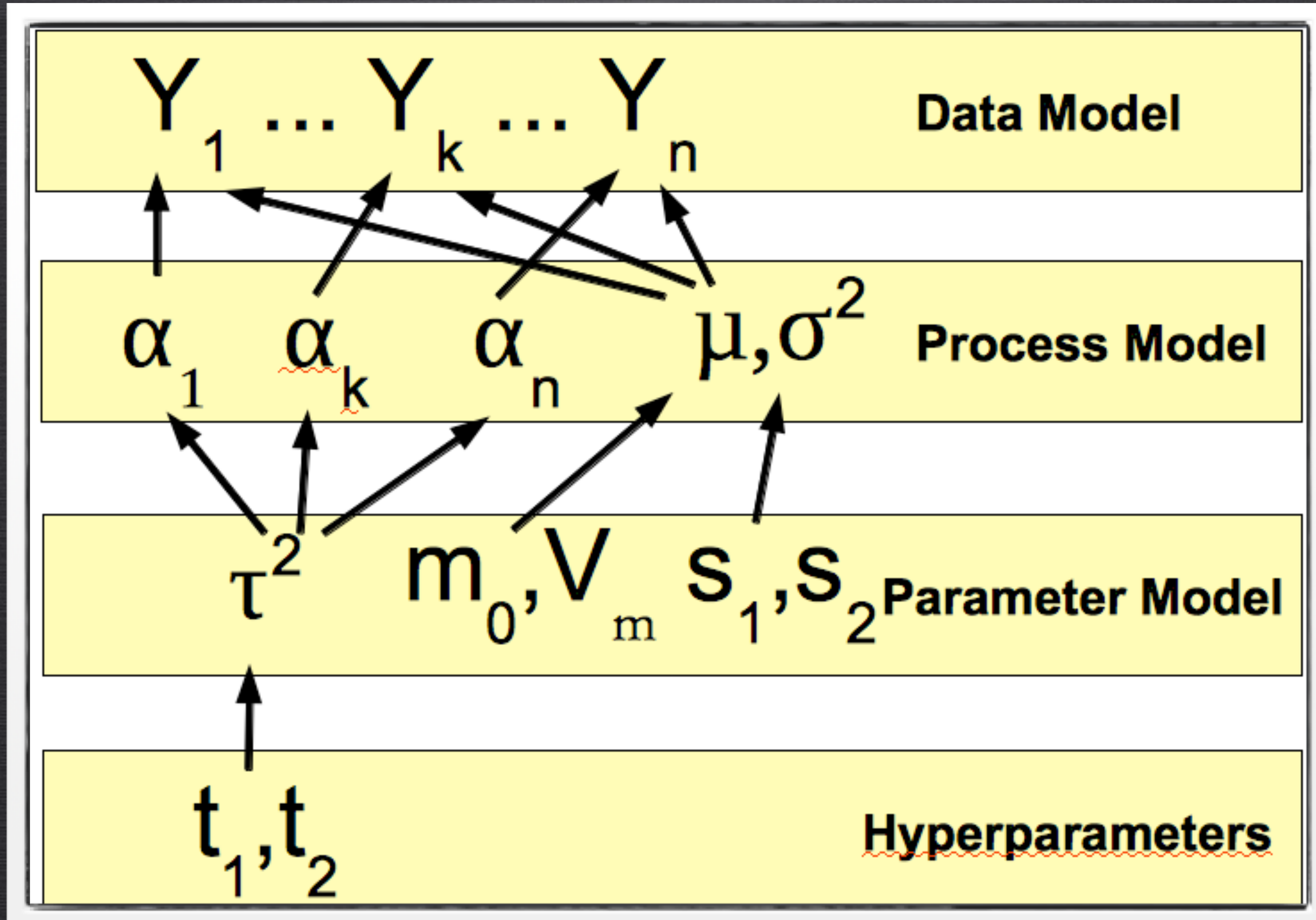


Y_1 ... Y_k ... Y_n
Data Model
μ_1 μ_k μ_n σ^2
Process Model
μ, τ^2
s_1, s_2
Parameter Model
m_0, V_m
t_1, t_2
Hyperparameters



