Characterizing Uncertainty

Lesson 6

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!

SYNOPSIS

This section dives into the Bayesian methods for characterizing and partitioning sources of error that take us far beyond the classic assumption of a constant Normal variance.

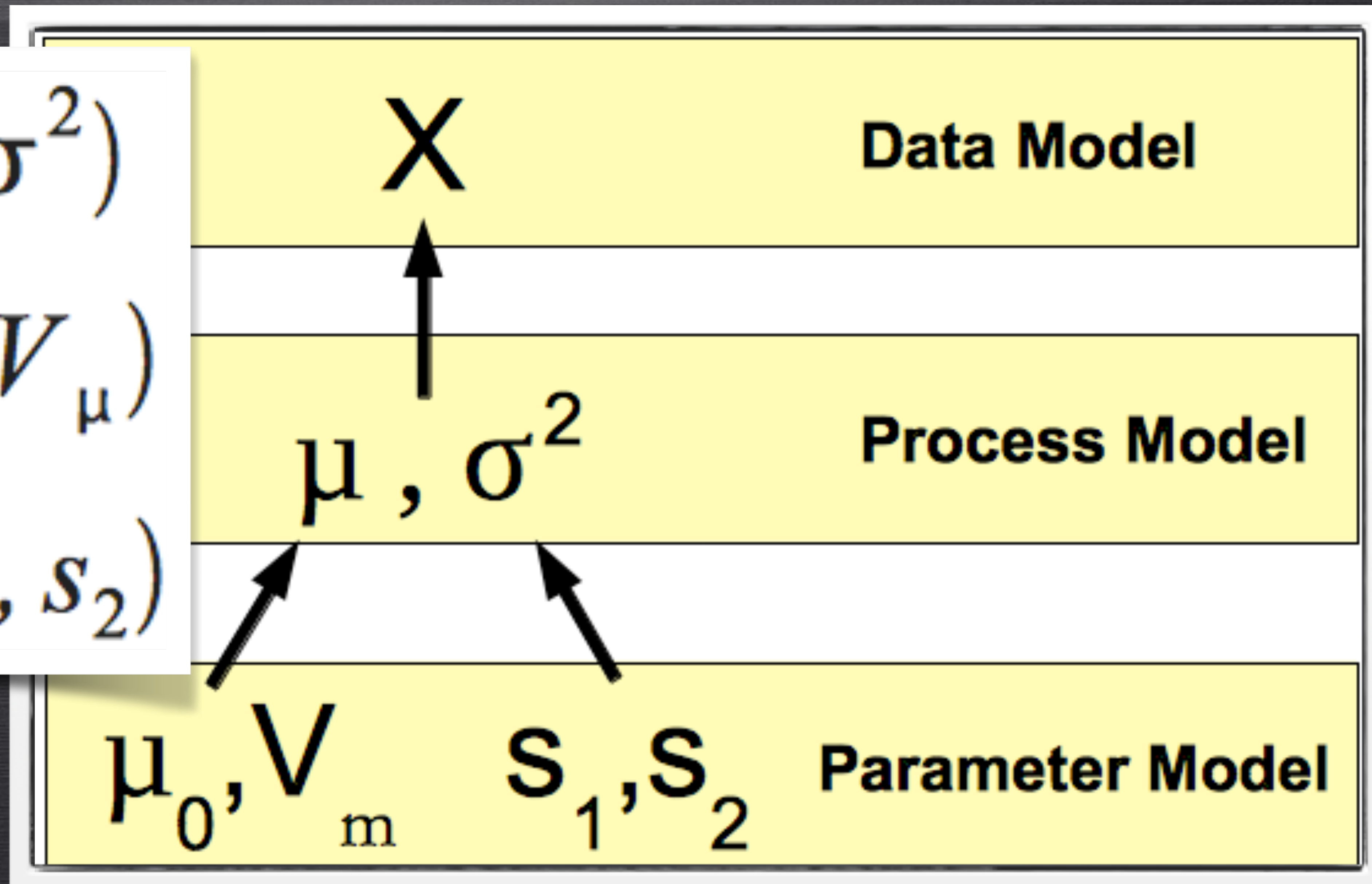
- Non-Gaussian
- **Errors in X variables**
- **Missing Data**
- **Hierarchical models**
- **State-Space (Wed)**
- Heteroskedasticity

