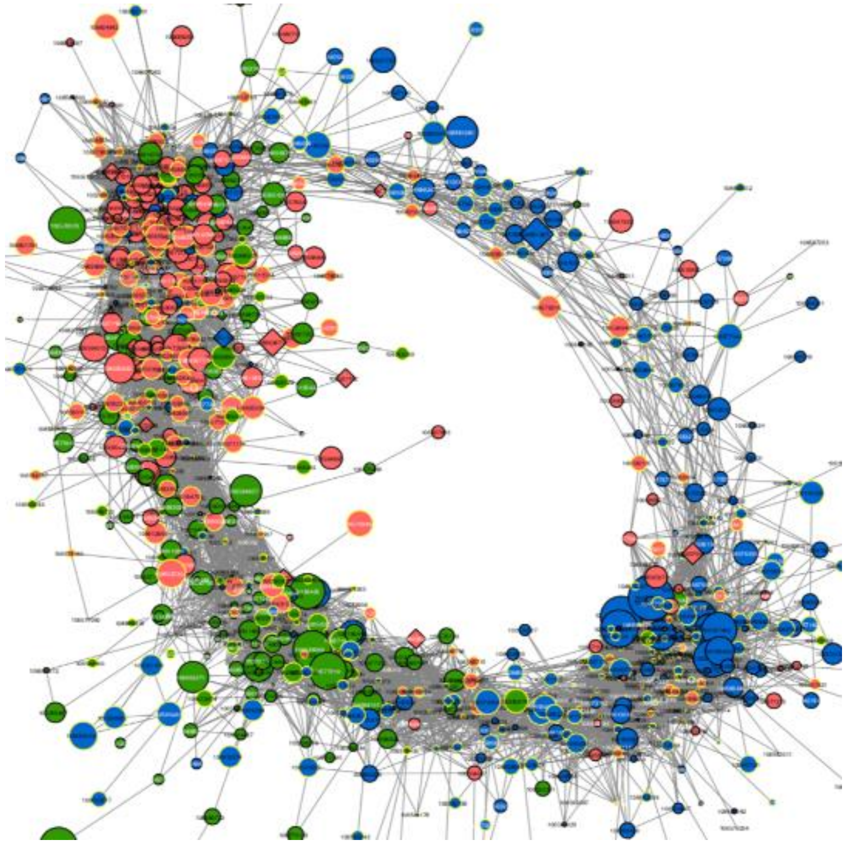
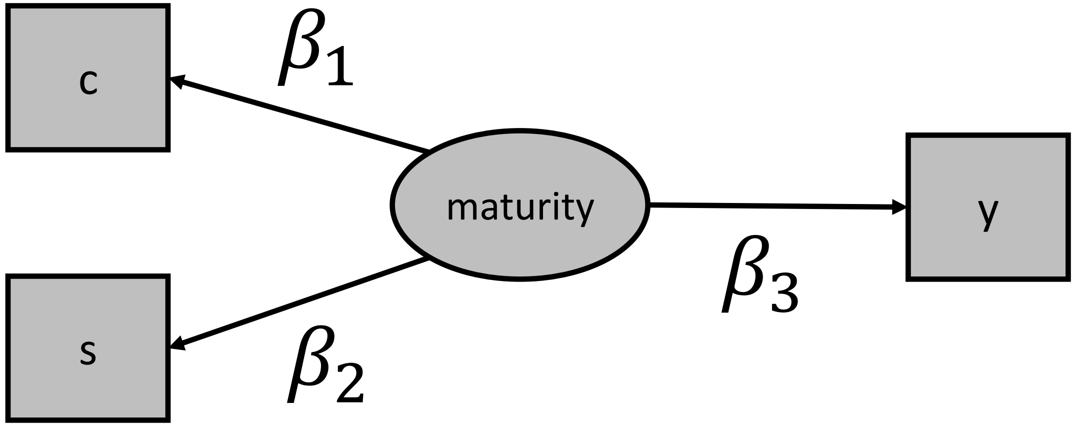
Latent variables





Building blocks

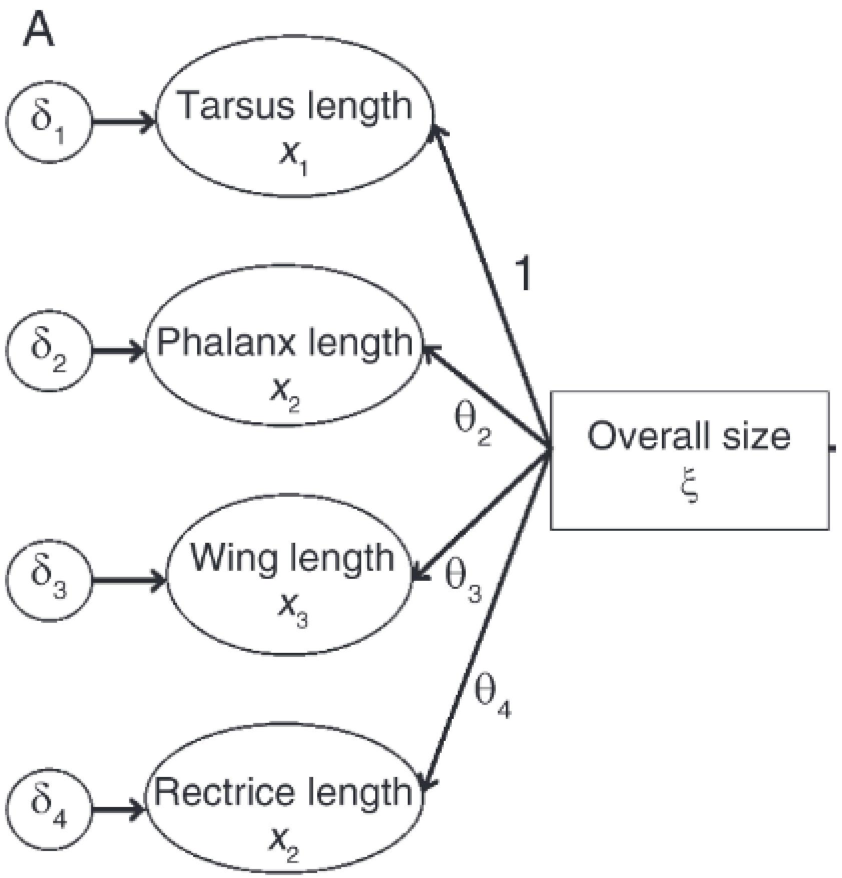
Sarah Cubaynes

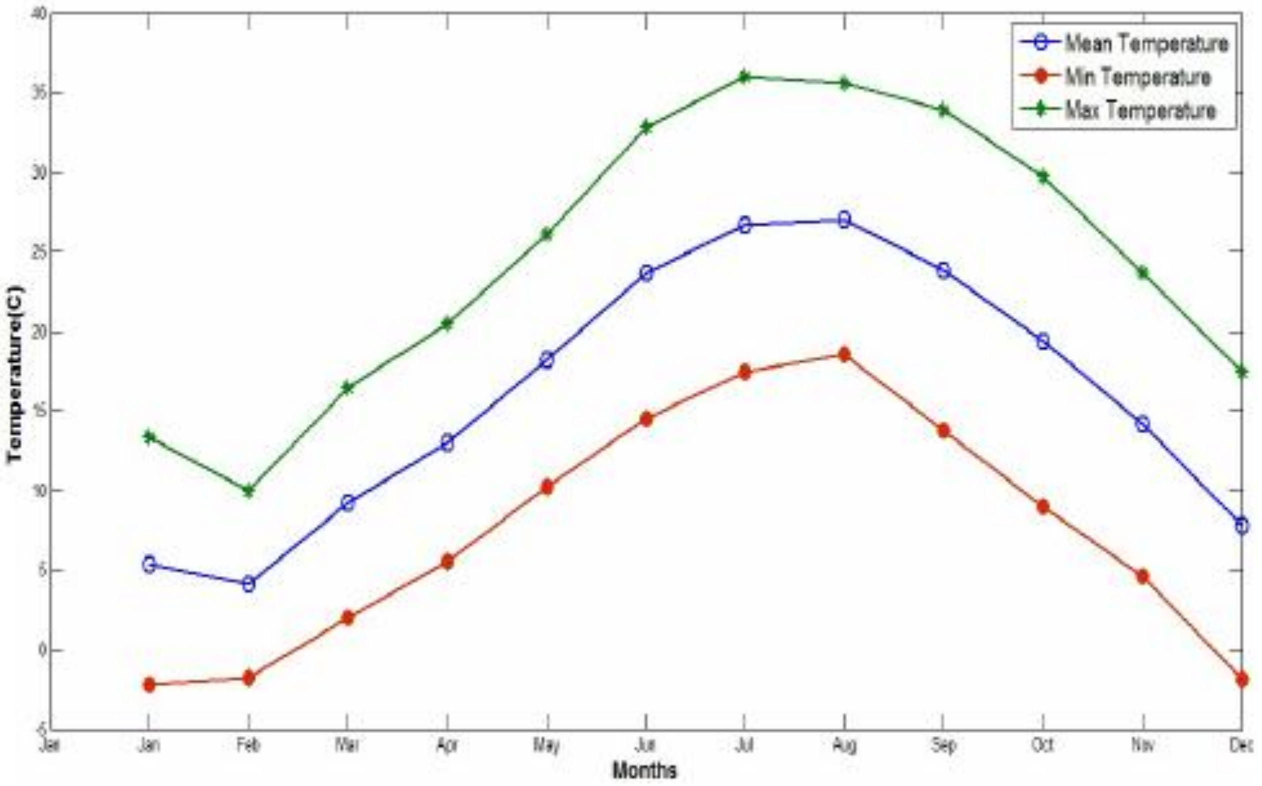
Martijn van de Pol

Cubaynes et al. (**2012**) *Ecology*

van de Pol et al. **(2021)** *Journal of Animal Ecology*

This module’s question: what if we’re kinda measuring the same thing?





What is a latent (or hidden) variable?

A random variable that is unmeasured but not necessarily unmeasureable.
-P Spirtes (2001)

A variable that is hypothesized to exist, but that has not been measured directly
-J Grace (2006)

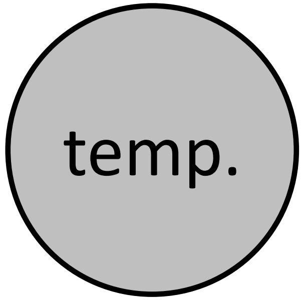
A variable that is not directly observable but is inferred from other variables that can be measured
-Generative AI (yesterday)

Variables that can only be inferred indirectly through a mathematical model from other observable variables
-Wikipedia (also yesterday)

What is a latent (or hidden) variable?

Everything is a latent variable **– LA Dyer**

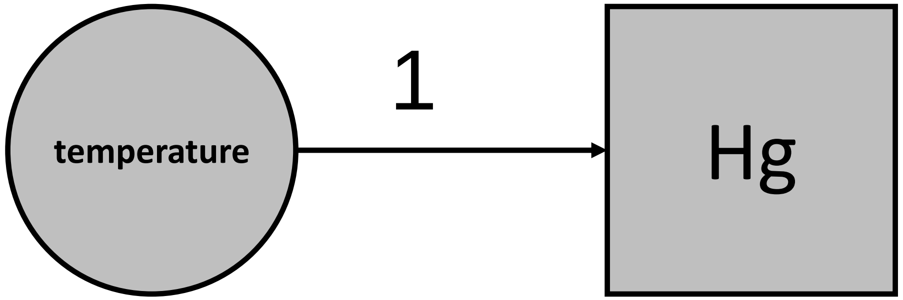
Is temperature a latent variable?