$$Y_k \sim N(\mu_g + \alpha_k, \sigma^2)$$

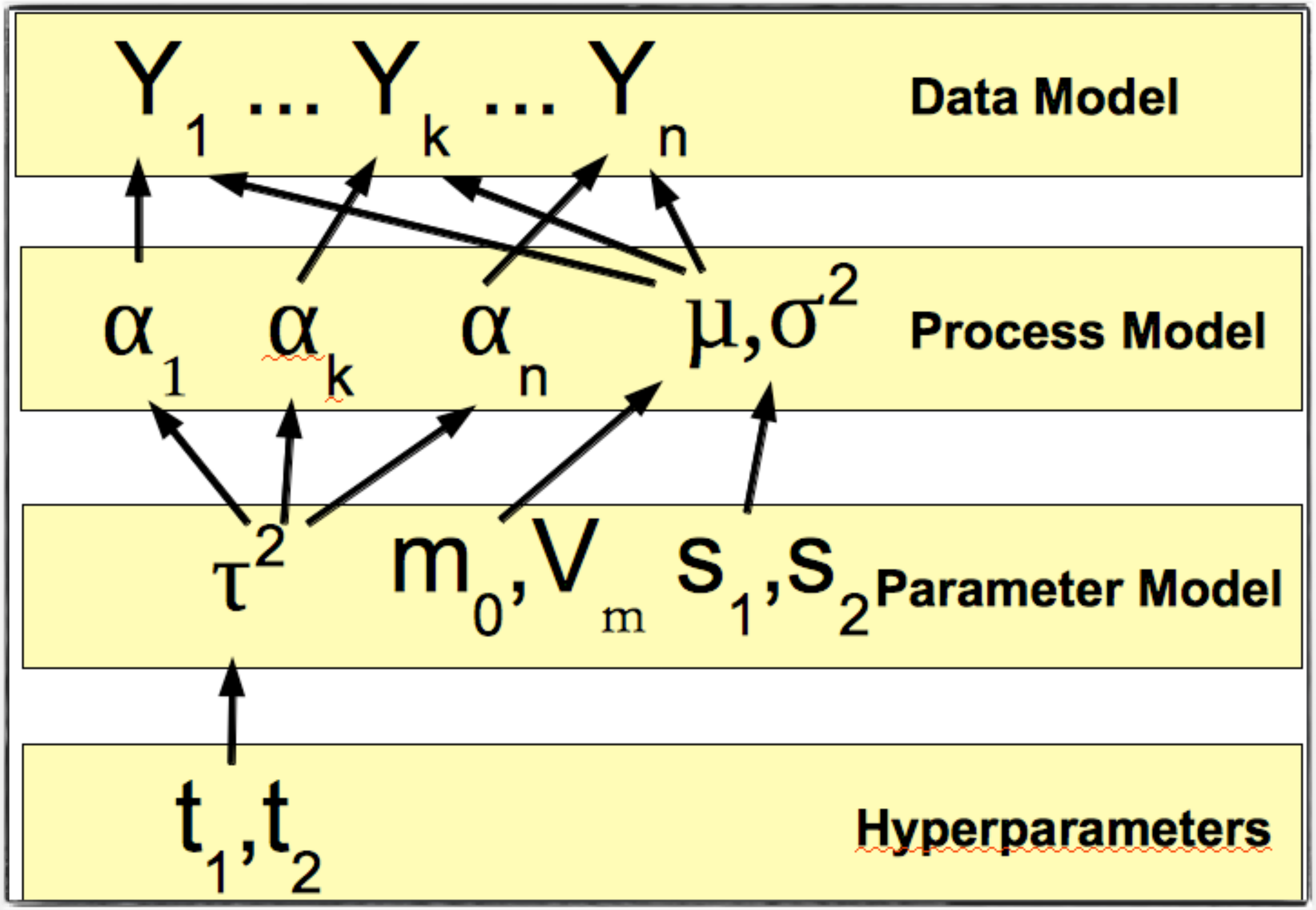
$$\alpha_k \sim N(0, \tau^2)$$



Random Effects

$$Y_k \sim N(\mu_g + \alpha_k, \sigma^2)$$

$$\alpha_k \sim N(0, \tau^2)$$



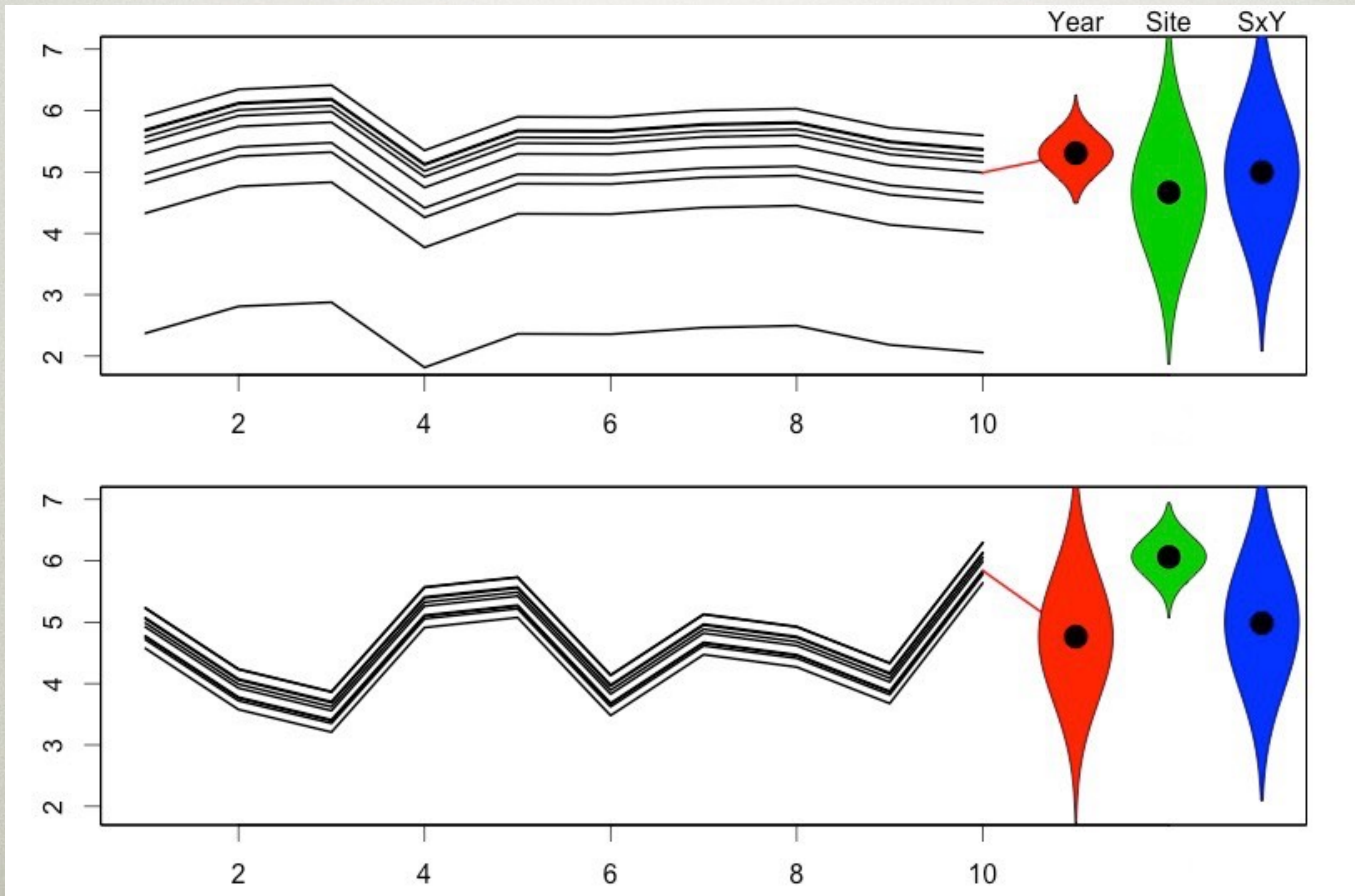
$$Y_k \sim N(\mu_g + \alpha_k, \sigma^2)$$
$$\alpha_k \sim N(0, \tau^2)$$

Random Effects

Notes:

* Can put variability on any param
* Need not be independent

Impacts on Inference



Explaining unexplained variance

- Random effects attempt to account for the unexplained variance associated with some group (plot, year, etc.) due to all the things that were not measured
- May point to scales that need additional explanation
- Adding covariates may explain some portion of this variance, but there's always something you didn't measure
- Sometimes additional fixed effects not justified (model selection)

STATE SPACE MODELS

(AKA HIDDEN MARKOV)