Latent variables

$$X \sim N(\mu, \sigma^2)$$

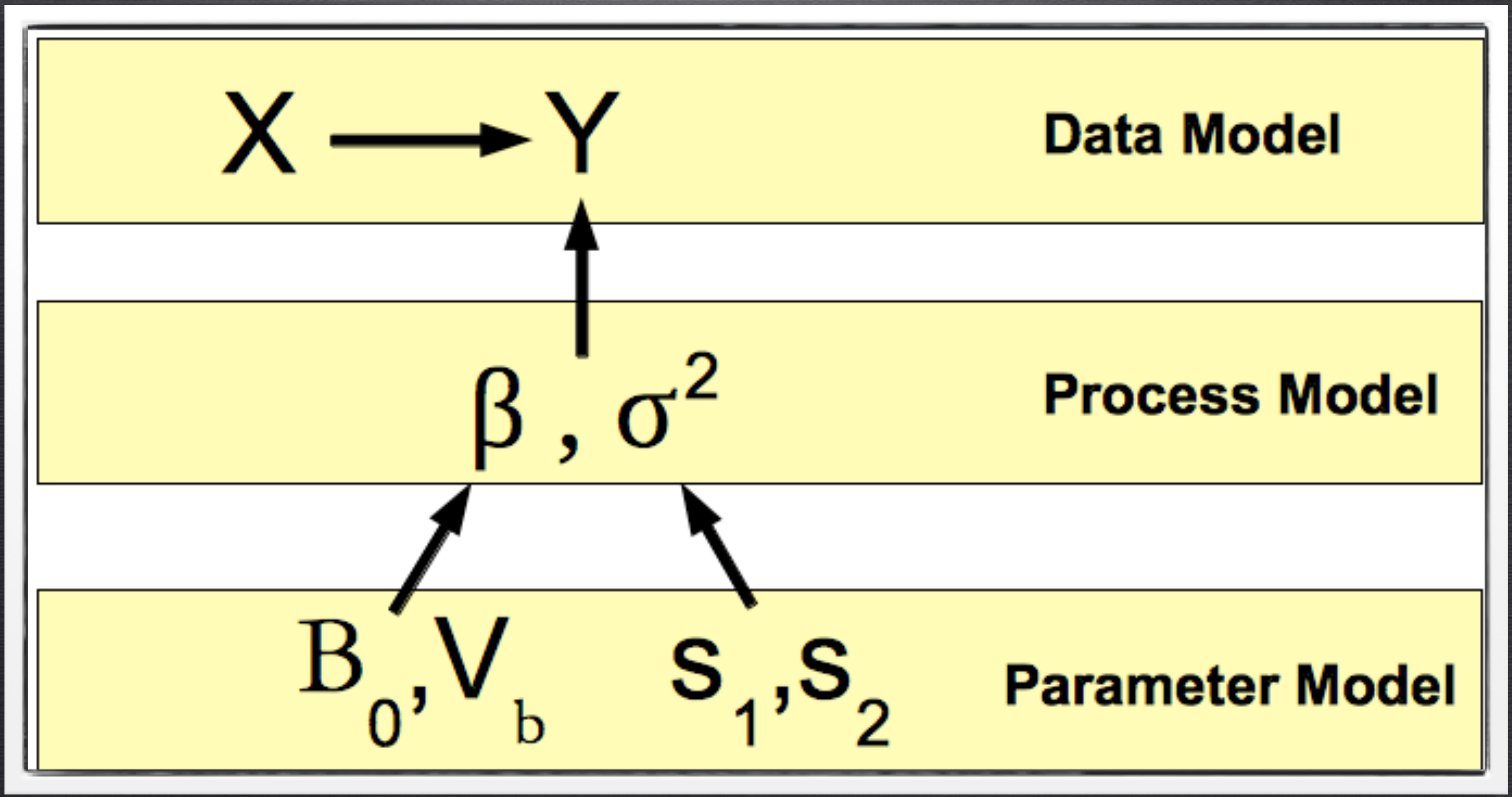
$$\mu \sim N(\mu_0, V_\mu)$$

$$\sigma^2 \sim IG(s_1, s_2)$$



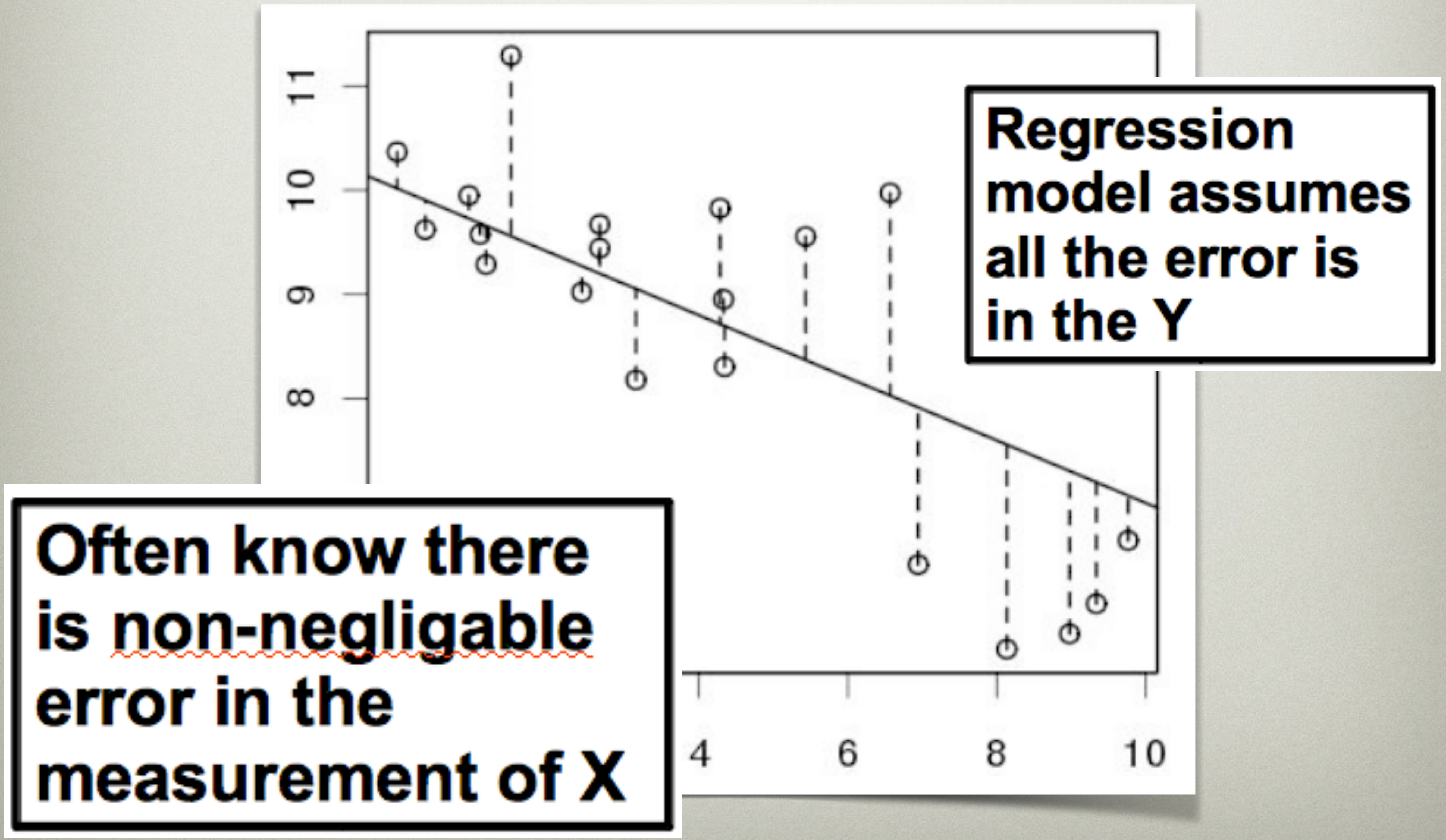
Graph Notation

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



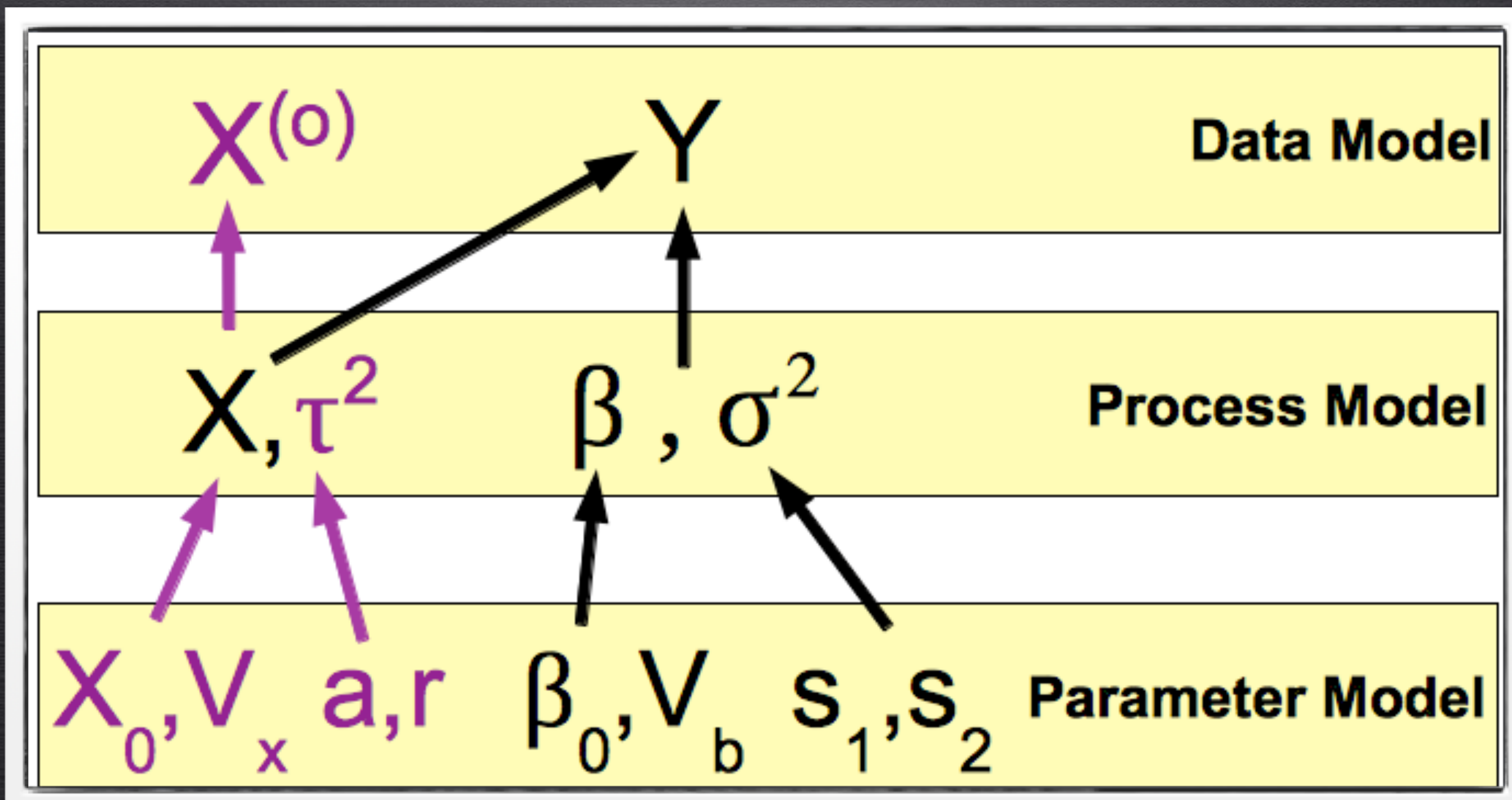