A random variable that is unmeasured but not necessarily unmeasureable.
-P Spirtes (2001)

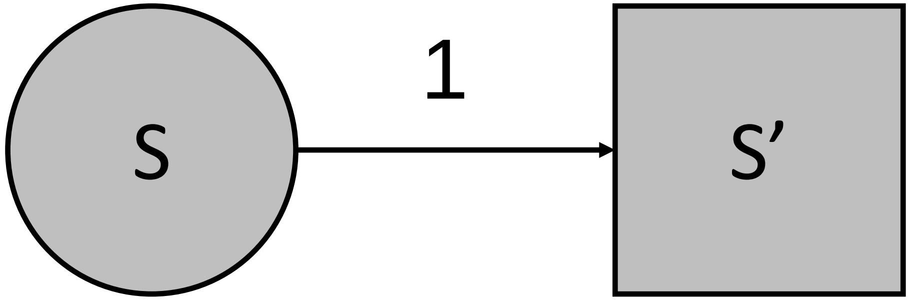
Temperature is the average kinetic energy of particles

Temperature is a latent variable



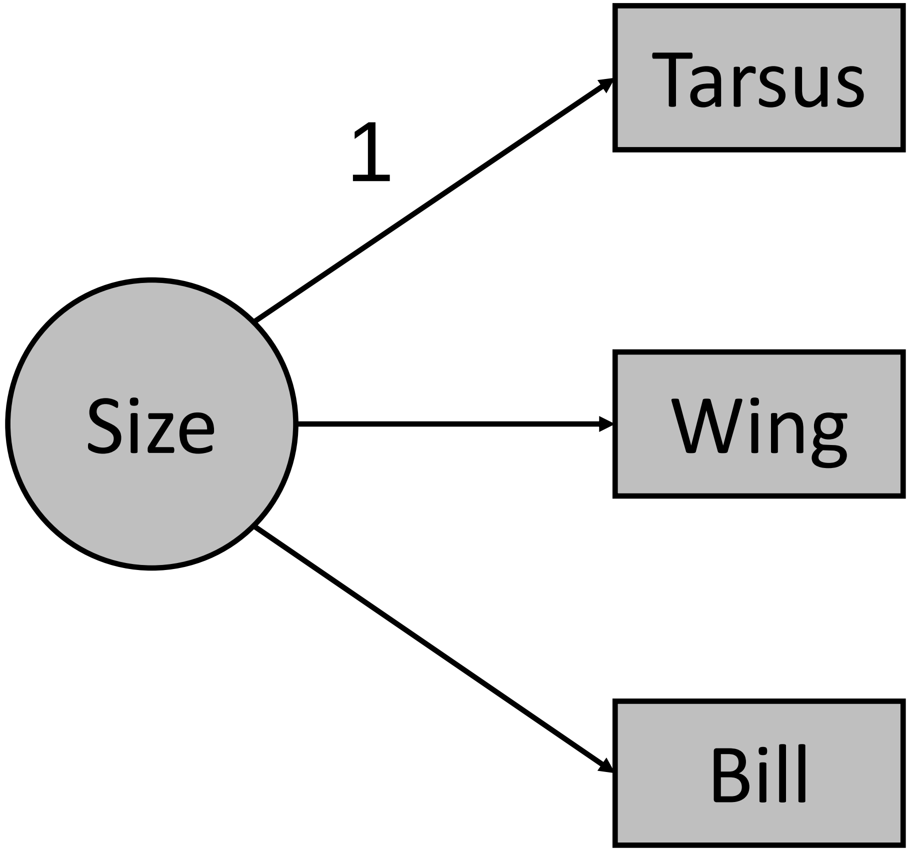
Temperature is the average kinetic energy of particles
We measure it (with error) via the expansion of mercury (or lasers)

Is survival a latent variable?



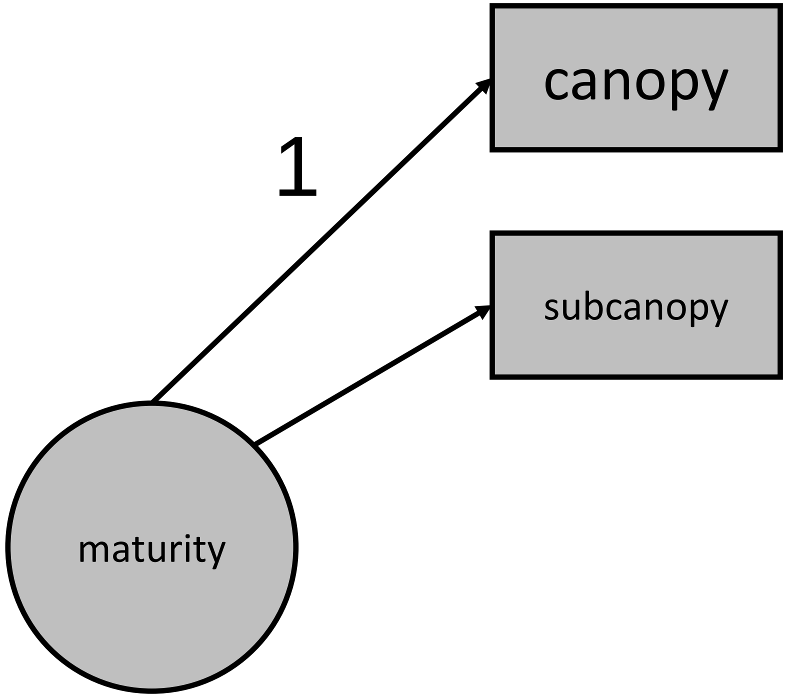
S: survival of a population, S’: survival of a marked sample

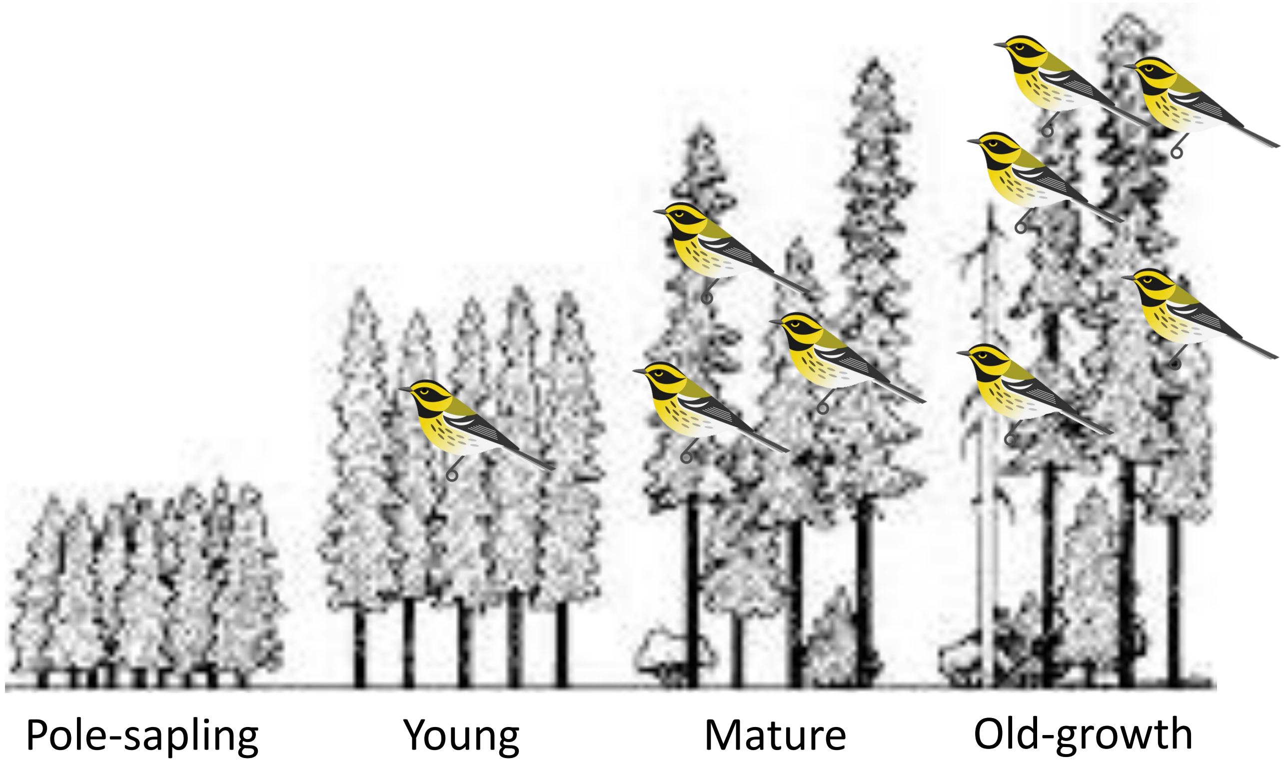
Size



‘Size’ is a human construct (i.e., latent variable)