DYNAMIC MODELS

- Future state depends on the current state

$$X_{t+1} = f(X_t, Z_t \mid \theta) + \epsilon$$

DYNAMIC MODELS

- Future state depends on the current state

$$X_{t+1} = f(X_t, Z_t \mid \theta) + \epsilon$$

state
variable

DYNAMIC MODELS

- Future state depends on the current state

$$X_{t+1} = f(X_t, Z_t \mid \theta) + \epsilon$$

state variable

covariates / drivers

DYNAMIC MODELS

- Future state depends on the current state

$$X_{t+1} = f(X_t, Z_t \mid \theta) + \epsilon$$

state
variable

covariates / drivers

parameters

DYNAMIC MODELS

- Future state depends on the current state

$$X_{t+1} = f(X_t, Z_t \mid \theta) + \epsilon$$

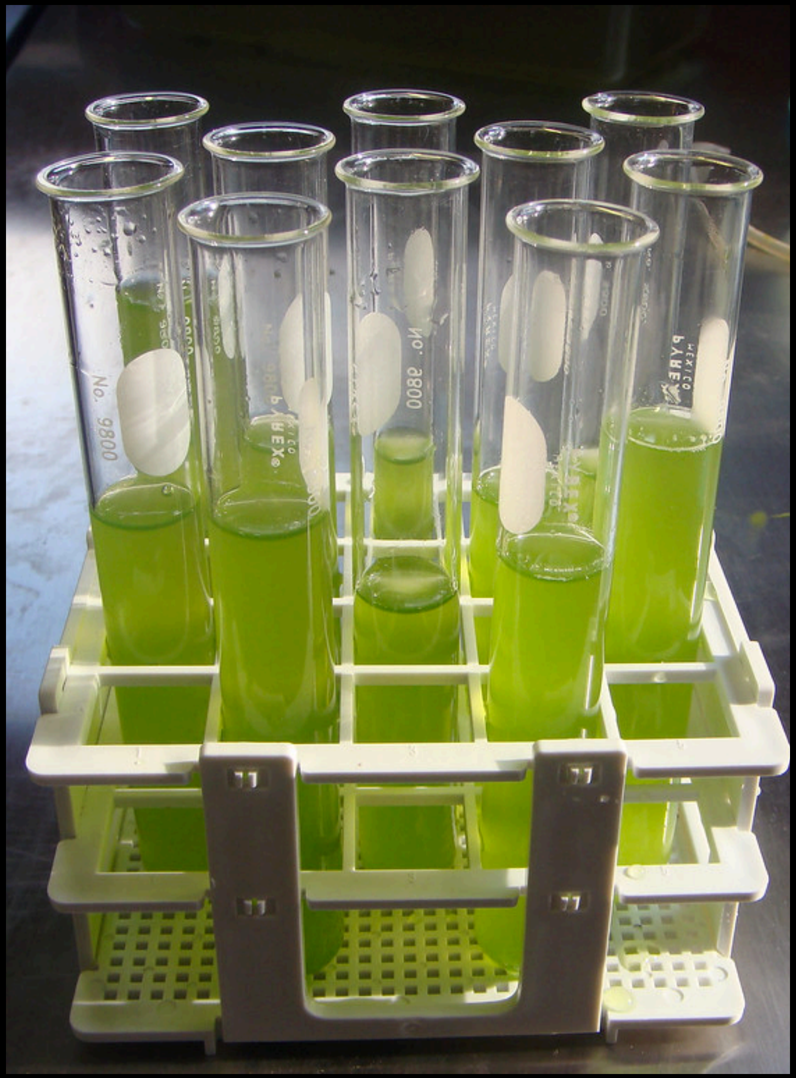
state variable

covariates / drivers

parameters

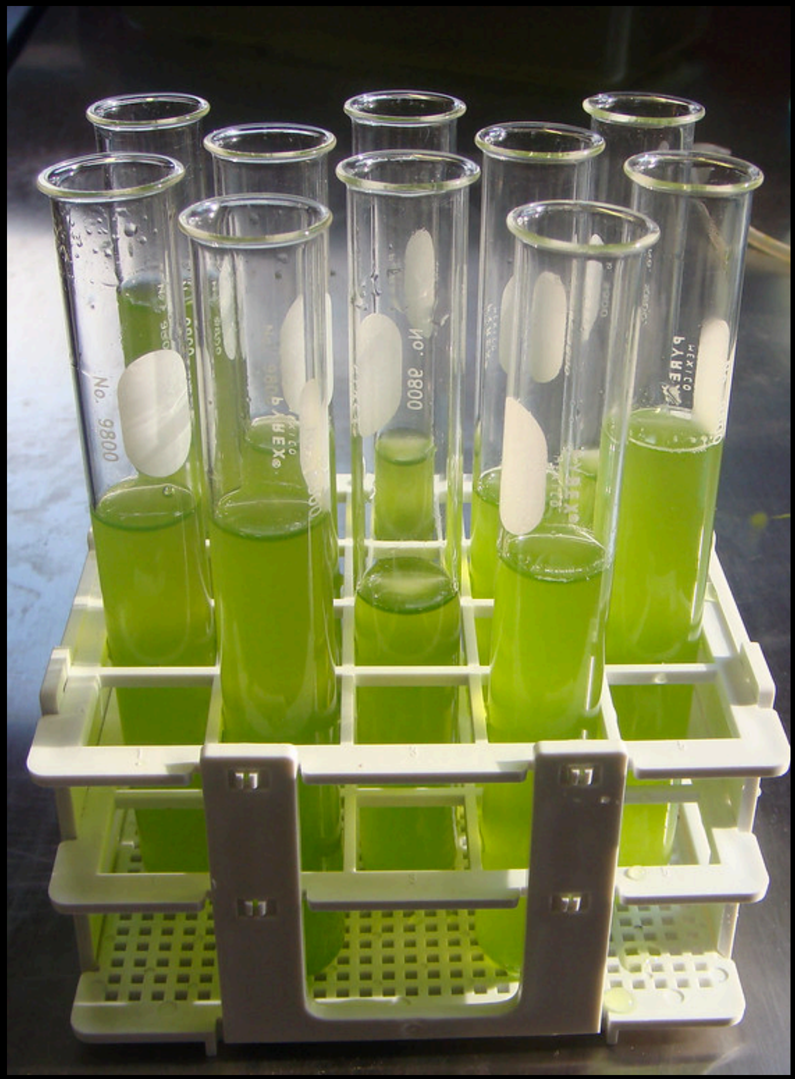
process error

OBSERVATION ERROR



State space models are for when you observe this Y_t

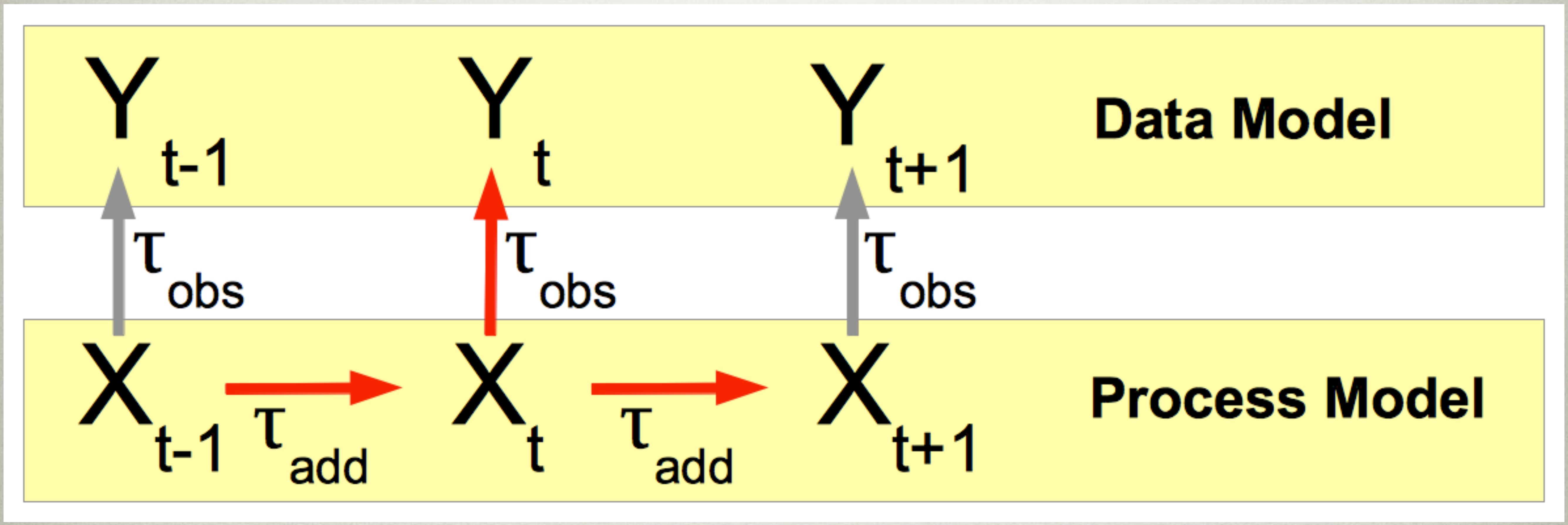
OBSERVATION ERROR



State space models are for when you observe this Y_t

But want to model this X_t

State Space



$Y_t = g(X_t | \phi)$ **Data Model**

$X_t = f(X_{t-1} | \theta)$ **Process Model**

Random Walk

$$X_t \sim N(X_{t-1}, \tau^2_{add})$$

$$Y_t \sim N(X_t, \tau^2_{obs})$$

$$\tau^2_{obs} \sim IG(a_{obs}, r_{obs})$$

$$\tau^2_{add} \sim IG(a_{add}, r_{add})$$

$$X_0 \sim N(X_{ic}, \tau_{IC})$$

Random Walk

- What is the conditional distribution of X?