LINEAR REGRESSION

Errors in Variables



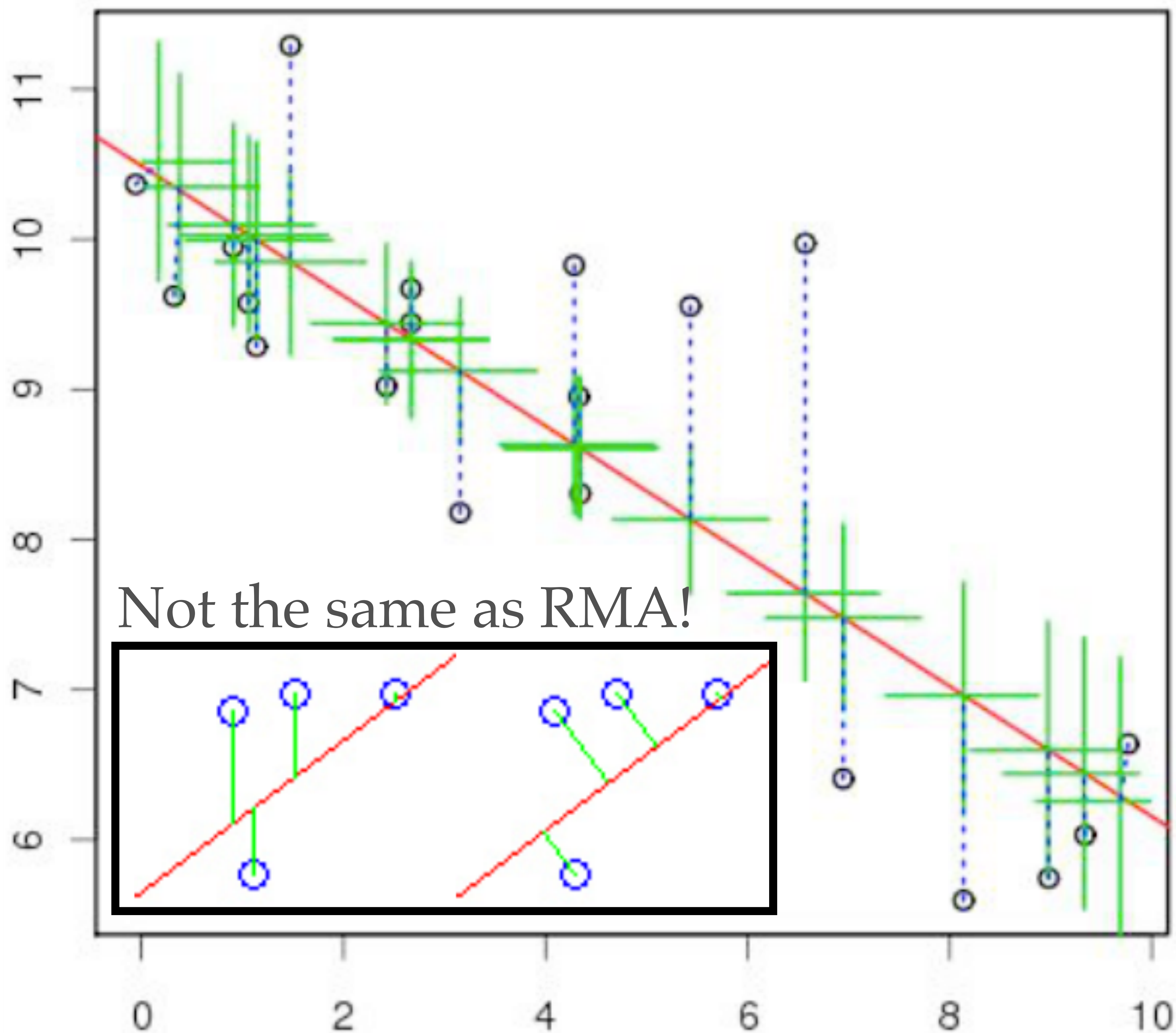
$$\vec{y} \sim N(\boldsymbol{X}\vec{\beta}, \sigma^2)$$

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



```
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)
  }
}
```