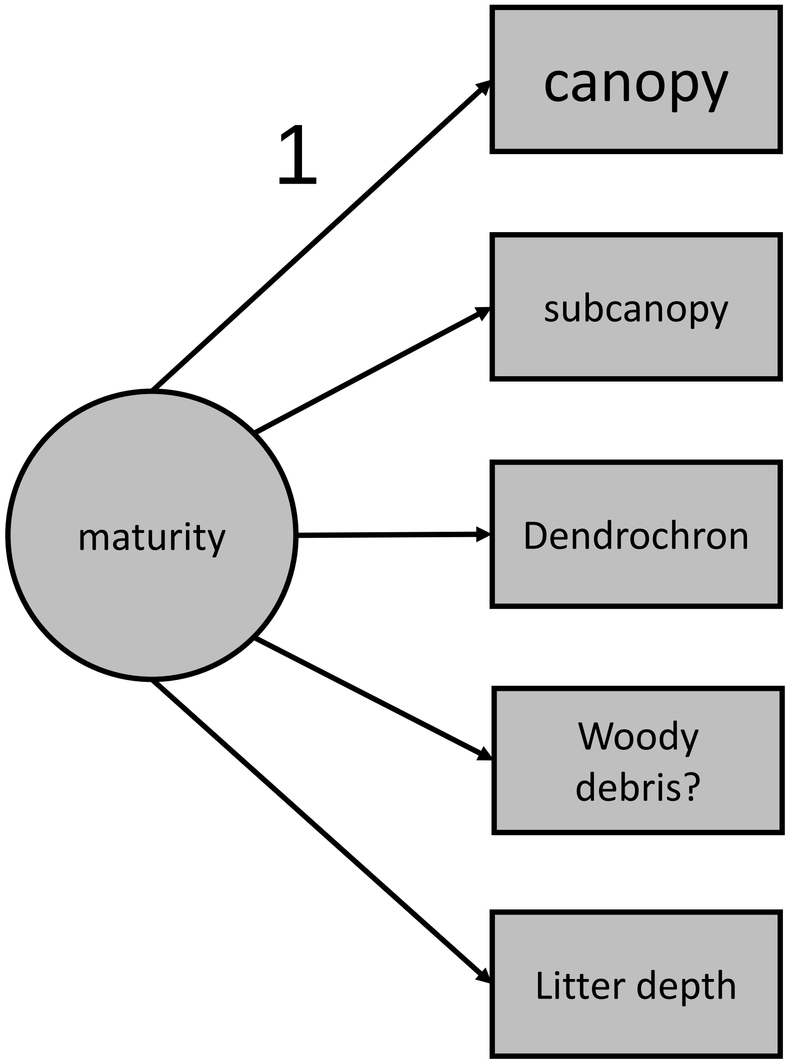
Forest maturity

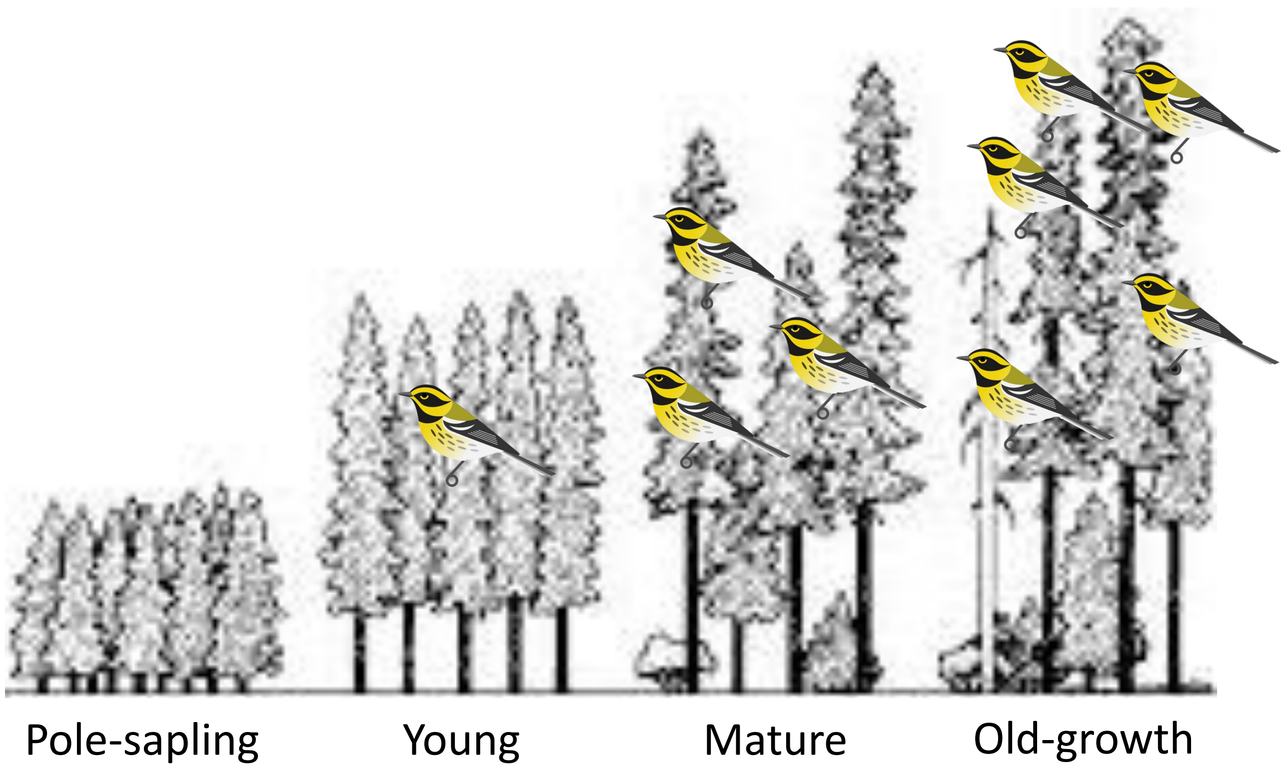




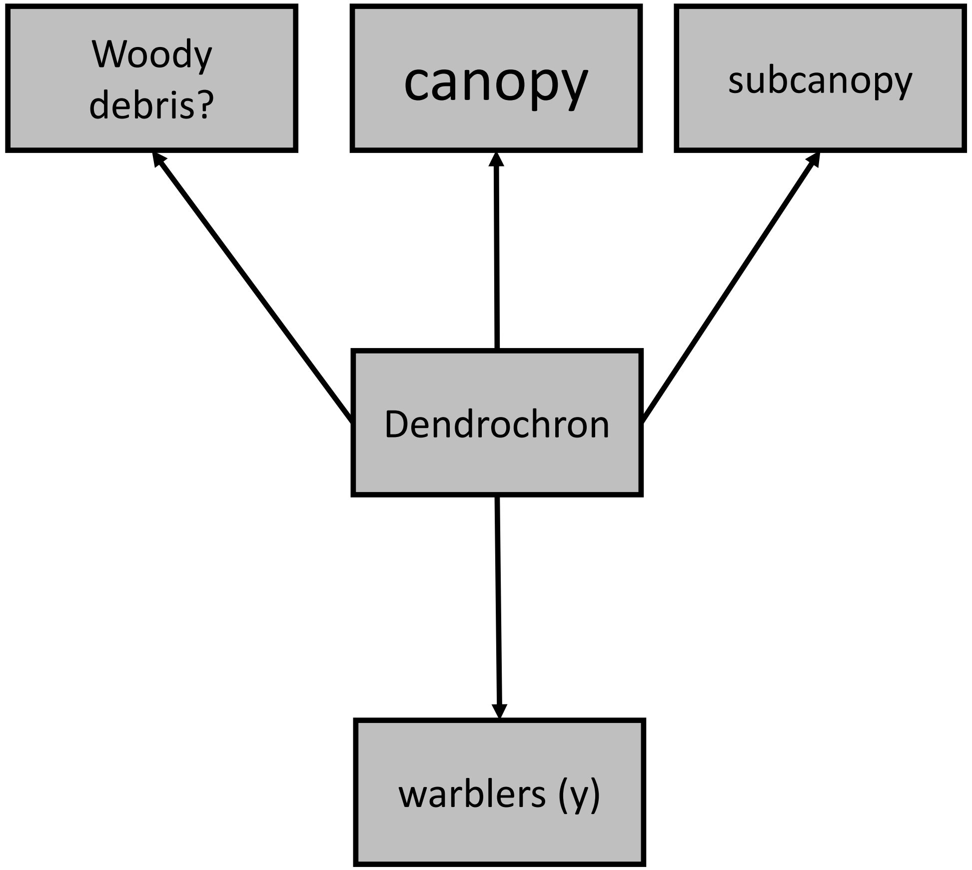
These ‘seral stages’ are human constructs

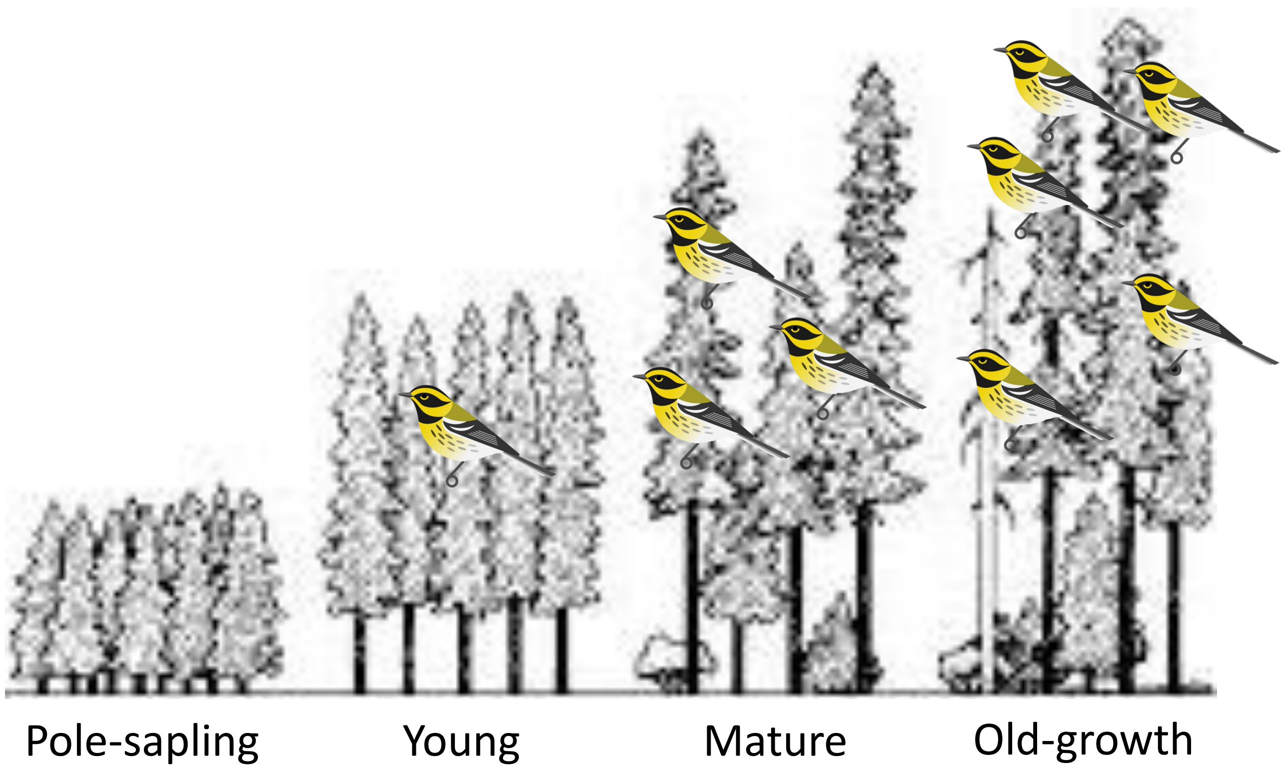
Forest maturity [expanded]





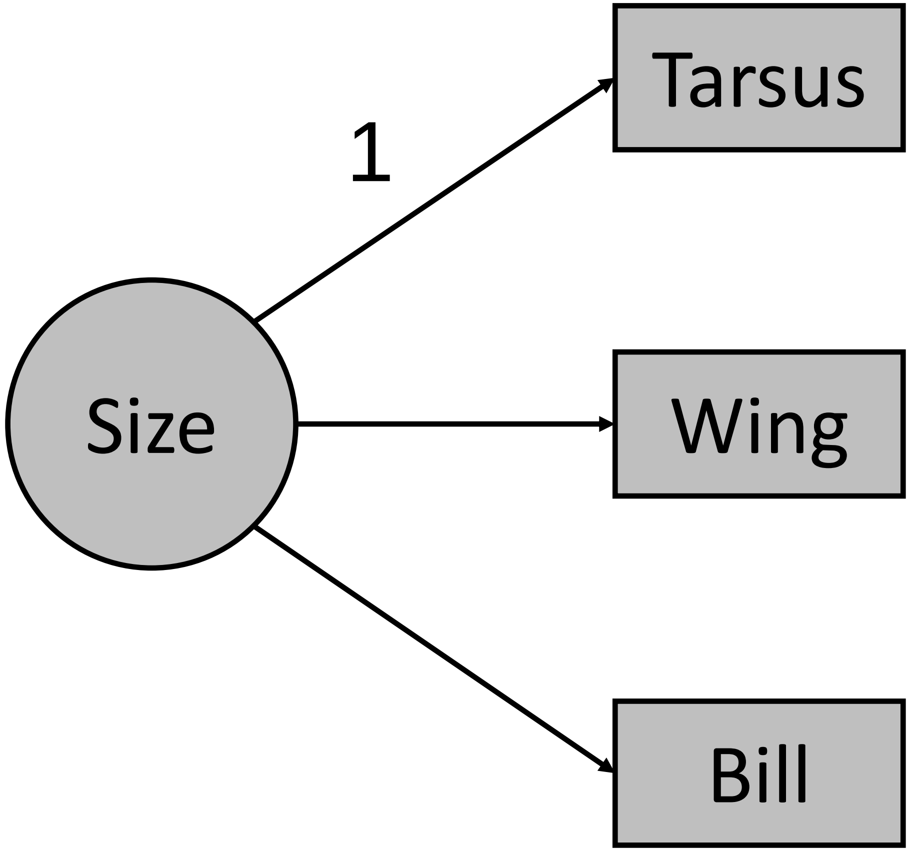
We could structure this differently





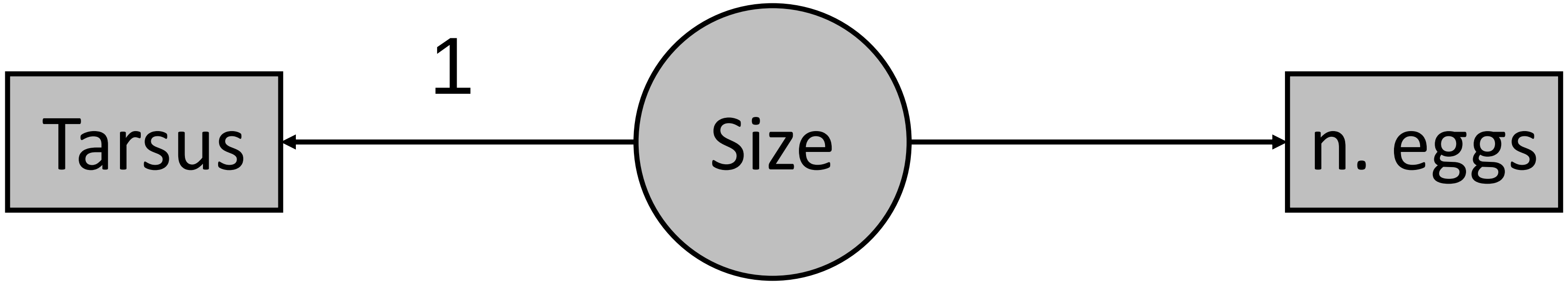
Path analysis: we don't have to use latent variables!

They're just very useful…

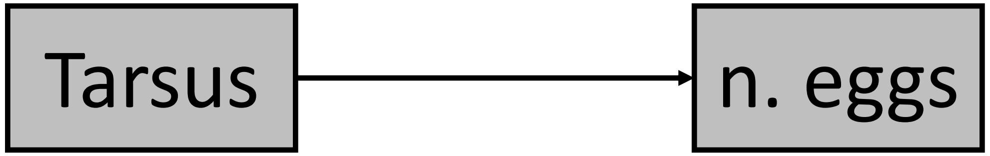


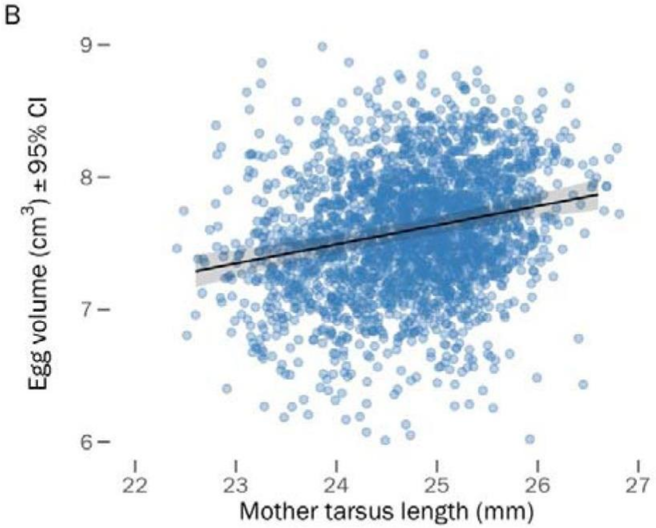
Further, I argue that <u>we use them all the time</u> subconsciously

Subconscious model



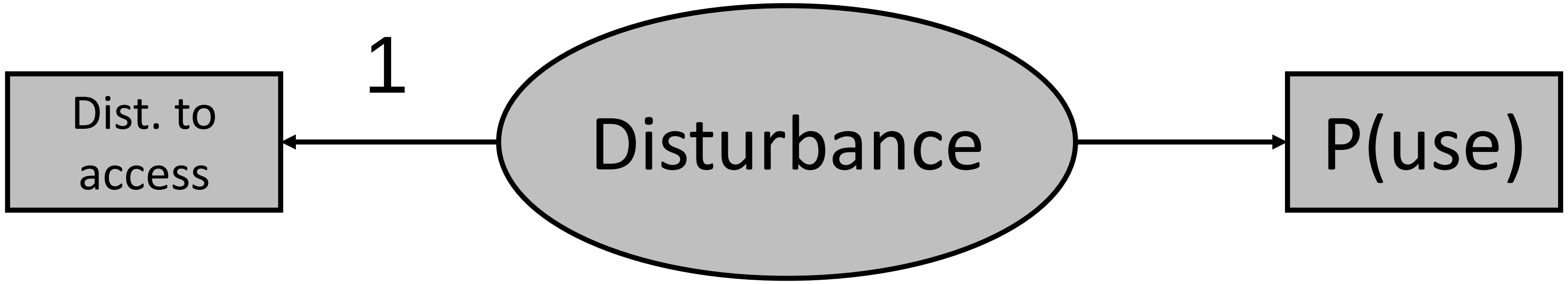
Actual model



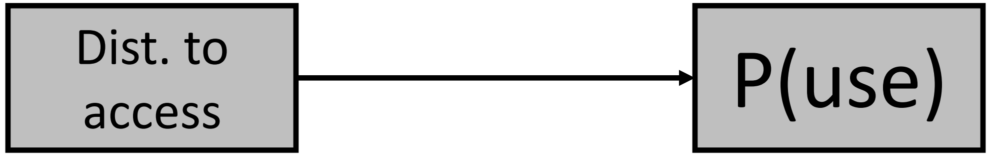


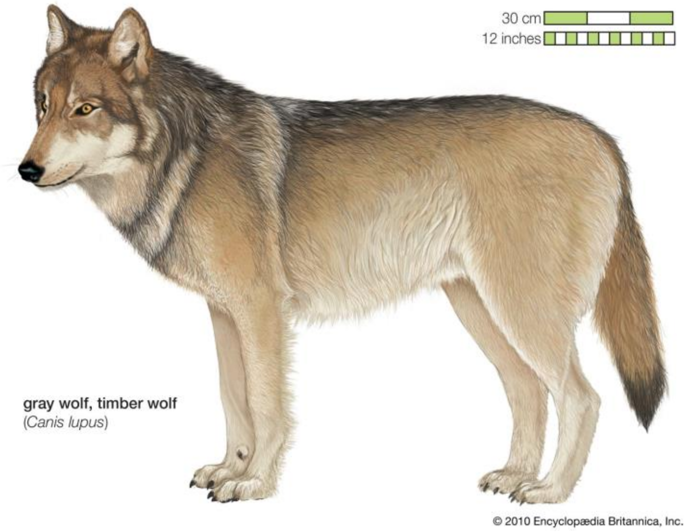
Further, I argue that we use them all the time subconsciously