$$X_t \sim N\left(X_{t-1}, \tau^2_{add}\right)$$
$$Y_t \sim N\left(X_t, \tau^2_{obs}\right)$$
$$\tau^2_{obs} \sim IG\left(a_{obs}, r_{obs}\right)$$
$$\tau^2_{add} \sim IG\left(a_{add}, r_{add}\right)$$
$$X_0 \sim N\left(X_{ic}, \tau_{IC}\right)$$

Random Walk

- What is the conditional distribution of X?

$$
\begin{aligned}
X_t \sim\, & N(X_t \mid X_{t-1}, \tau^2_{add}) \times \\
& N(X_{t+1} \mid X_t, \tau^2_{add}) \times \\
& N(Y_t \mid X_t, \tau^2_{obs})
\end{aligned}
$$

$$
\begin{aligned}
X_t &\sim N(X_{t-1}, \tau^2_{add}) \\
Y_t &\sim N(X_t, \tau^2_{obs}) \\
\tau^2_{obs} &\sim IG(a_{obs}, r_{obs}) \\
\tau^2_{add} &\sim IG(a_{add}, r_{add}) \\
X_0 &\sim N(X_{ic}, \tau_{IC})
\end{aligned}
$$

Random Walk

- What is the conditional distribution of X?

$$X_t \sim N(X_t|X_{t-1}, \tau^2_{add}) \times$$
$$N(X_{t+1}|X_t, \tau^2_{add}) \times$$
$$N(Y_t|X_t, \tau^2_{obs})$$

- Special Cases

$$X_t \sim N(X_{t-1}, \tau^2_{add})$$
$$Y_t \sim N(X_t, \tau^2_{obs})$$
$$\tau^2_{obs} \sim IG(a_{obs}, r_{obs})$$
$$\tau^2_{add} \sim IG(a_{add}, r_{add})$$
$$X_0 \sim N(X_{ic}, \tau_{IC})$$

Random Walk

- What is the conditional distribution of X?

$$
\begin{aligned}
X_t \sim\ & N(X_t | X_{t-1}, \tau^2_{add}) \times \\
& N(X_{t+1} | X_t, \tau^2_{add}) \times \\
& N(Y_t | X_t, \tau^2_{obs})
\end{aligned}
$$

$$
\begin{aligned}
X_t &\sim N(X_{t-1}, \tau^2_{add}) \\
Y_t &\sim N(X_t, \tau^2_{obs}) \\
\tau^2_{obs} &\sim IG(a_{obs}, r_{obs}) \\
\tau^2_{add} &\sim IG(a_{add}, r_{add}) \\
X_0 &\sim N(X_{ic}, \tau_{IC})
\end{aligned}
$$

- Special Cases
  - First

Random Walk

- What is the conditional distribution of X?

$$X_t \sim N(X_t | X_{t-1}, \tau^2_{add}) \times$$
$$N(X_{t+1} | X_t, \tau^2_{add}) \times$$
$$N(Y_t | X_t, \tau^2_{obs})$$

$$X_t \sim N(X_{t-1}, \tau^2_{add})$$
$$Y_t \sim N(X_t, \tau^2_{obs})$$
$$\tau^2_{obs} \sim IG(a_{obs}, r_{obs})$$
$$\tau^2_{add} \sim IG(a_{add}, r_{add})$$
$$X_0 \sim N(X_{ic}, \tau_{IC})$$

- Special Cases
    - First
    - Last

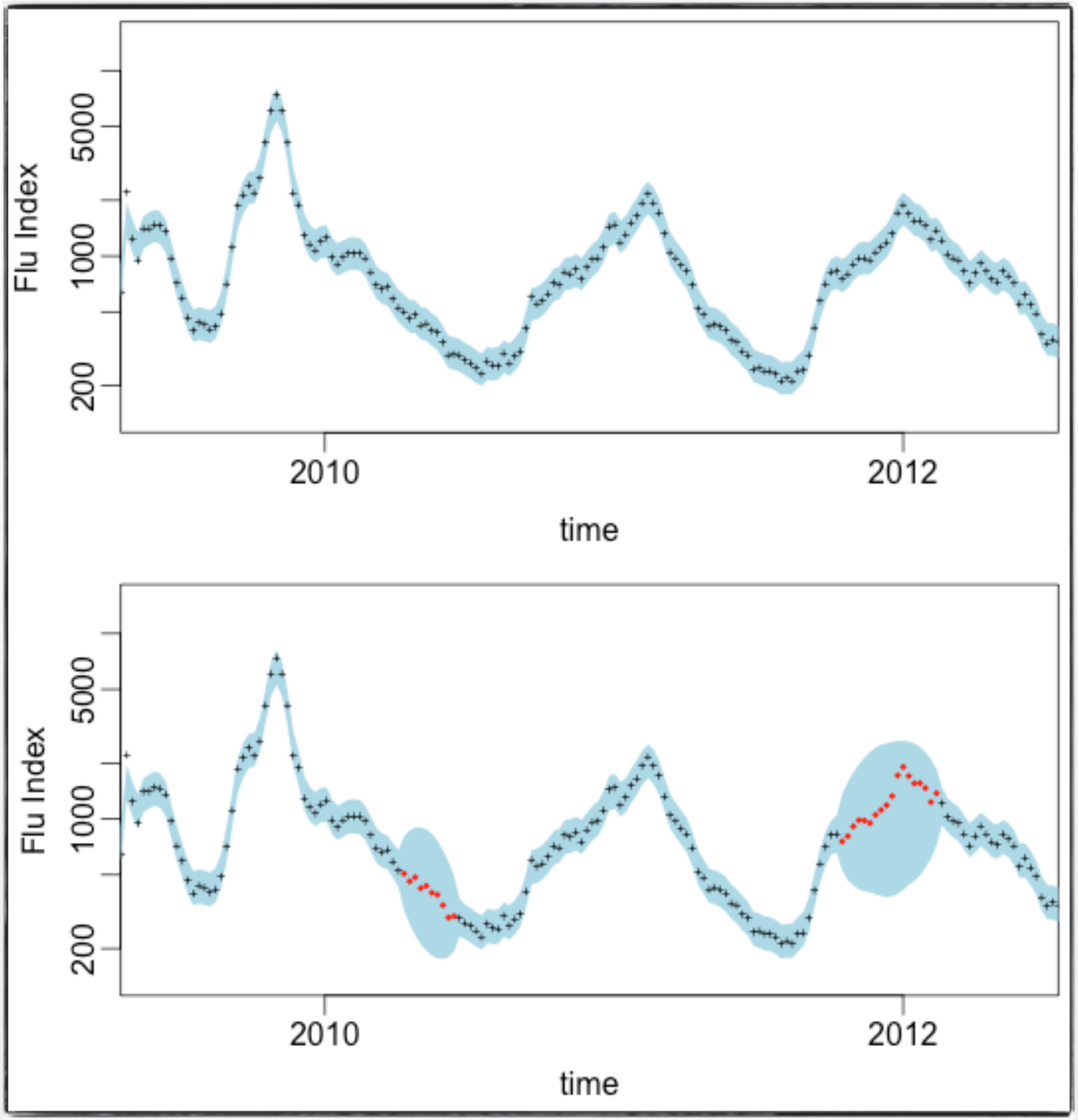
Random Walk

- What is the conditional distribution of X?

$$X_t \sim N(X_t | X_{t-1}, \tau^2_{add}) \times$$
$$N(X_{t+1} | X_t, \tau^2_{add}) \times$$
$$N(Y_t | X_t, \tau^2_{obs})$$