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

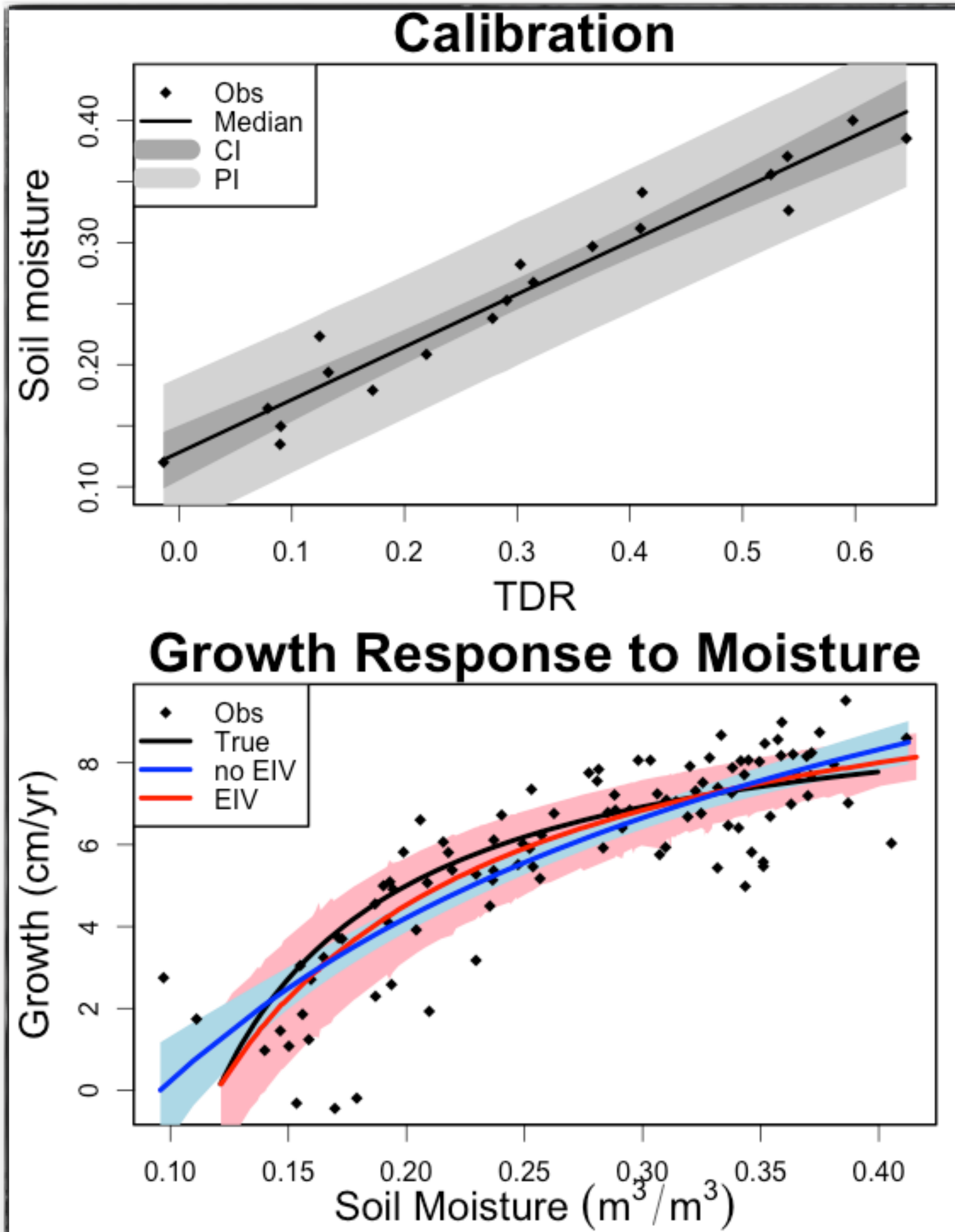
Additional Thoughts on EIV

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

- 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



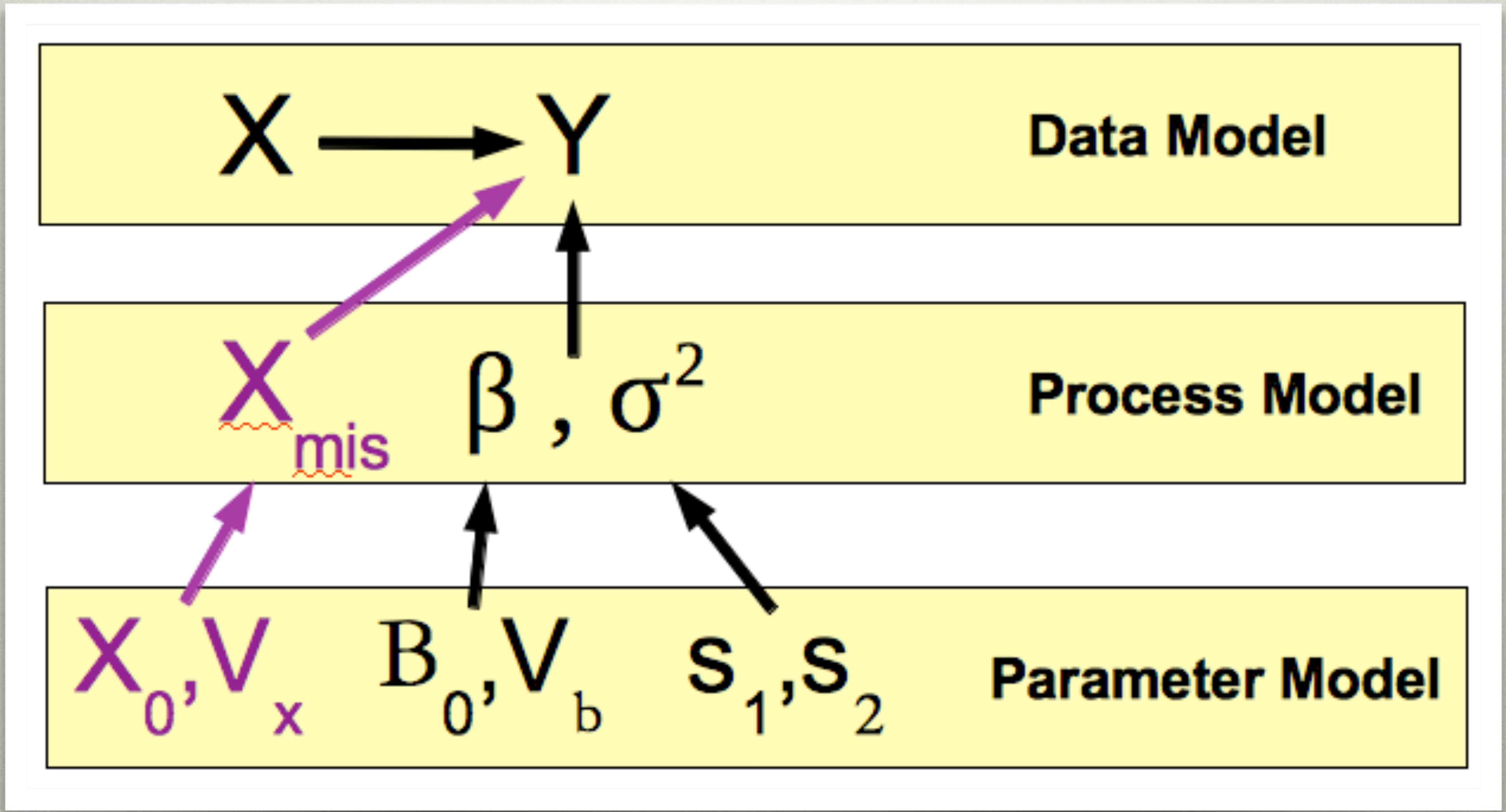
Calibration
Obs
Median
CI
PI
Soil moisture
0.10
0.20
0.30
0.40
0.0
0.1
0.2
0.3
0.4
0.5
0.6
TDR
Growth Response to Moisture
Obs
True
no EIV
EIV
Growth (cm/yr)
0
2
4
6
8
0.10
0.15
0.20
0.25
0.30
0.35
0.40
Soil Moisture (m3/m3)

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



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

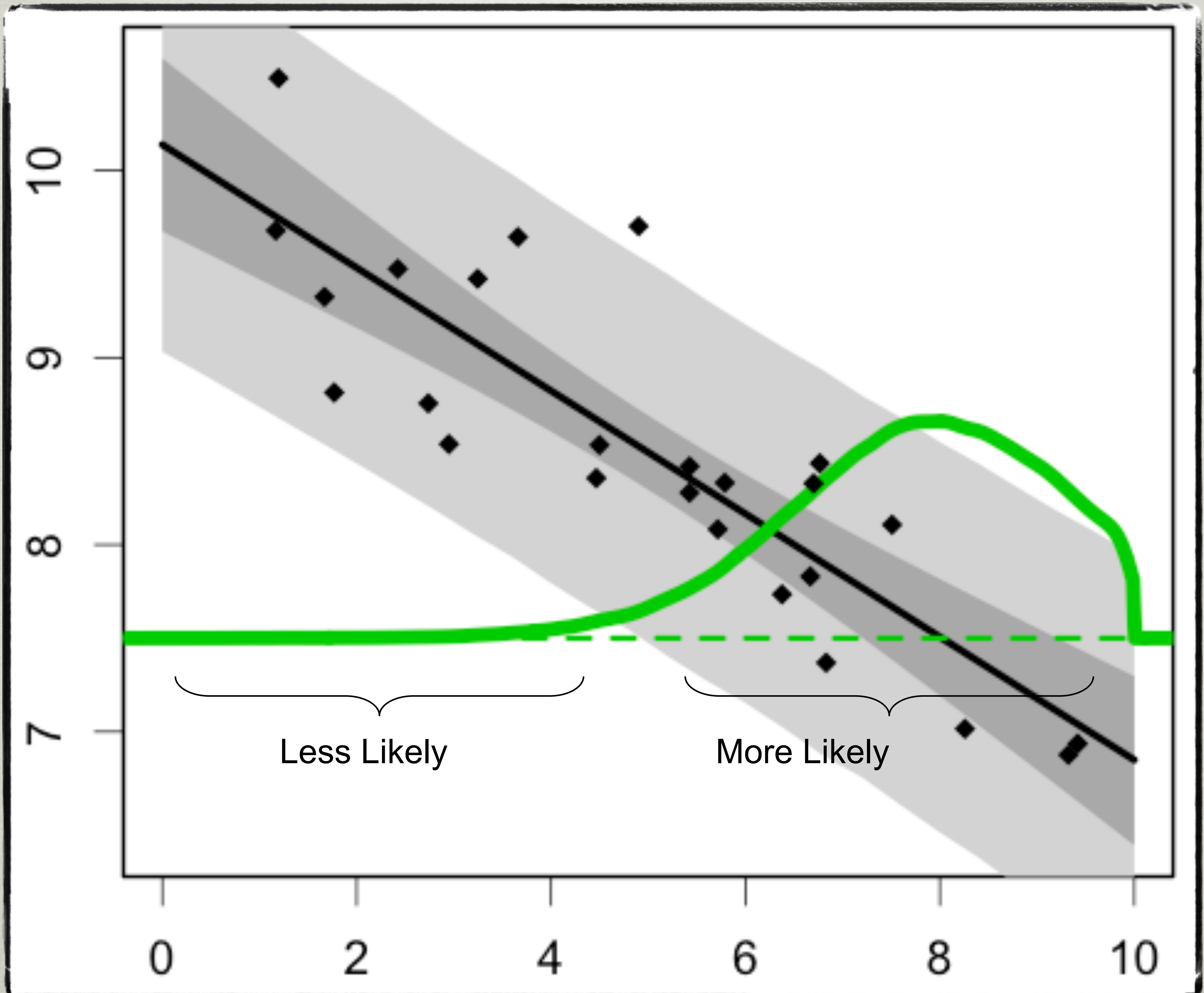
JAGS example: Simple Regression

```
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)
  }
}
```