Subconscious model



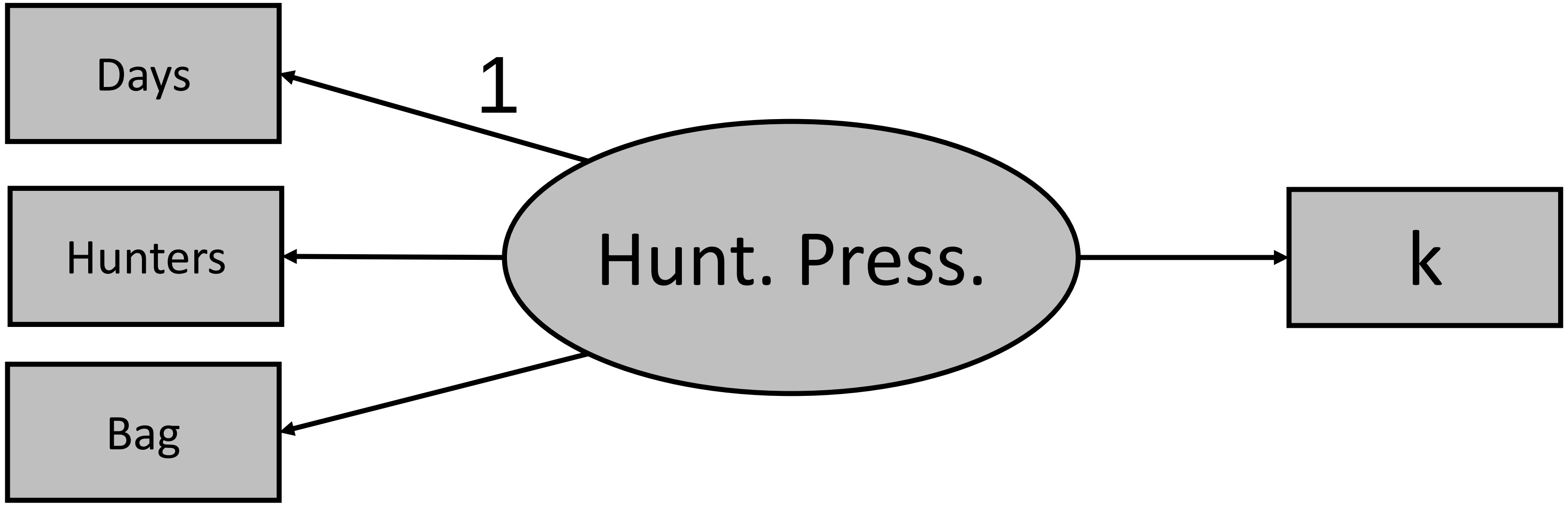
Actual model



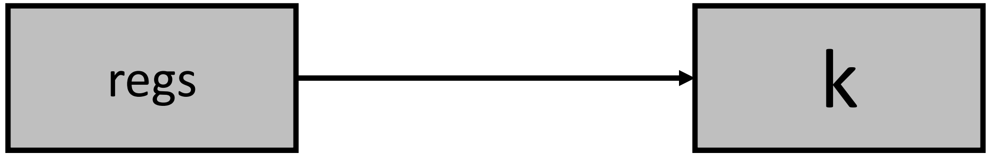


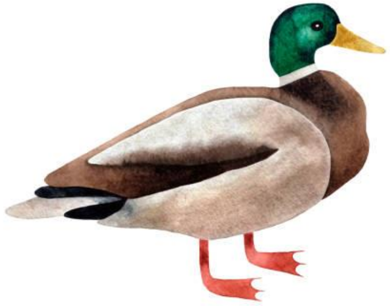
Further, I argue that we use them all the time subconsciously

Subconscious model

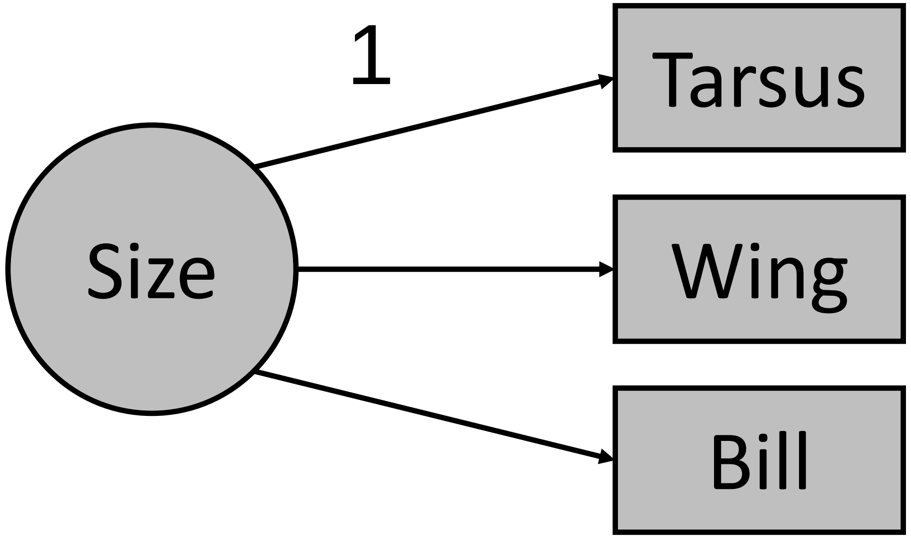


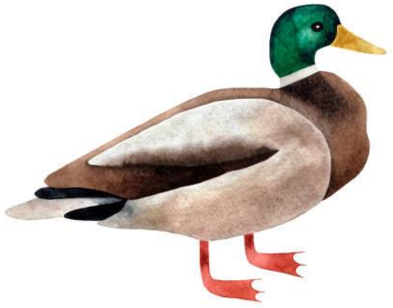
Actual model

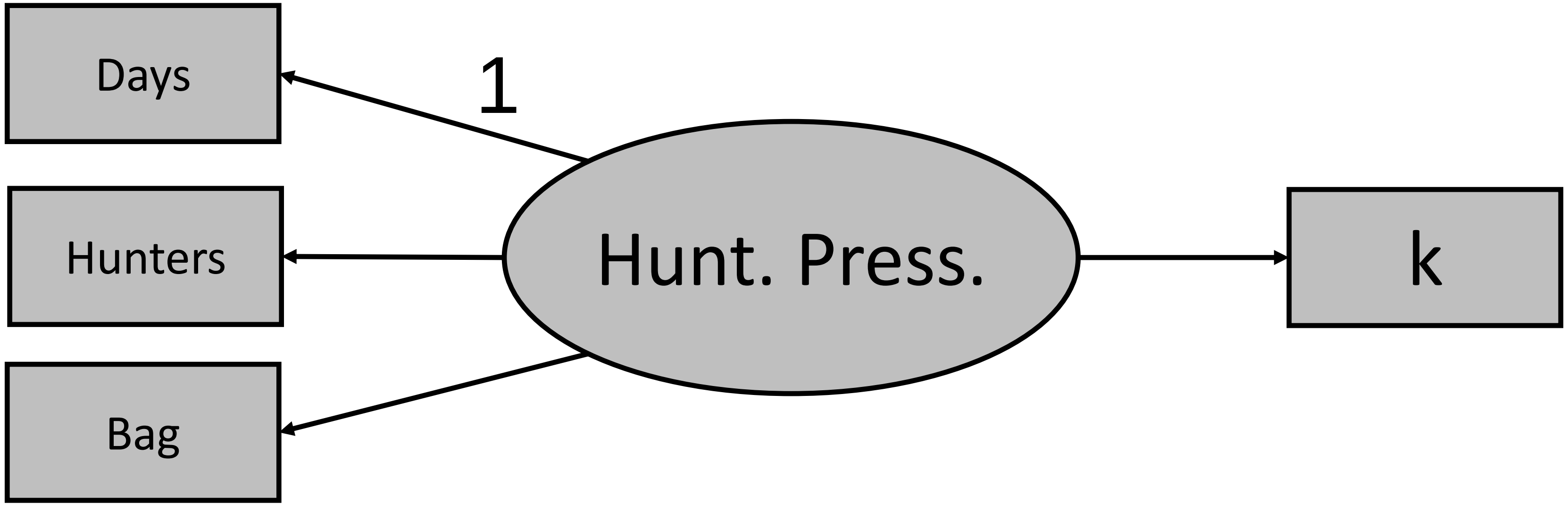




Arrow directionality





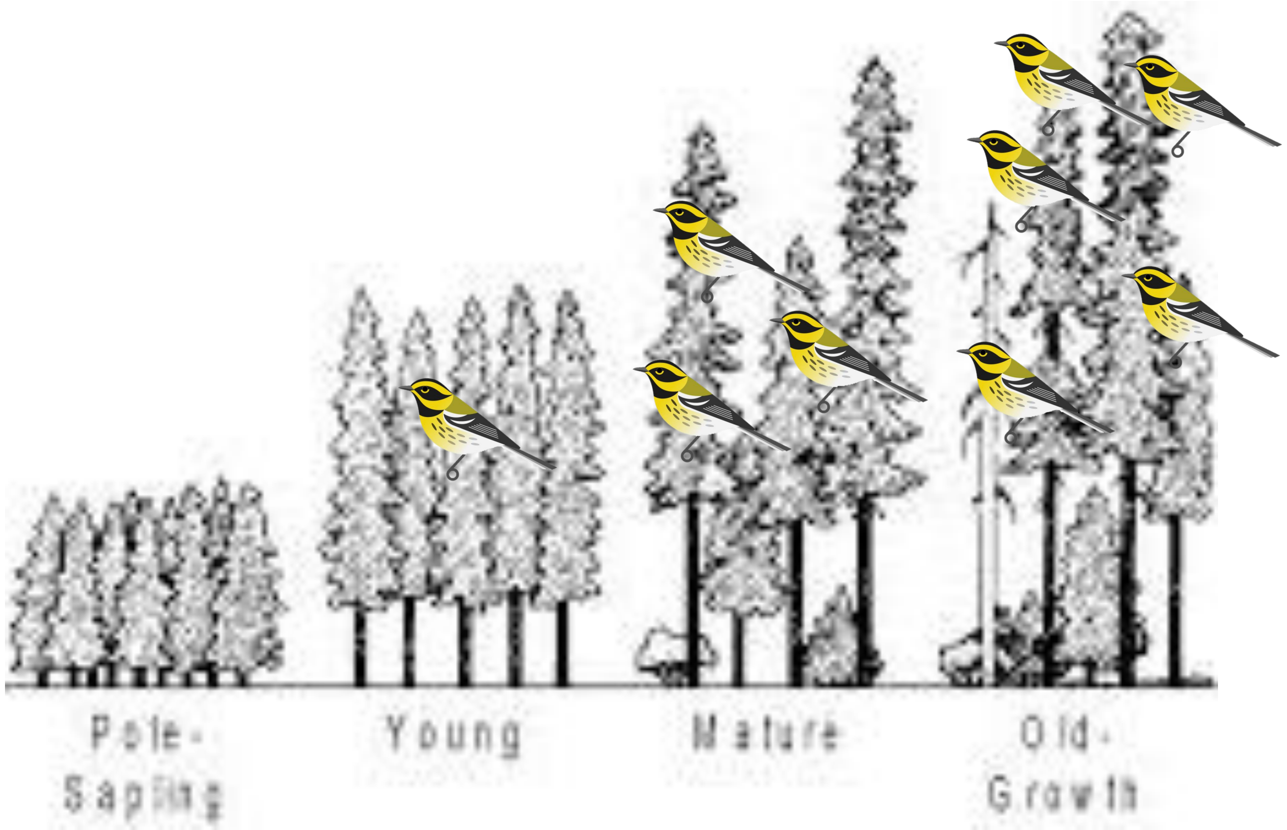


1. **Latent variables are intuitive, and we already informally use them all the time**

2. **They can be used to link multiple measurements of similar processes**

We can make some assumptions about our latent variable(s)

Our first example: forest age as a latent variable



We generally assume they're normally distributed

$$\boldsymbol{m} \sim \text{normal}(\mu, \sigma_m^2)$$

They're really kind of like random effects…

We assume that they are zero-centered b/c they're human constructs

i.e., what should the scale of forest maturity be?