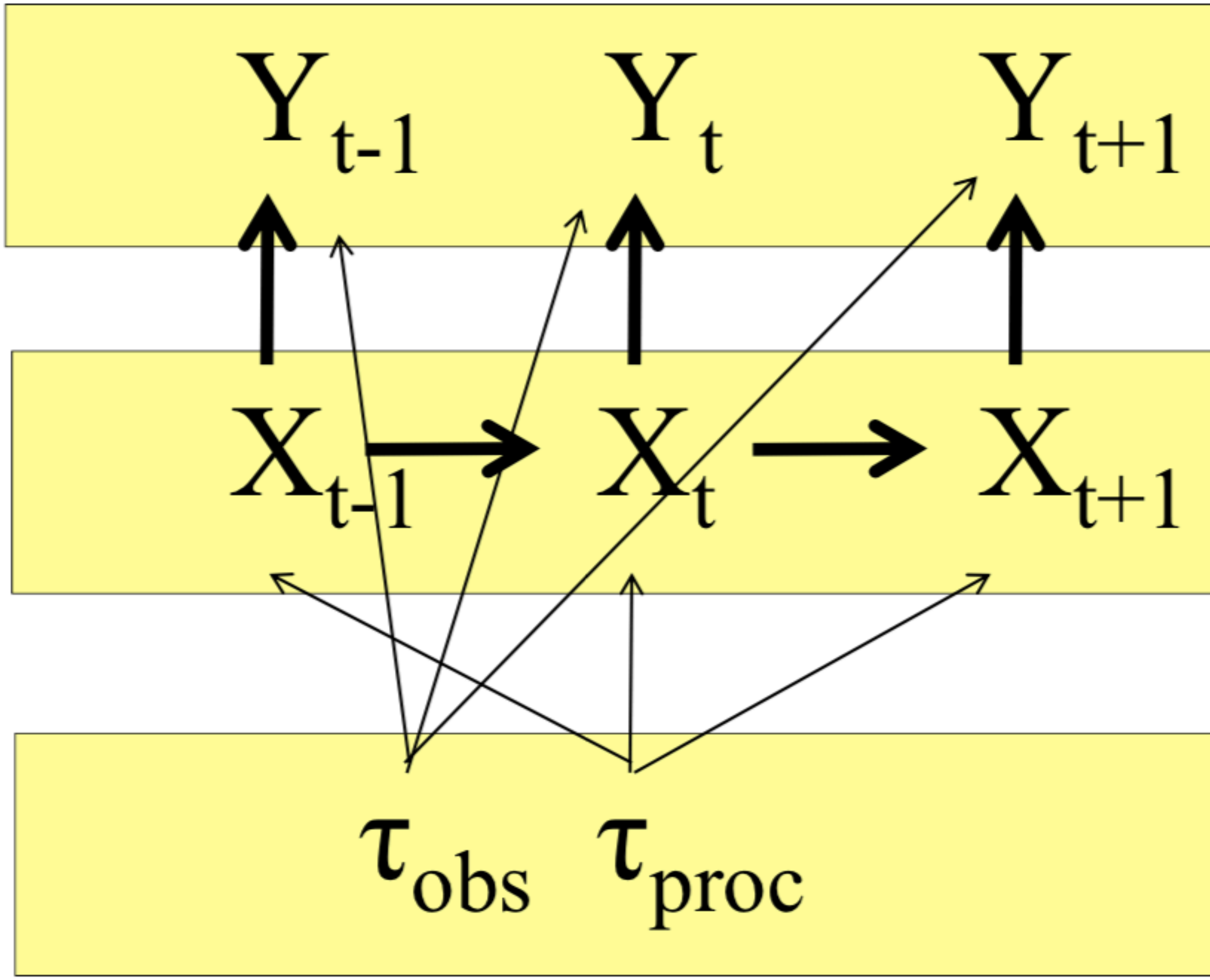
$$X_t \sim N(X_{t-1}, \tau^2_{add})$$
$$Y_t \sim N(X_t, \tau^2_{obs})$$
$$\tau^2_{obs} \sim IG(a_{obs}, r_{obs})$$
$$\tau^2_{add} \sim IG(a_{add}, r_{add})$$
$$X_0 \sim N(X_{ic}, \tau_{IC})$$