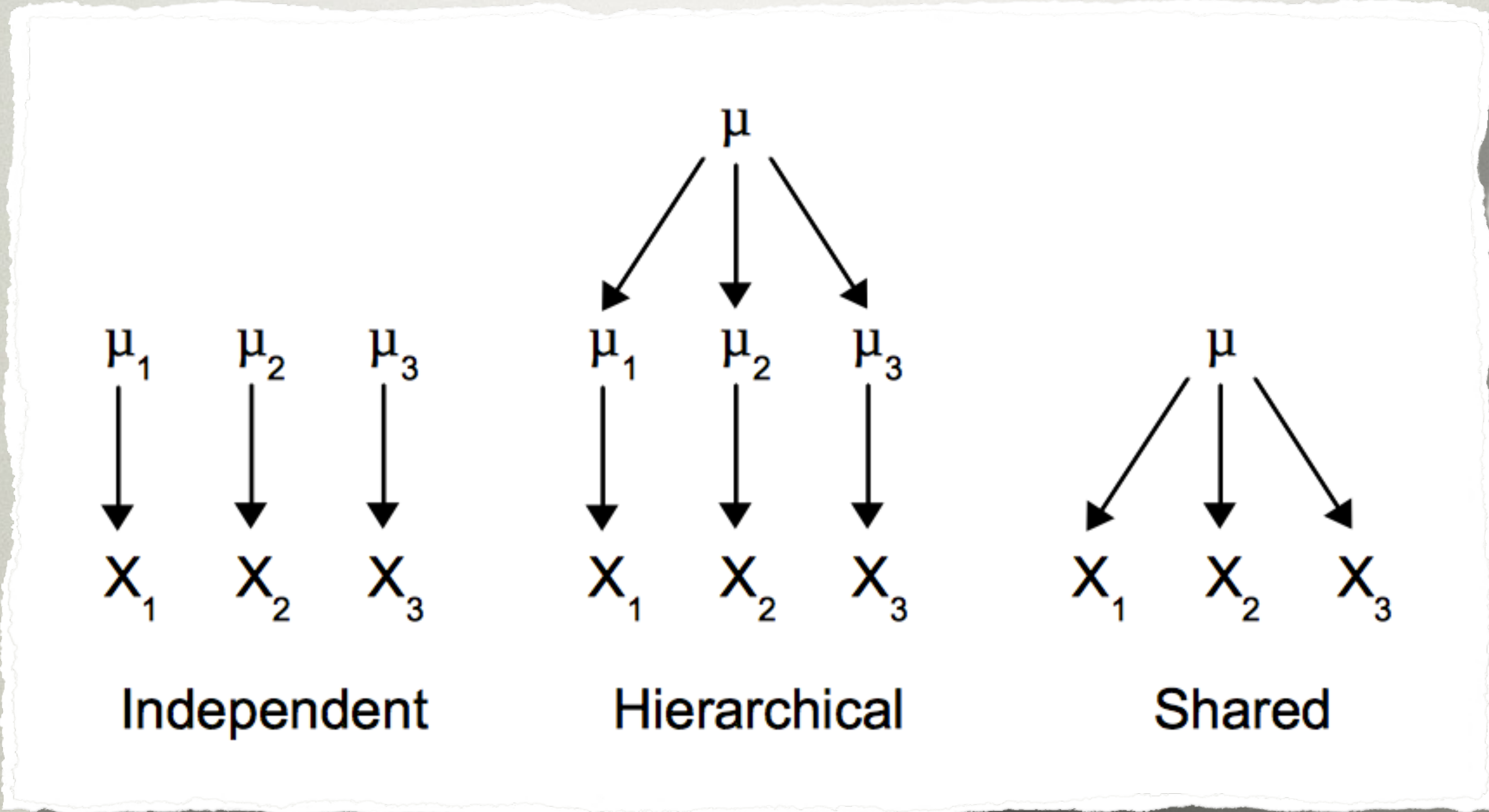
Vector giving indices of missing values

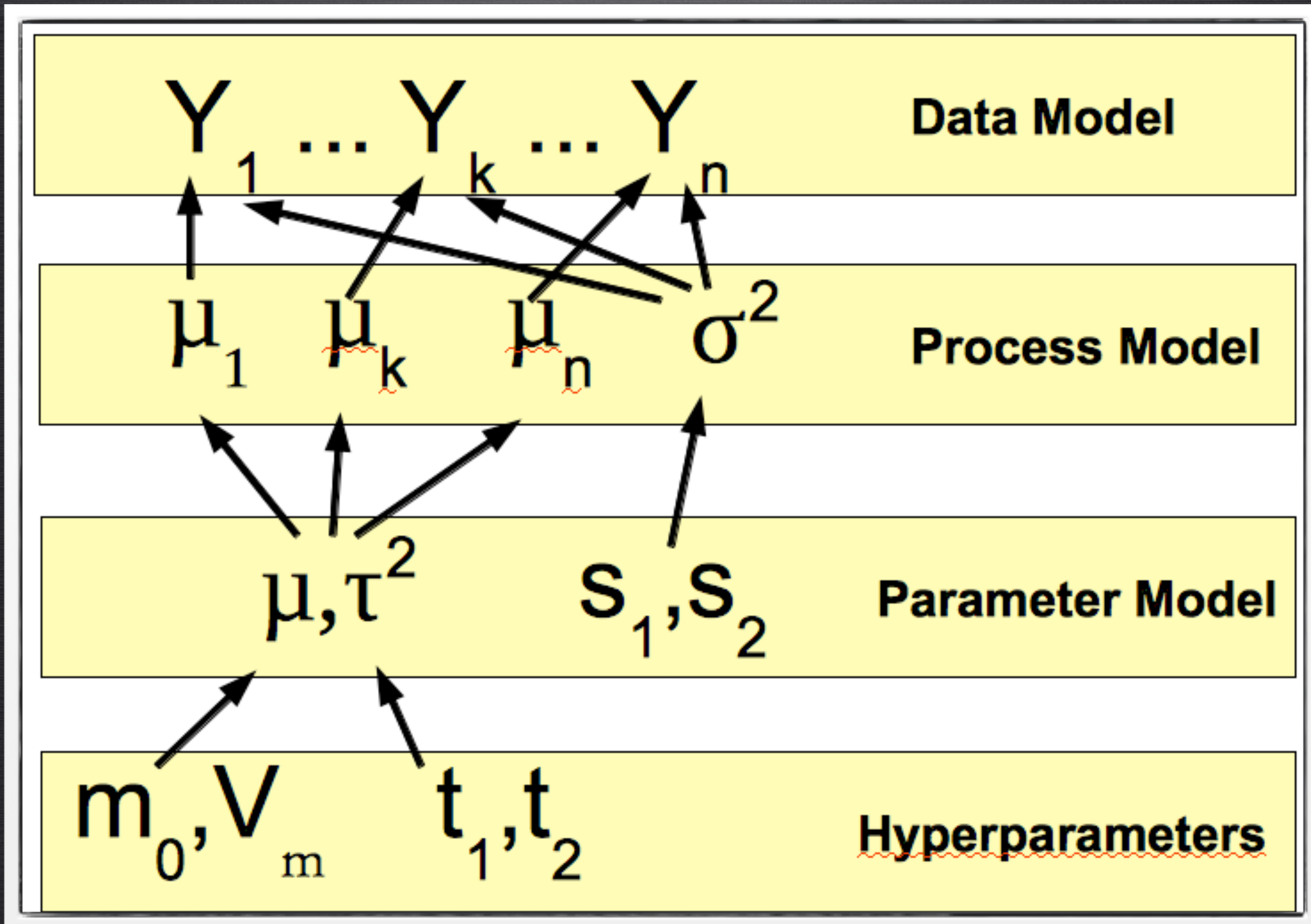
	X	Y
	4.68	8.46
	2.95	8.55
	9.09	7.01
	8.15	9.06
	1.76	11.38
	4.23	9.12
	7.73	7.3
	2.43	8.02
	6.46	8.45
	4.06	8.95
	2.42	9.62
	0.6	9.15
	8.17	7.51
	0.22	10.8
	4.93	9.78
	2.99	10.71
	8.36	8.89
	6.4	8.21
	8.17	6.22
	6.46	5.4
	1.82	10.05
	9.52	7.96
	2.44	9.63
	6.84	7.05
	7.42	8.73
$mis = 26$	NA	7.5