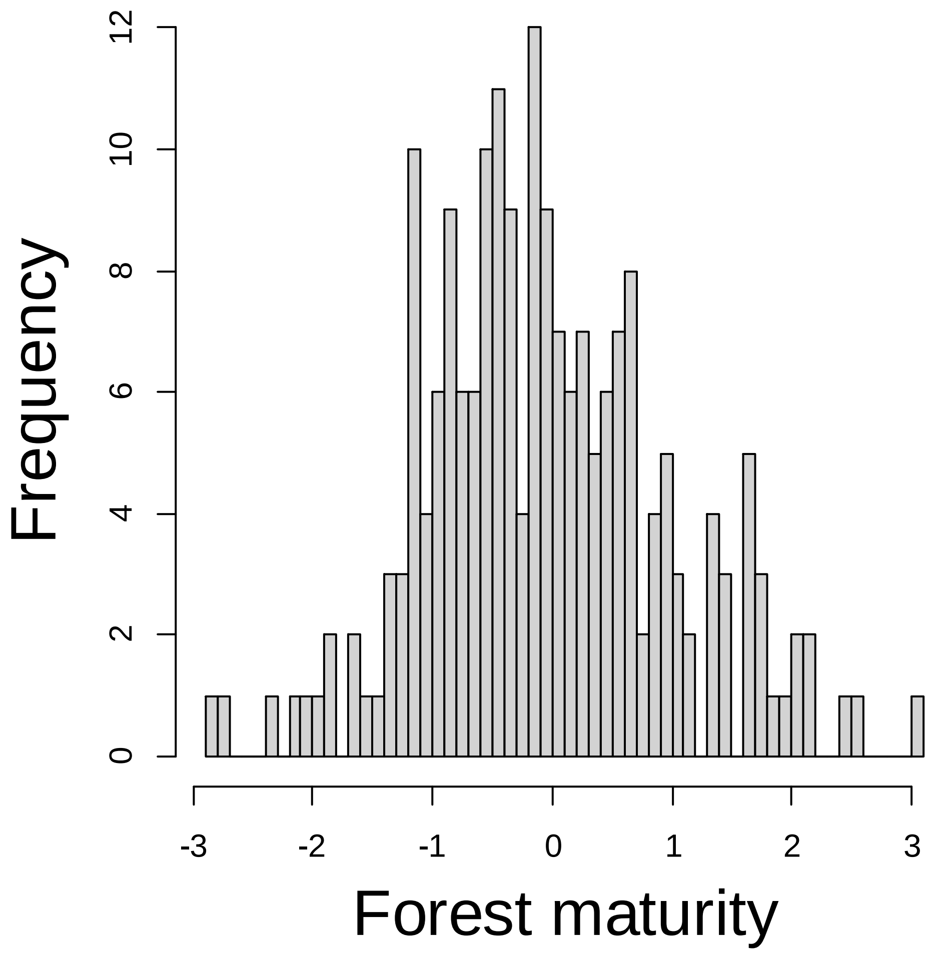
$$\boldsymbol{m} \sim \mathrm{normal}(0, \sigma_m^2)$$

Assigning an intercept would be entirely subjective, plus the math is easier if $\mu = 0$

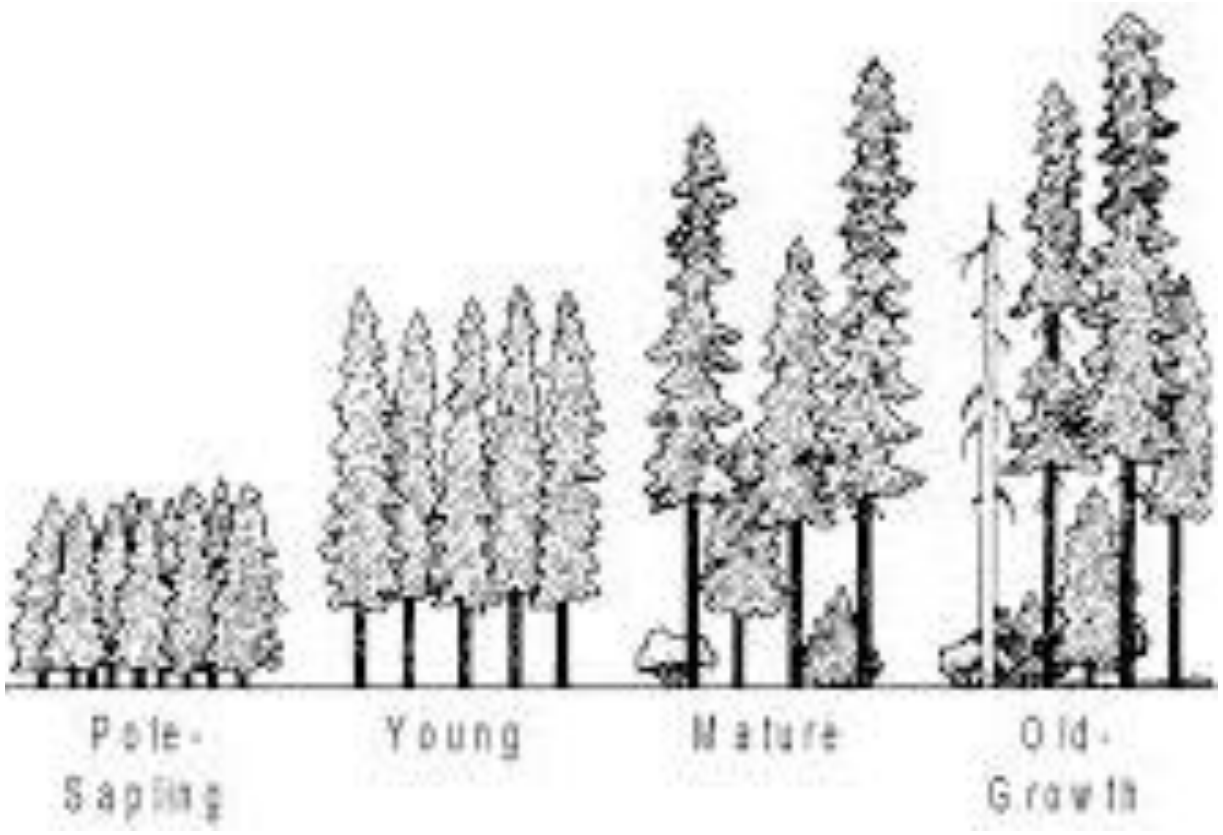
Let’s simulate some data

Data: counts (y) of ‘yellow-footed weeble-wobbles’ at sites with different canopy (c) and sub-canopy (s) heights

Step 1: simulate variation in forest maturity



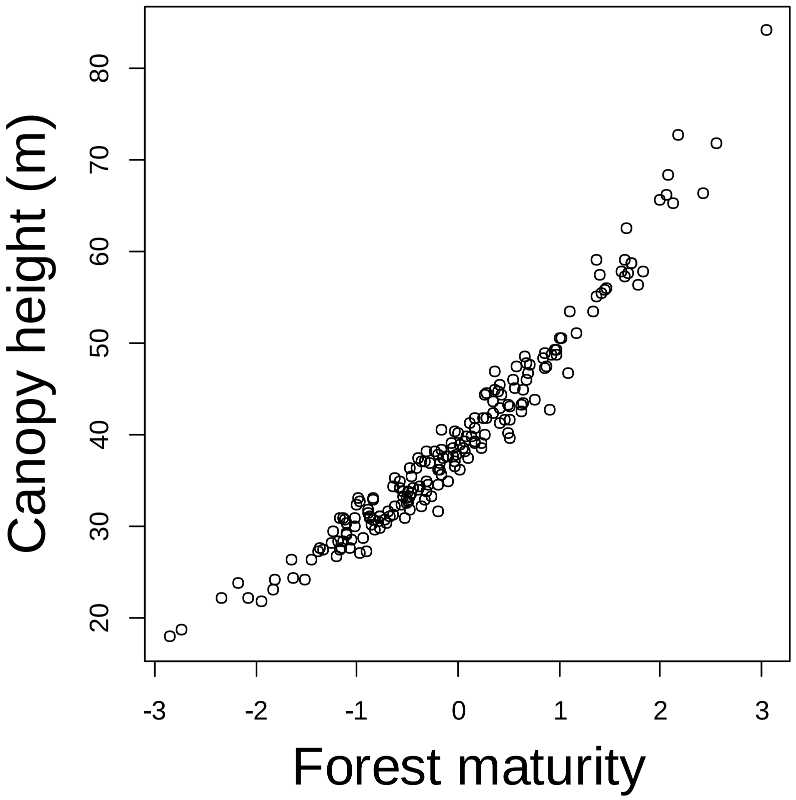
$$\boldsymbol{m} \sim \mathrm{normal}(0, 1)$$

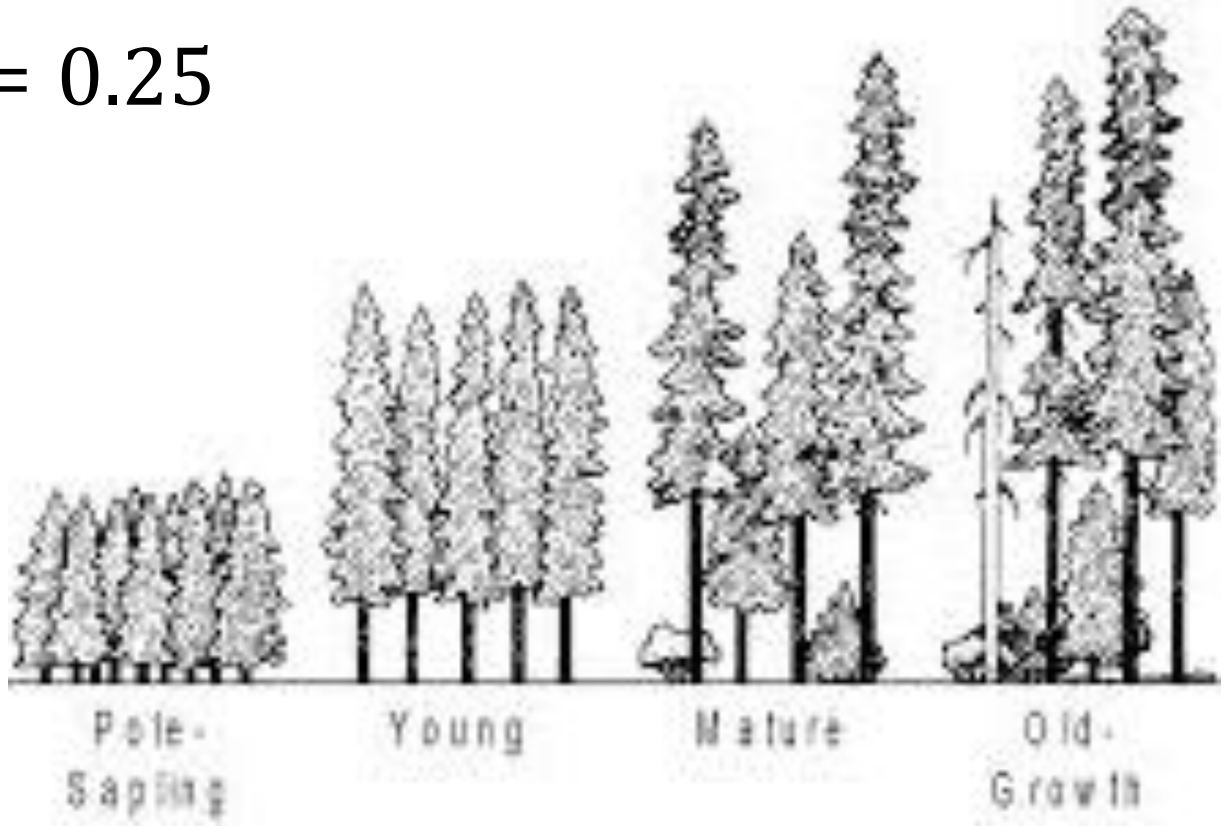


Step 2: simulate variation in canopy height (c)

$$\boldsymbol{c} \sim \text{lognormal}(\alpha_1 + \beta_1 \boldsymbol{m}, \sigma_c = 0.05)$$