- Special Cases
  - First
  - Last
  - Missing Y

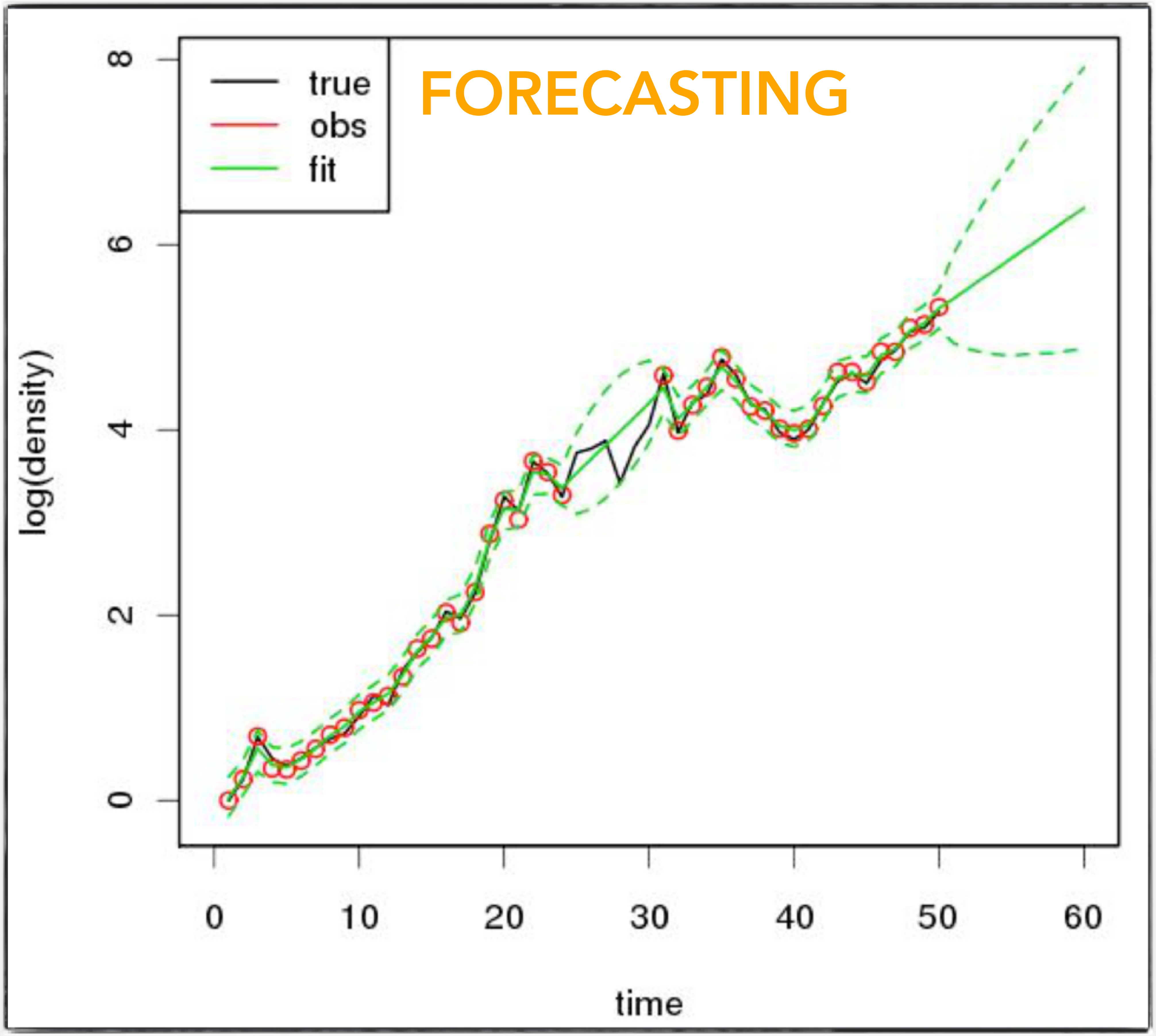


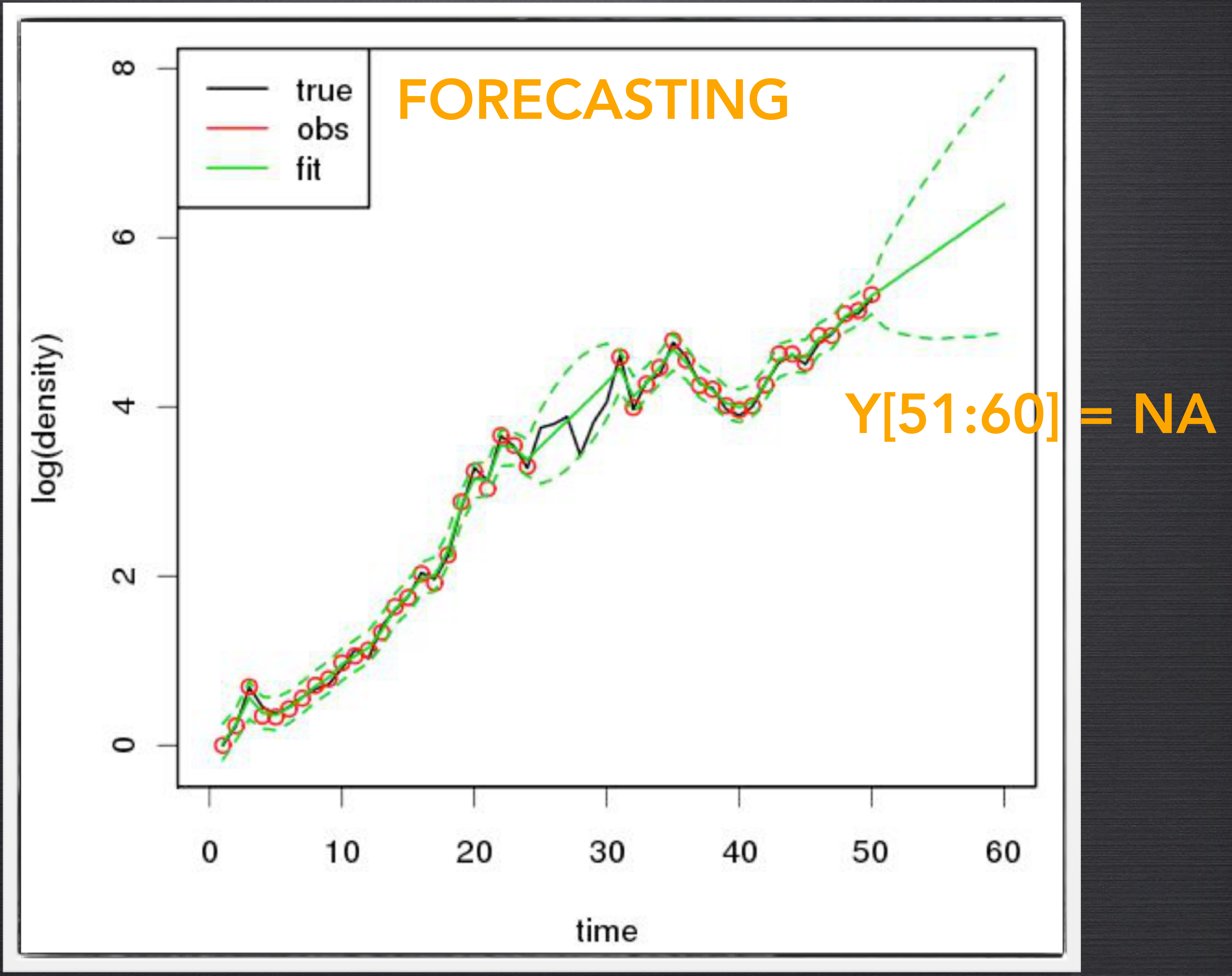
Flu Index
5000
1000
200
2010
2012
time
Flu Index
5000
1000
200
2010
2012
time

```
RandomWalk = "
model{

  #### Data Model
  for(i in 1:n){
    y[i] ~ dnorm(x[i],tau_obs)
  }

  #### Process Model
  for(i in 2:n){
    x[i]~dnorm(x[i-1],tau_proc)
  }

  #### Priors
  x[1] ~ dnorm(x_ic,tau_ic)
  tau_obs ~ dgamma(a_obs,r_obs)
  tau_proc ~ dgamma(a_proc,r_proc)
}
```