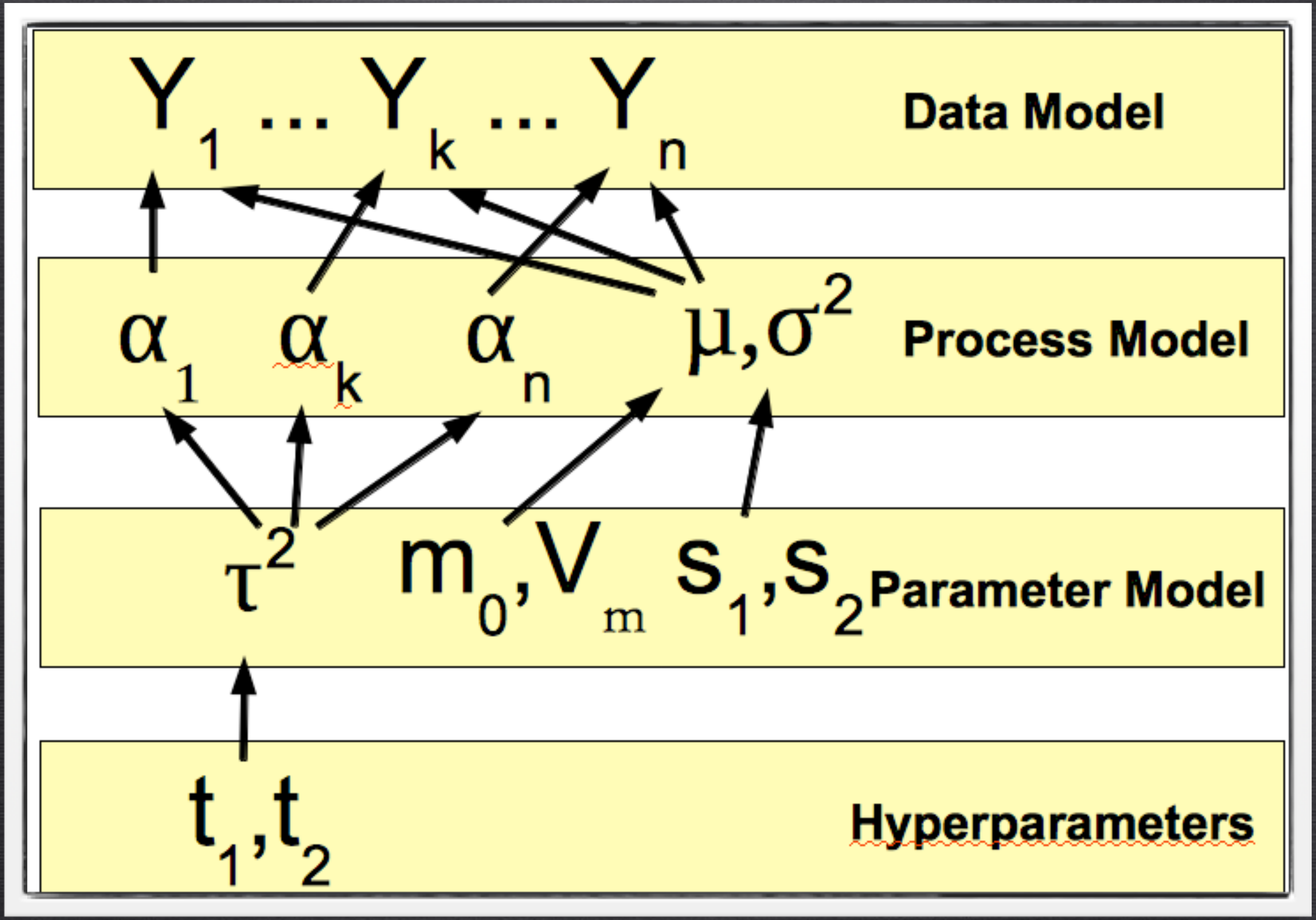
10
9
8
7
Less Likely
More Likely
0
2
4
6
8
10