$$\alpha_1 = 3.65$$
$$\beta_1 = 0.25$$

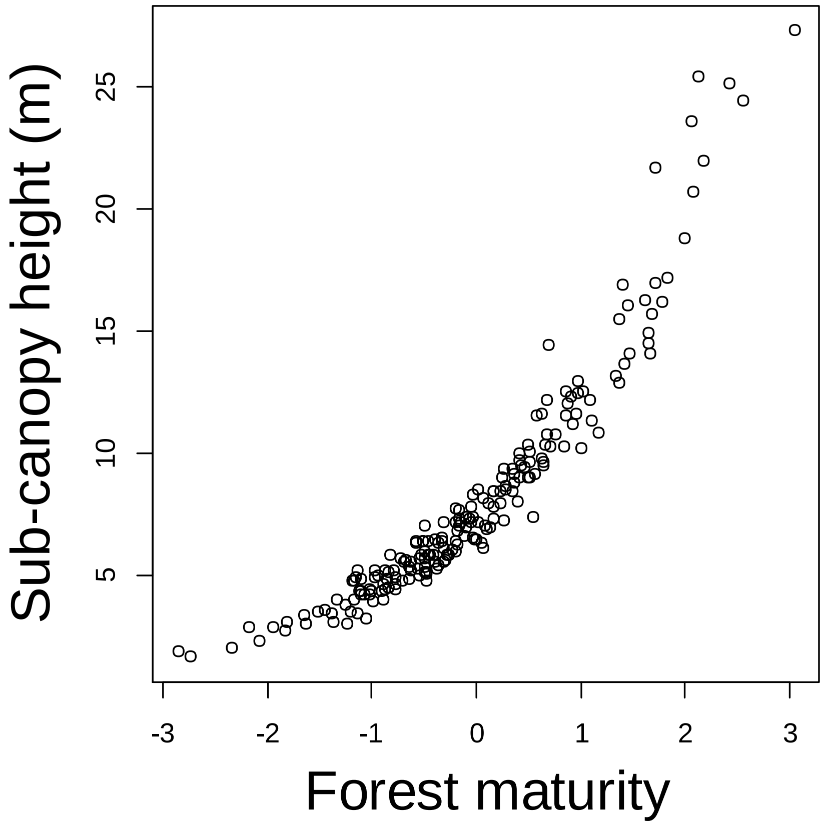


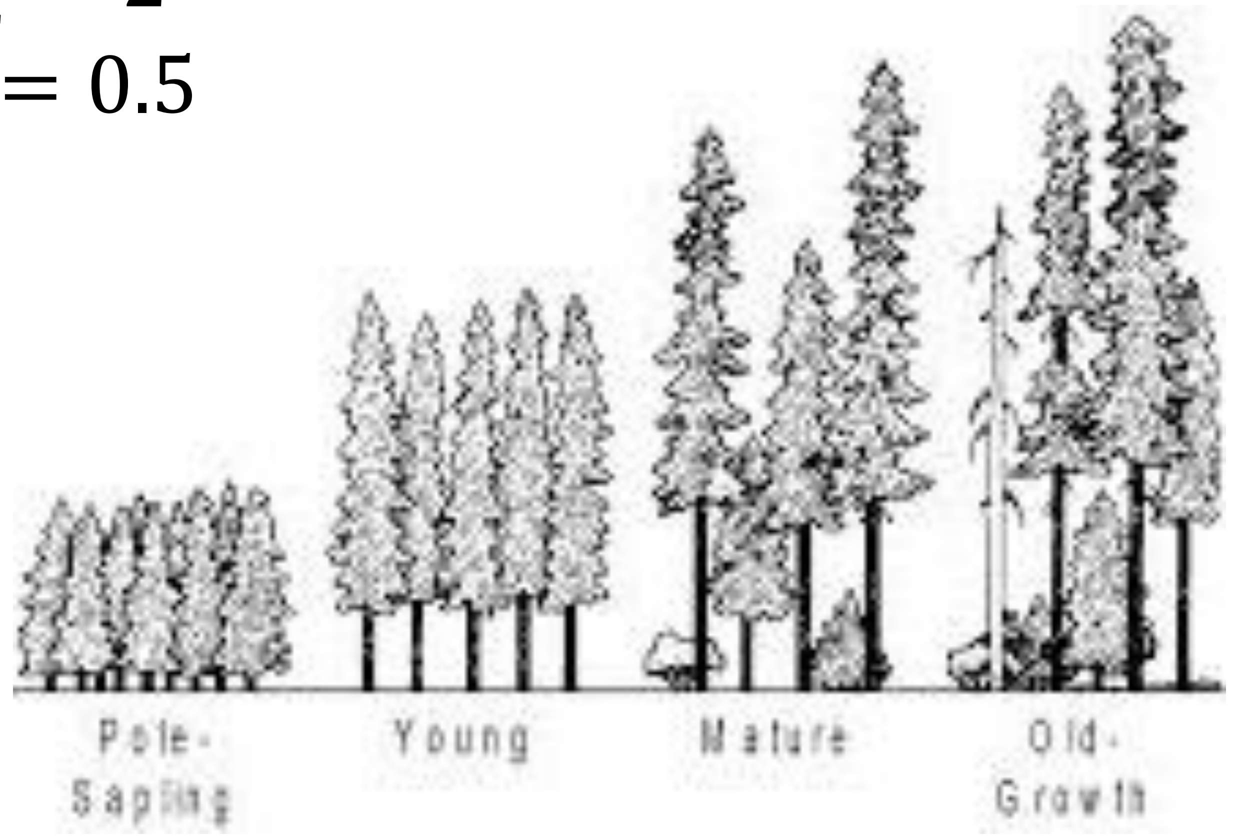


Step 3: simulate variation in sub-canopy height (s)

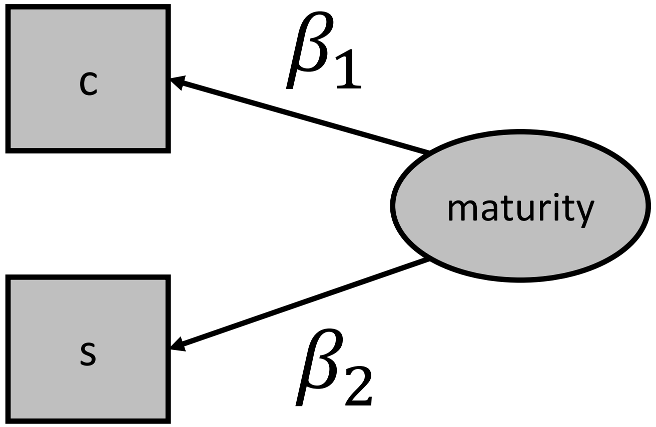
$$\boldsymbol{s} \sim \mathrm{lognormal}(\alpha_2 + \beta_2 \boldsymbol{m}, \sigma_s = 0.05)$$

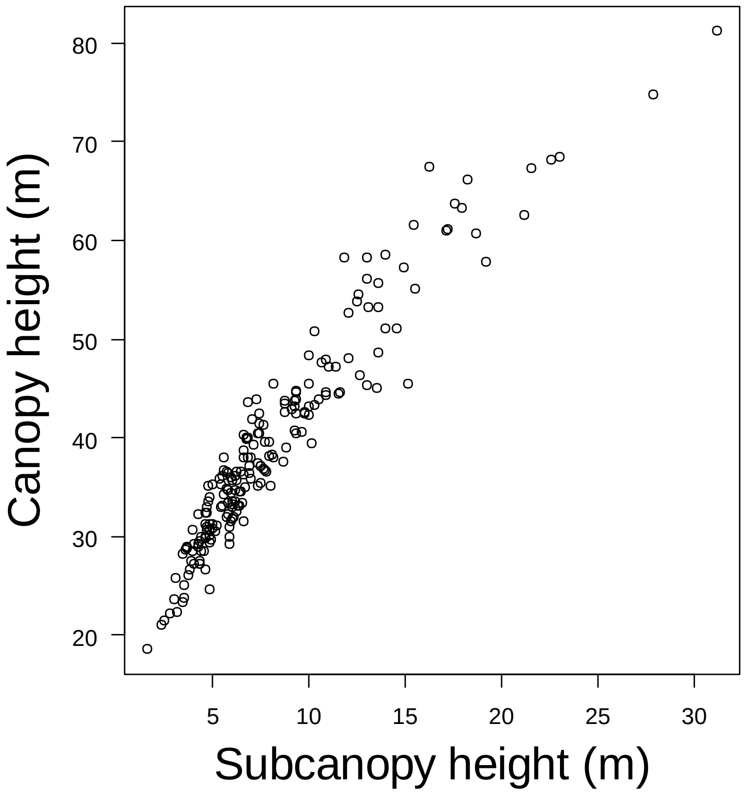
$$\alpha_2 = 2$$
$$\beta_2 = 0.5$$



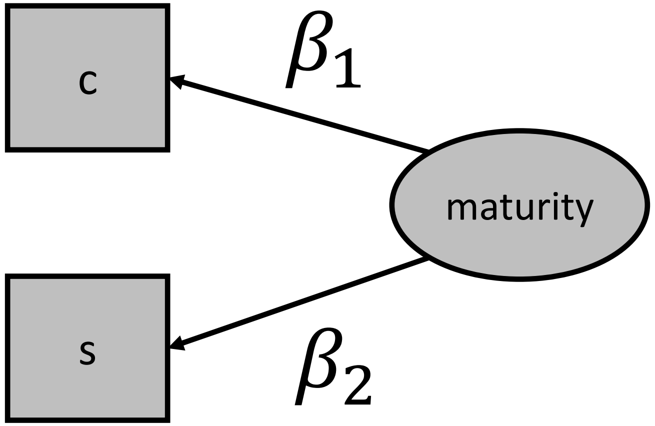


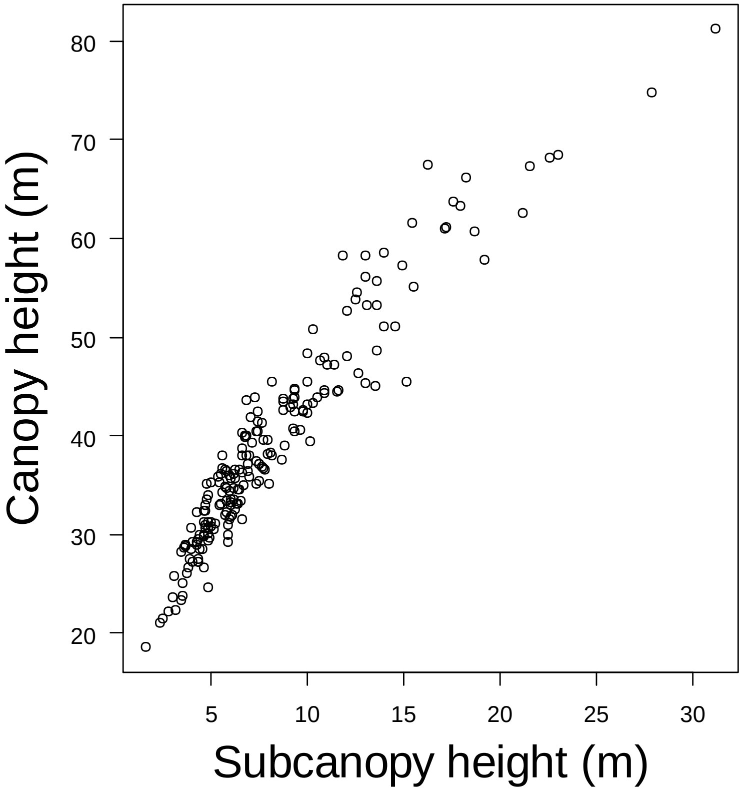
The hypothesis: older forests will have greater canopy heights and greater sub-canopy heights





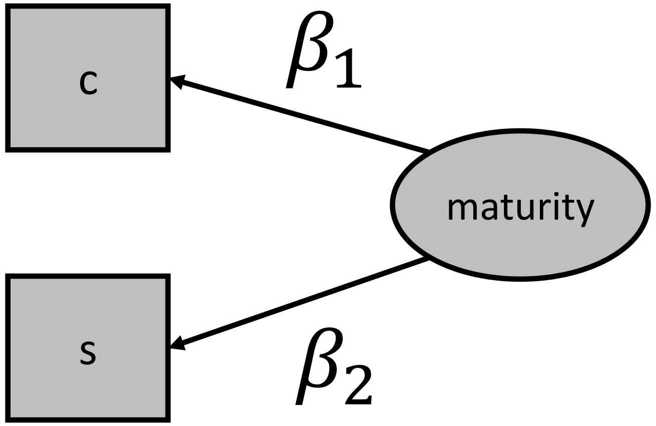
<u>The most important caveat</u>: if things aren't collinear, then you can't assign them to a latent variable





A note on drawing graphs

Squares or rectangles represent measured variables



This oval represents a latent variable

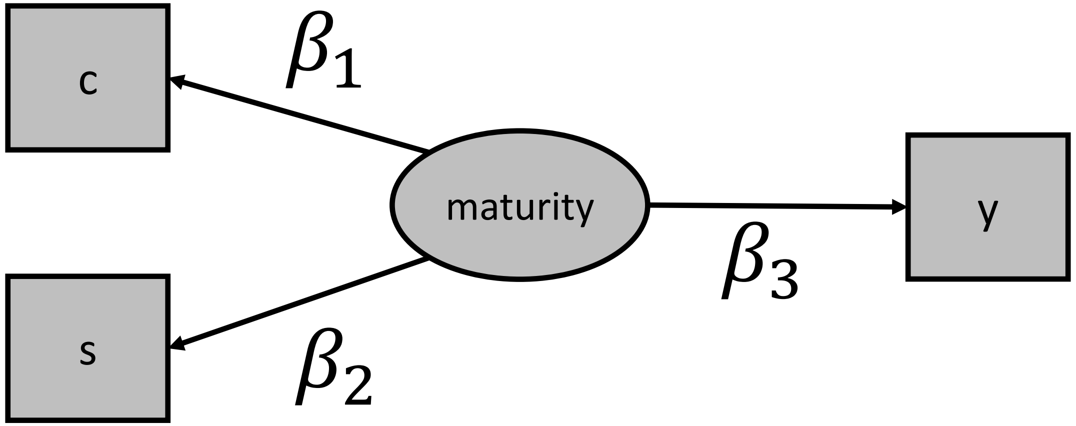
Arrows represent paths (linear models). The direction of the arrow indicates how to parameterize the relationship