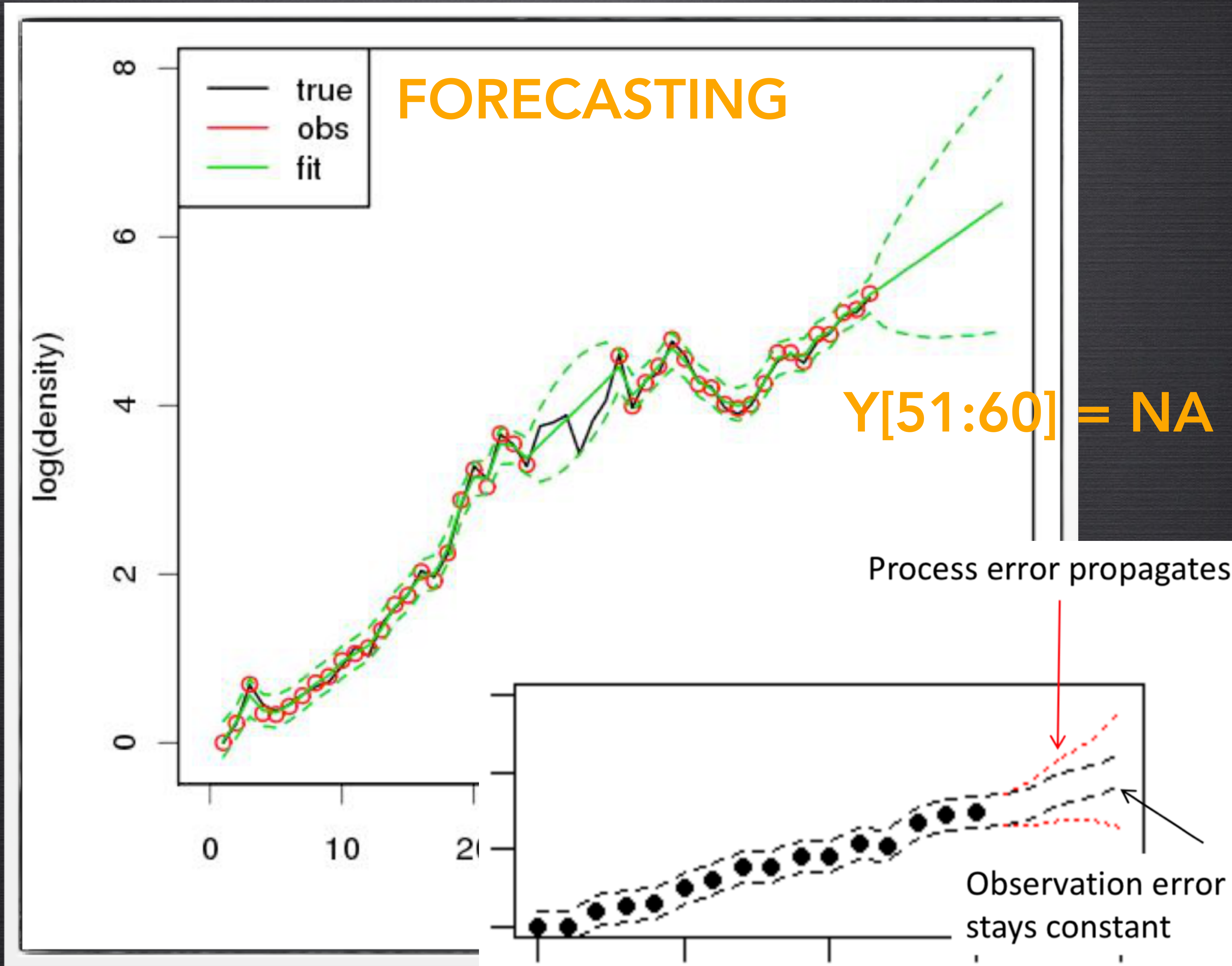


FORECASTING

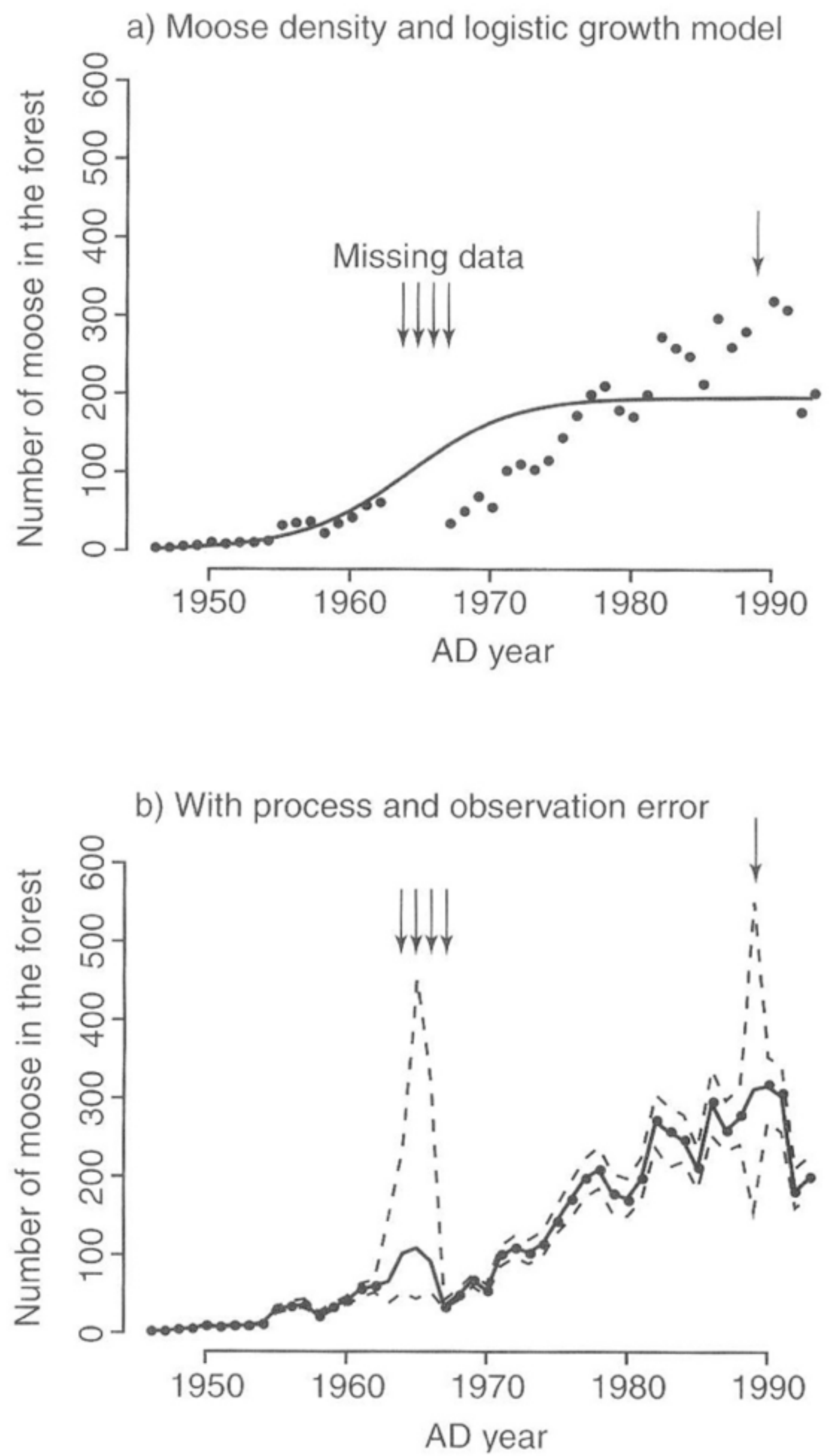


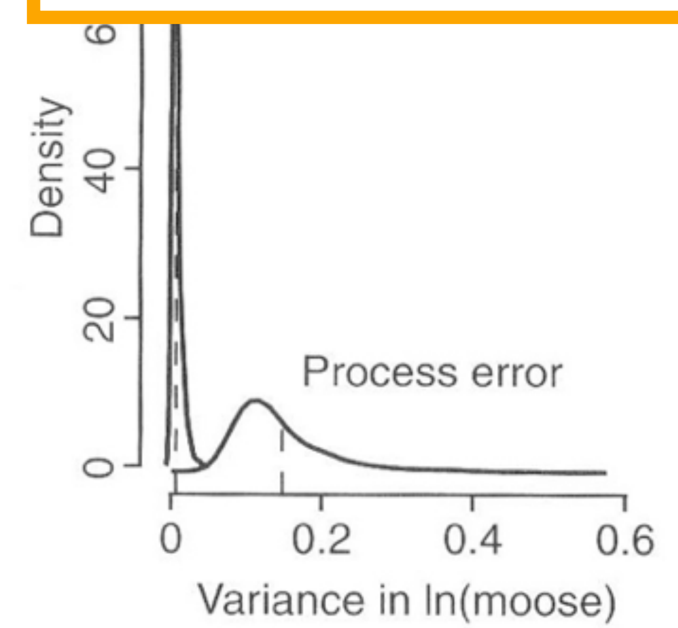


true
obs
fit
FORECASTING
Y[51:60] = NA
log(density)
0
2
4
6
8
0
10
20
30
40
50
60
time



FORECASTING
Y[51:60] = NA
true
obs
fit
log(density)
0
2
4
6
8
0
10
Process error propagates
Observation error stays constant