Hierarchical Models





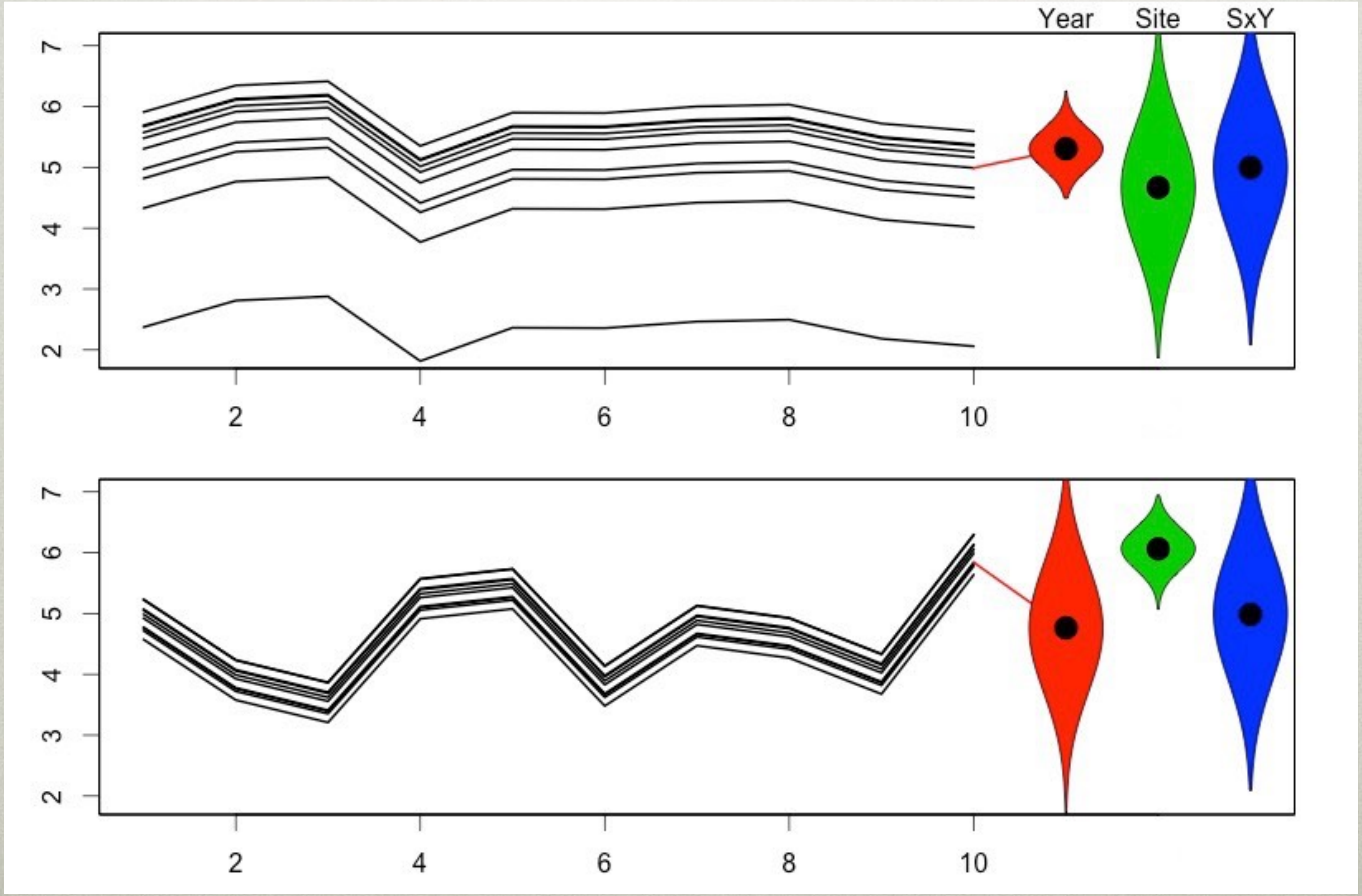
Y1 ... Yk ... Yn
Data Model
μ1 μk μn σ2
Process Model
μ,τ2
s1,s2
Parameter Model
m0,Vm
t1,t2
Hyperparameters