Step 4: simulate variation in warbler counts (y)

$$\boldsymbol{y} \sim \text{Poisson}(e^{\alpha_3 + \beta_3 \boldsymbol{m}})$$

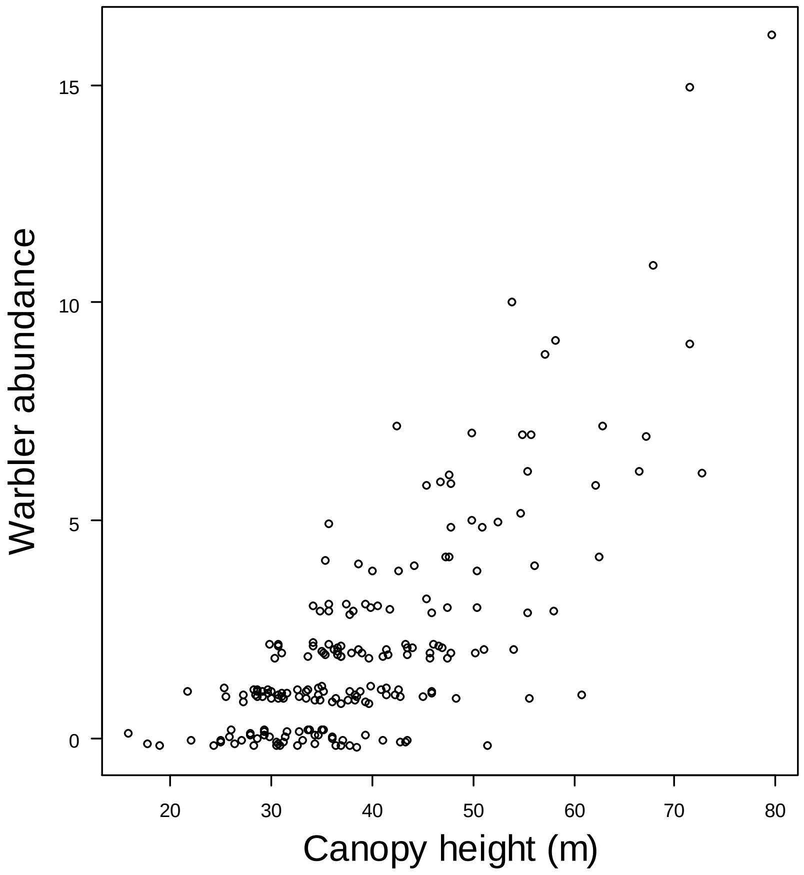
$$\alpha_3 = 0.5$$

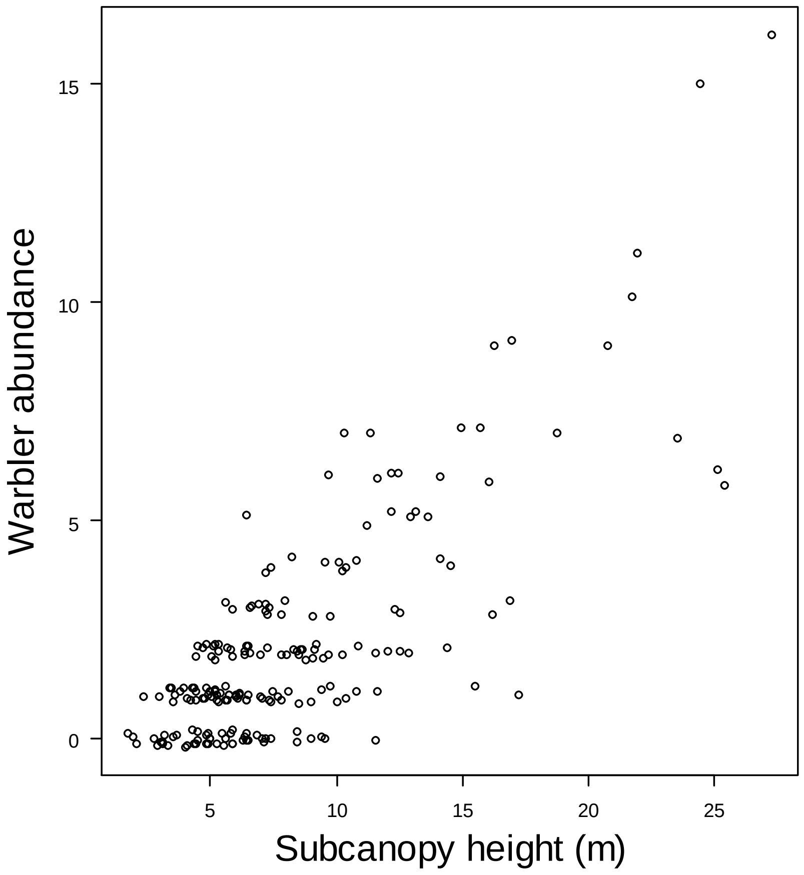
$$\beta_3 = 0.75$$

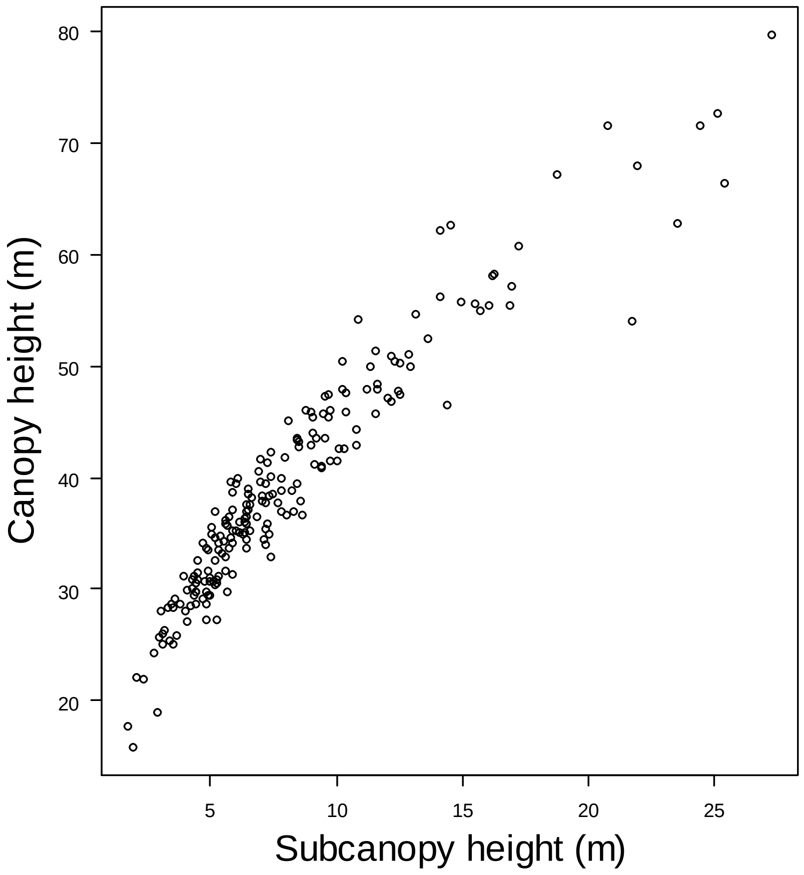


Step 4: simulate variation in warbler counts (y)

$$\boldsymbol{y} \sim \text{Poisson}(e^{\alpha_3 + \beta_3 \boldsymbol{m}})$$







The ecological hypothesis: older forests will have more birds

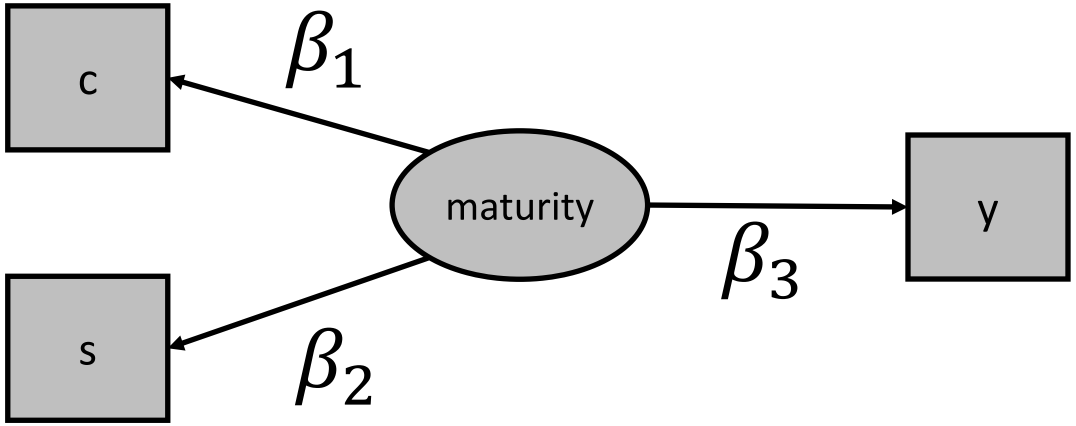
Our (first) model

$$\boldsymbol{m} \sim \mathrm{normal}(0, \sigma_m^2)$$

$$\boldsymbol{c} \sim \mathrm{normal}(\alpha_1 + \beta_1 \boldsymbol{m}, \sigma_c^2)$$

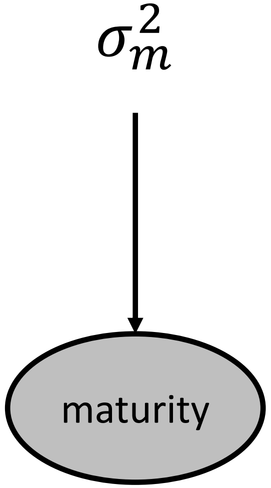
$$\boldsymbol{s} \sim \mathrm{normal}(\alpha_2 + \beta_2 \boldsymbol{m}, \sigma_s^2)$$

$$\boldsymbol{y} \sim \mathrm{normal}(\alpha_3 + \beta_3 \boldsymbol{m}, \sigma_y^2)$$



Our (first) model

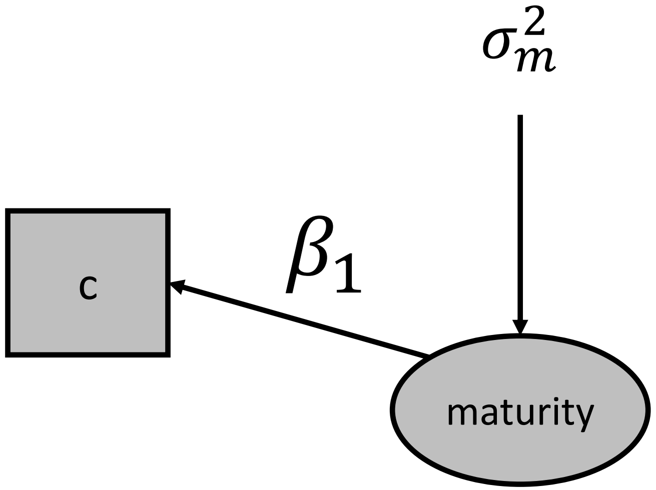
$$\boldsymbol{m} \sim \text{normal}(0, \sigma_m^2)$$



Our (first) model

$$\boldsymbol{m} \sim \text{normal}(0, \sigma_m^2)$$

$$\boldsymbol{c} \sim \text{normal}(\alpha_1 + \beta_1 \boldsymbol{m}, \sigma_c^2)$$

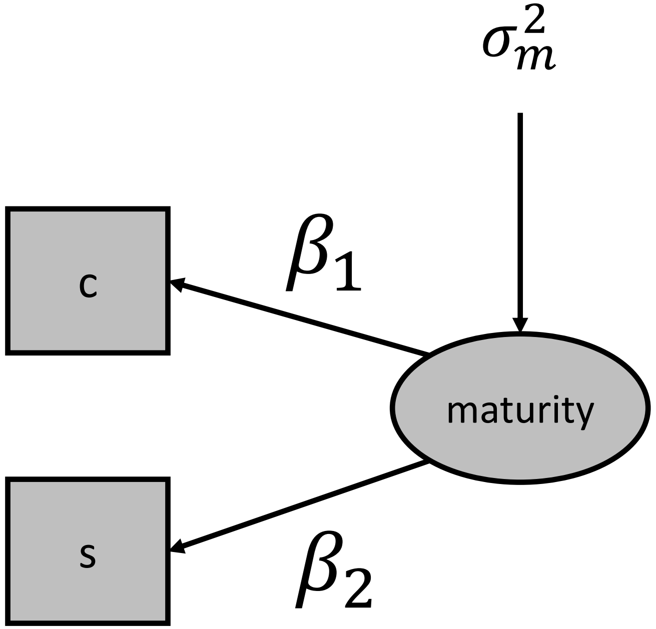


Our (first) model

$$\boldsymbol{m} \sim \text{normal}(0, \sigma_m^2)$$

$$\boldsymbol{c} \sim \text{normal}(\alpha_1 + \beta_1 \boldsymbol{m}, \sigma_c^2)$$

$$\boldsymbol{s} \sim \text{normal}(\alpha_2 + \beta_2 \boldsymbol{m}, \sigma_s^2)$$