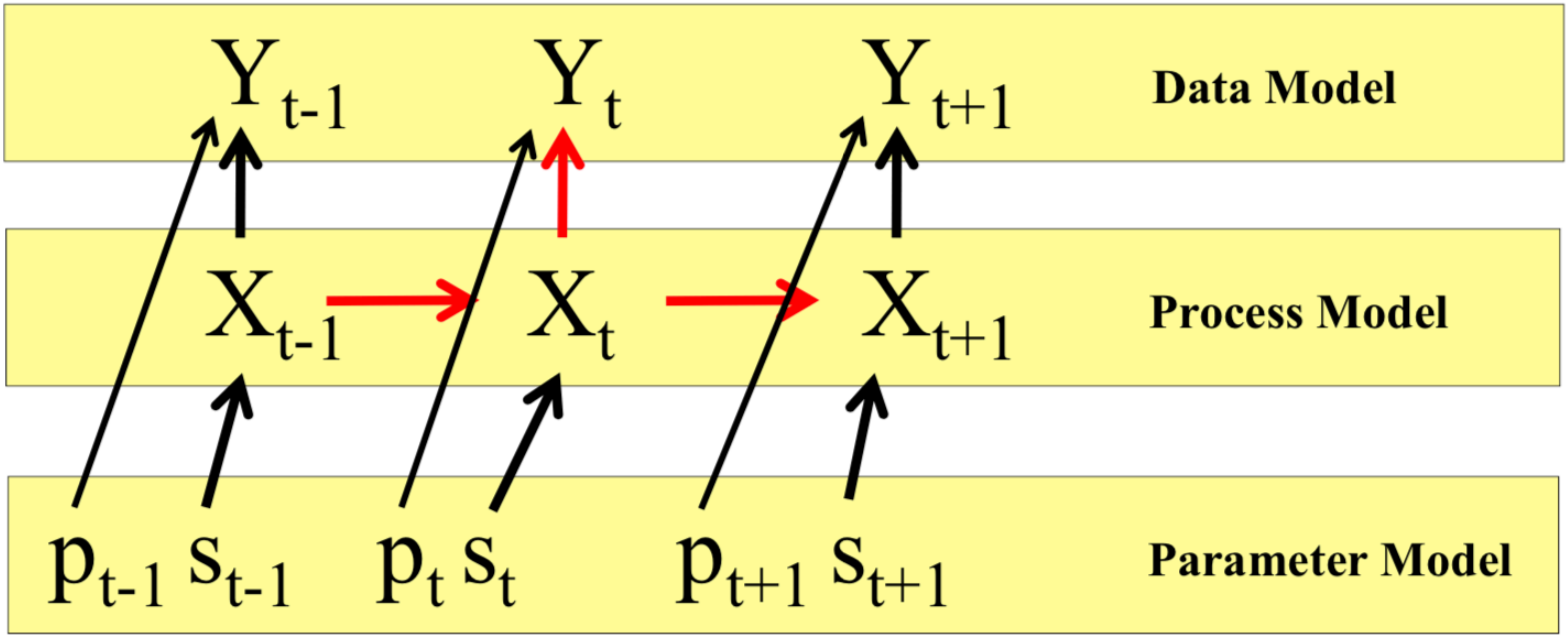



```
model
{
  for (i in 1:N) {
    y[t] ~ dnorm(x[t], tau)
  }
  for (i in 2:N) {
    x[t] ~ dnorm(mu[t], sigma)
    mu[t] <- x[t-1] + r * x[t-1]*(1-x[t-1]/K)
  }
  x[1] ~ dnorm(0.00000E+00, 0.01)
  r ~ dnorm(0.00000E+00, 0.01)
  K ~ dnorm(100,0.01)
  tau ~ dgamma(0.01, 0.01)
  sigma ~ dgamma(0.01, 0.01)
}
```

Generality of the State Space Framework

- Neither X nor Y need be Normal
- X and Y don't need to be the same type of data
- X and Y don't need to have the same time scale
- Handles missing data (gaps) and irregularly spaced data
- Handles multiple data sources (Y's), which don't need to be the same type or synchronous
- Handles time-integrated observations

Mark Recapture State Space



- Suppose an individual record consists of capture data

$$Y_i = [1,0,1,0,0]$$

- This is compatible with the following survival

$$X_i = [1,1,1,0,0]$$
$$X_i = [1,1,1,1,0]$$
$$X_i = [1,1,1,1,1]$$

Spatial Process

Raster:

Markov

Random

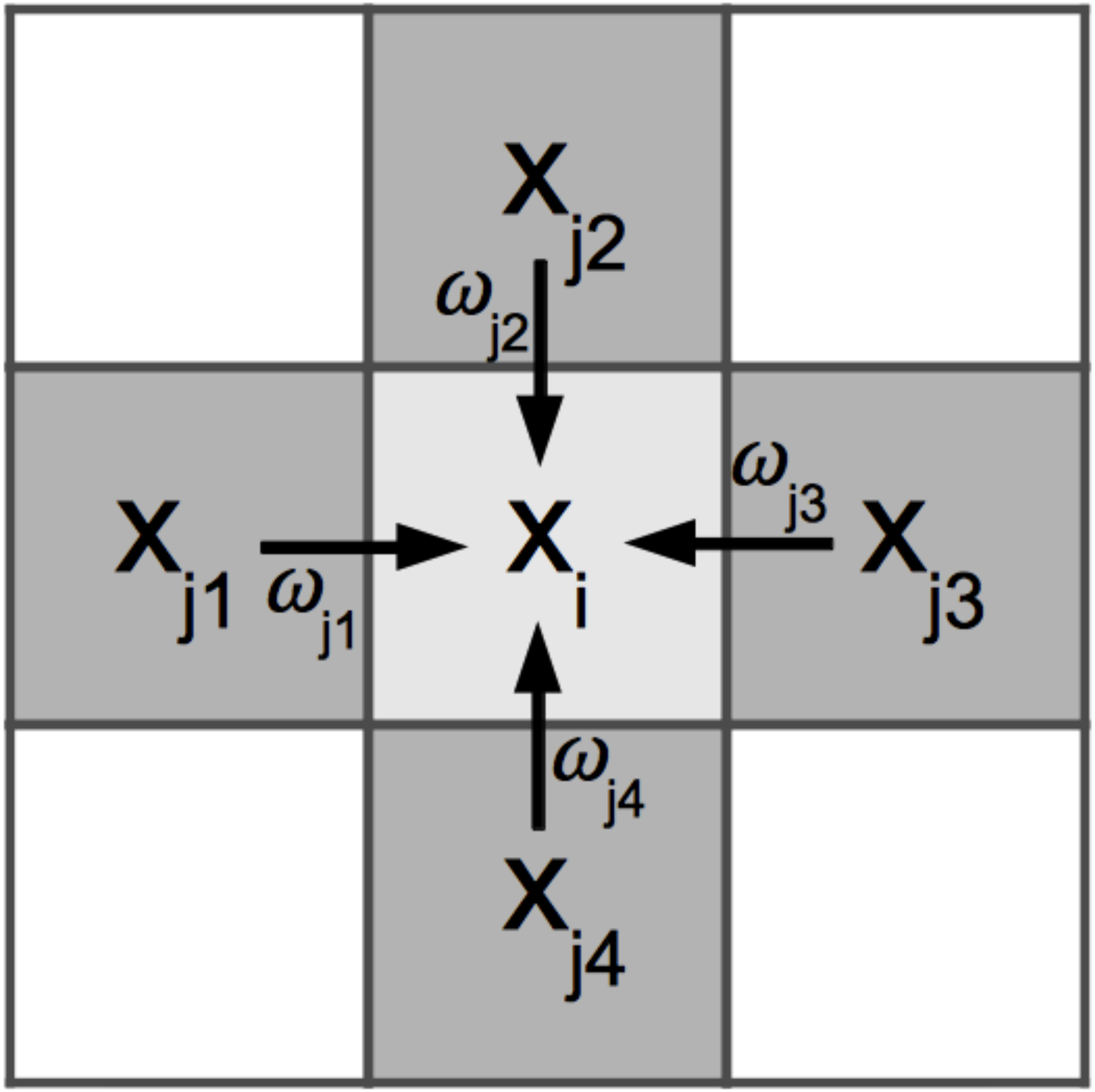
Field

Vector:

Conditional

Autoregressive

Alt to Kriging



$$x_i = \frac{1}{\sum \omega_j} \sum \omega_j x_j$$