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

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

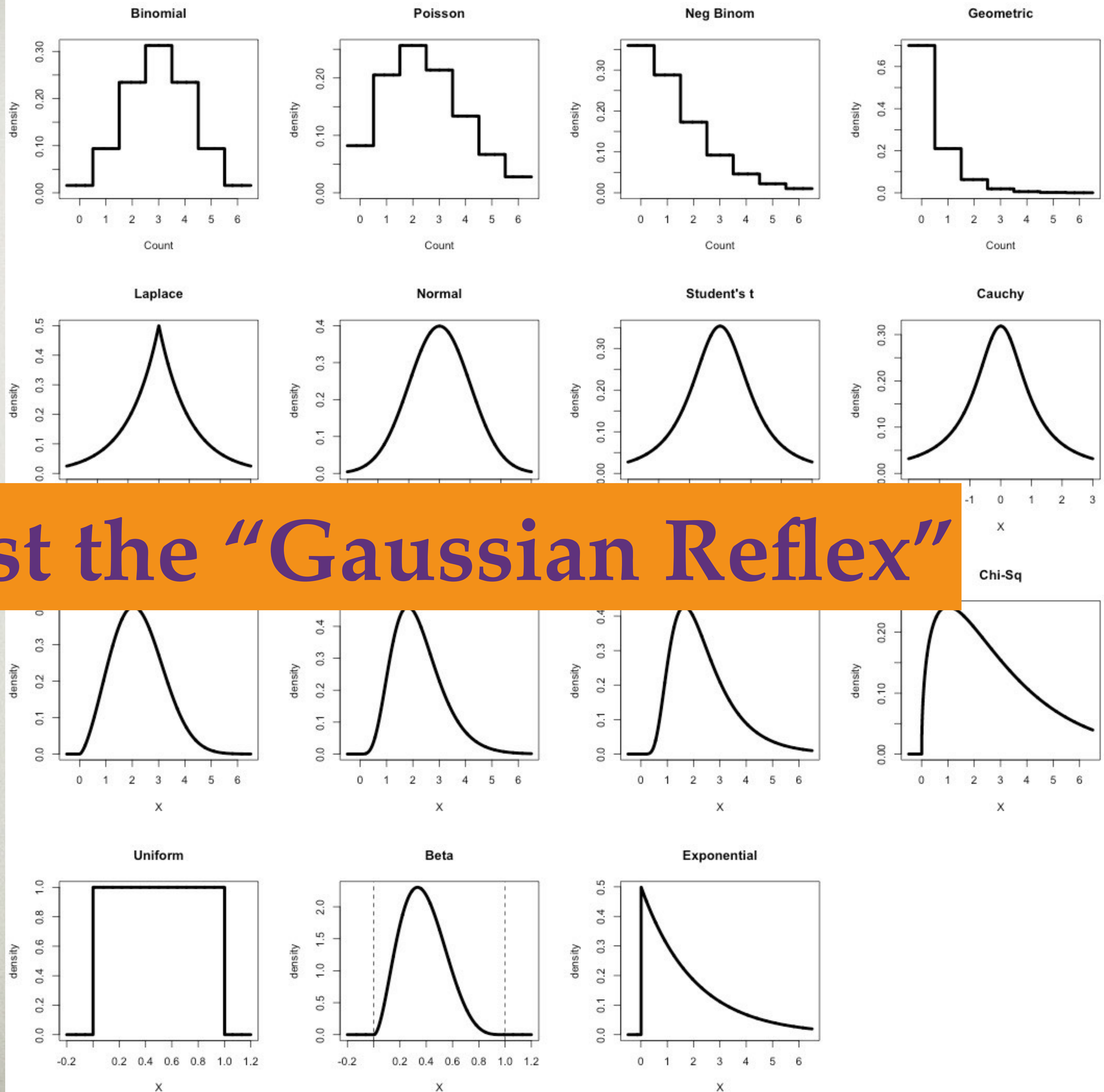
Impacts on Inference



Explaining unexplained variance

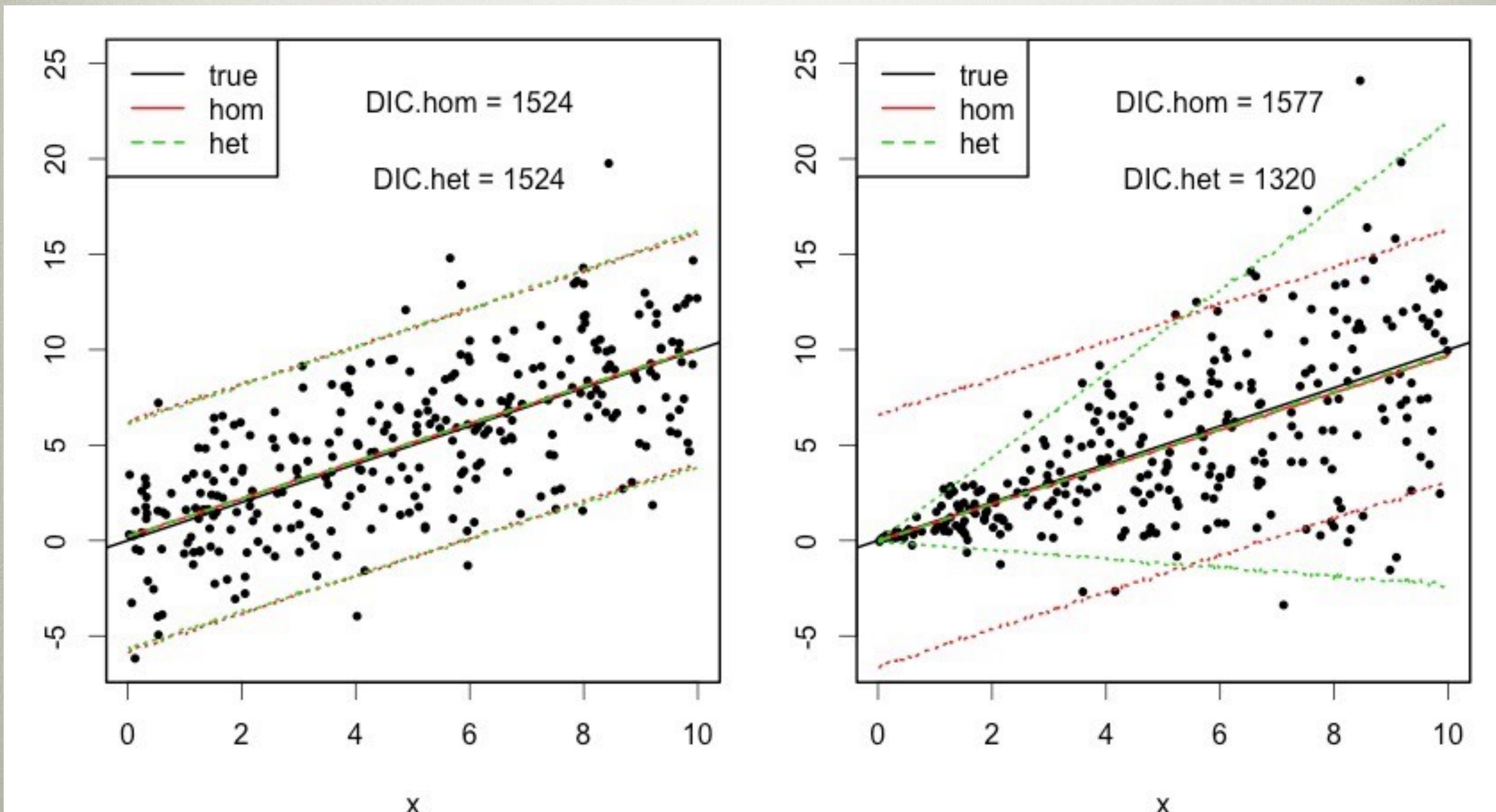
- Random effects attempt to account for the unexplained variance associated with some group (plot, year, etc.) due to all the things that were not measured
- May point to scales that need additional explanation
- Adding covariates may explain some portion of this variance, but there's always something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

Choosing a Distribution

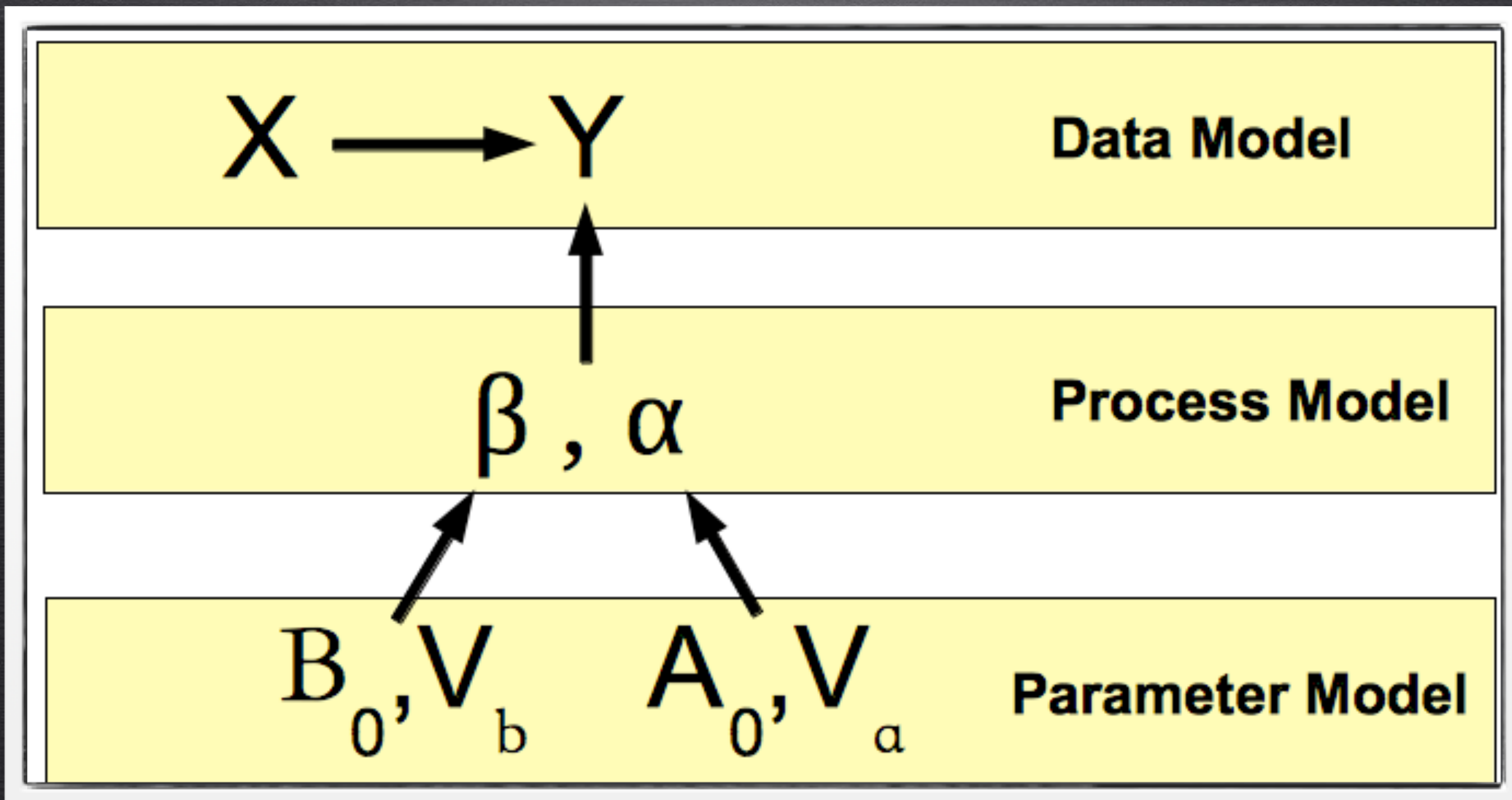


Resist the "Gaussian Reflex"

Heteroskedasticity



$$y \sim N(\beta_1 + \beta_2 x, (\alpha_1 + \alpha_2 x)^2)$$



Example: Linear varying SD

```
model{
  for(i in 1:2) { beta[i] ~ dnorm(0,0.001)}  ## priors
  prec ~ gamma(s1,s2)
  for(i in 1:n){
          mu[i] <- beta[1]+beta[2]*x[i]
          y[i] ~ dnorm(mu[i],prec)
  }
}
```

```
model{
  for(i in 1:2) { beta[i] ~ dnorm(0,0.001)}  ## priors
  for(i in 1:2) { alpha[i] ~ dlnorm(0,0.001)}  ## was prec ~ gamma(a1,a2)
  for(i in 1:n){
          prec[i] <- 1/pow(alpha[1] + alpha[2]*x[i],2)
          mu[i] <- beta[1]+beta[2]*x[i]
          y[i] ~ dnorm(mu[i],prec[i])
  }
}
```