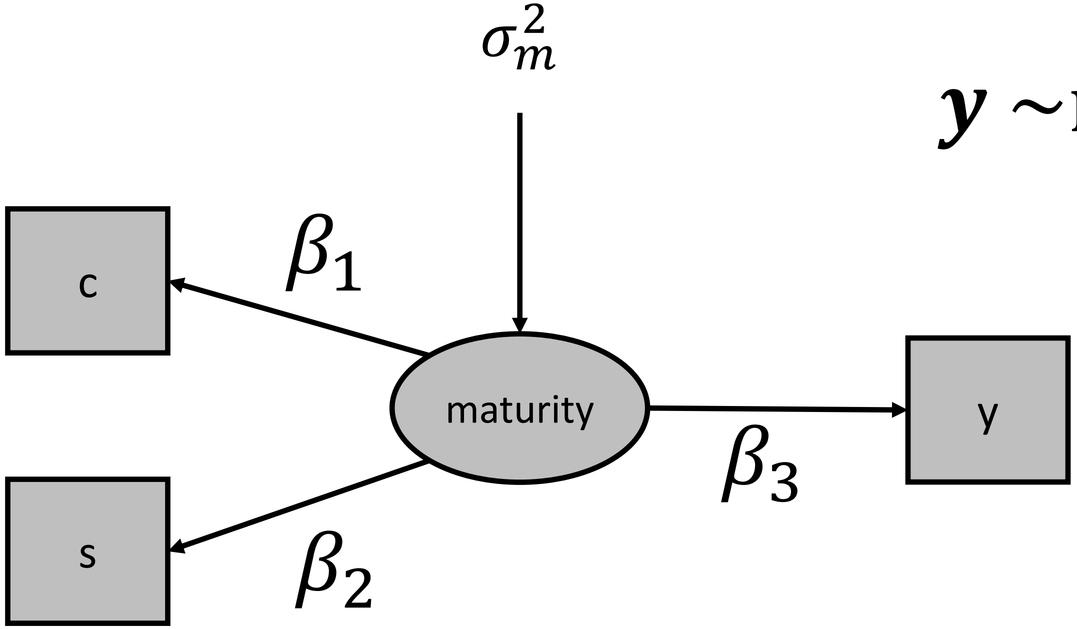
Our (first) model

$$\boldsymbol{m} \sim \text{normal}(0, \sigma_m^2)$$

$$\boldsymbol{c} \sim \text{normal}(\alpha_1 + \beta_1 \boldsymbol{m}, \sigma_c^2)$$

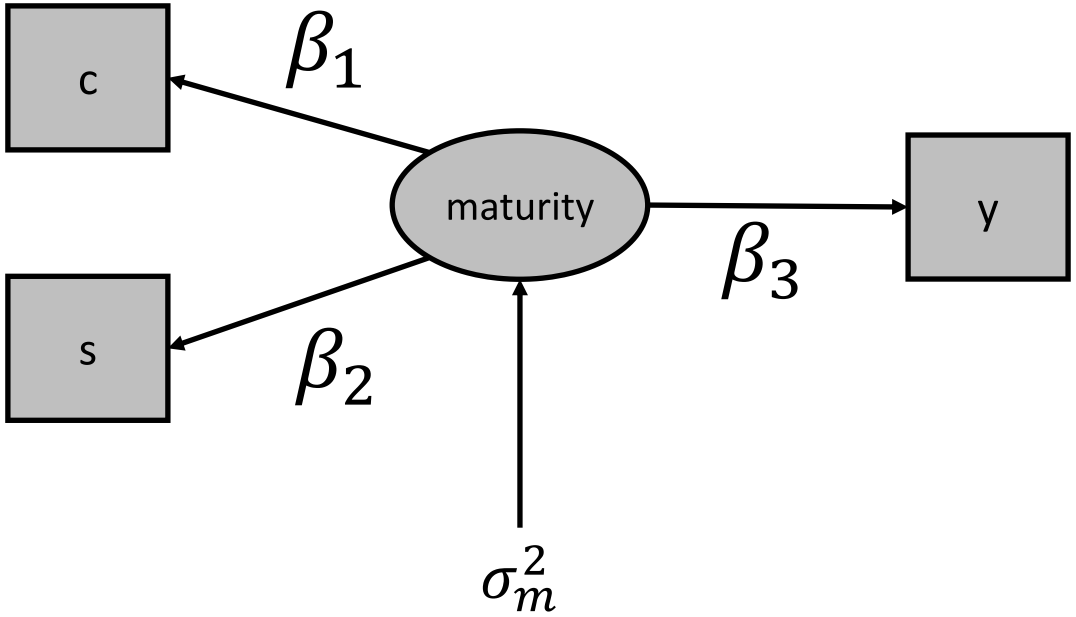
$$\boldsymbol{s} \sim \text{normal}(\alpha_2 + \beta_2 \boldsymbol{m}, \sigma_s^2)$$

$$\boldsymbol{y} \sim \text{normal}(\alpha_3 + \beta_3 \boldsymbol{m}, \sigma_y^2)$$



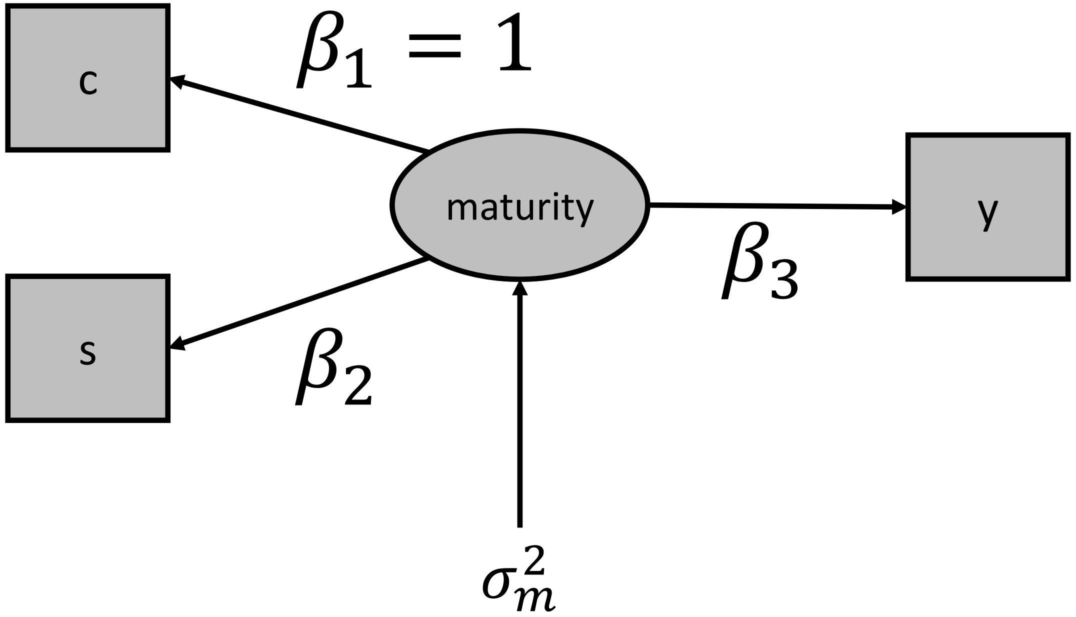
There is one very non-intuitive thing to discuss

We must fix a ‘loading’ to 1



There is one very non-intuitive thing to discuss

We must fix a ‘loading’ to 1



Why?!

Why?!

Well, so the model will be identifiable…

What are the implications of that?

What are the implications of that?

1. The latent variable will be on the same scale as whatever path we fix = 1

What are the implications of that?

1. The latent variable will be on the same scale as whatever path we fix = 1.

2. Our estimates of parameter relationships will be a function of that scale.

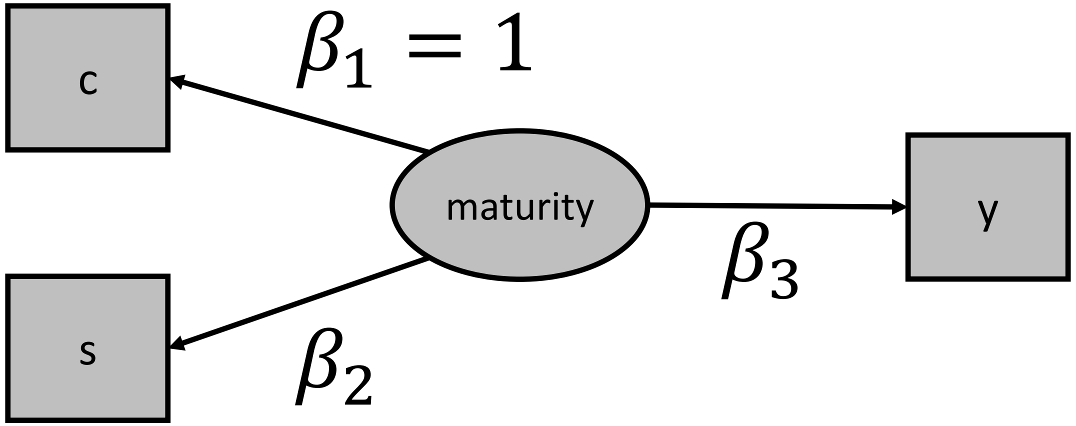
What are the implications of that?

1. The latent variable will be on the same scale as whatever path we fix = 1

2. Our estimates of parameter relationships will be a function of that scale.

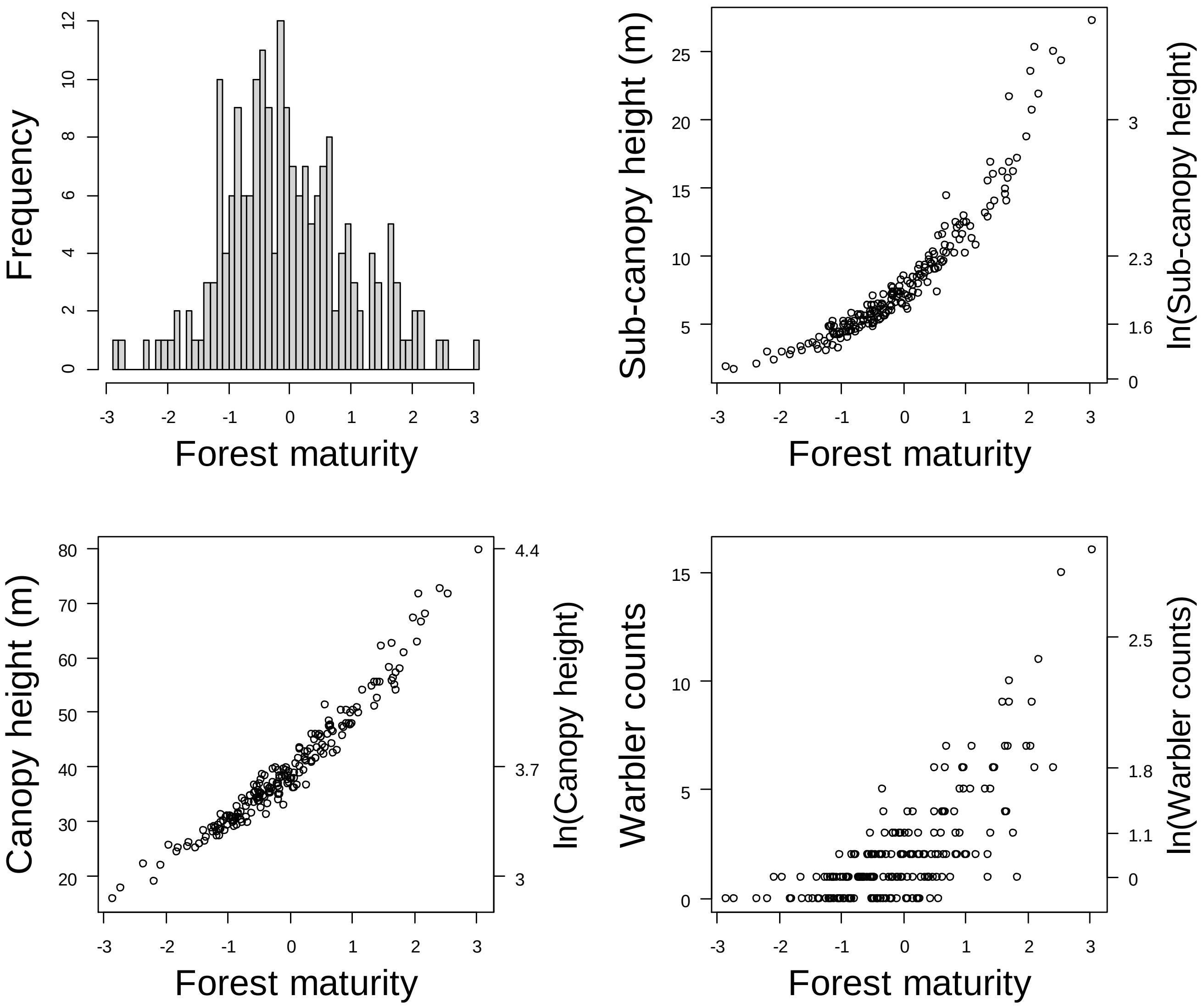
3. That’s it. It won’t change our predictions (i.e., warbler counts)

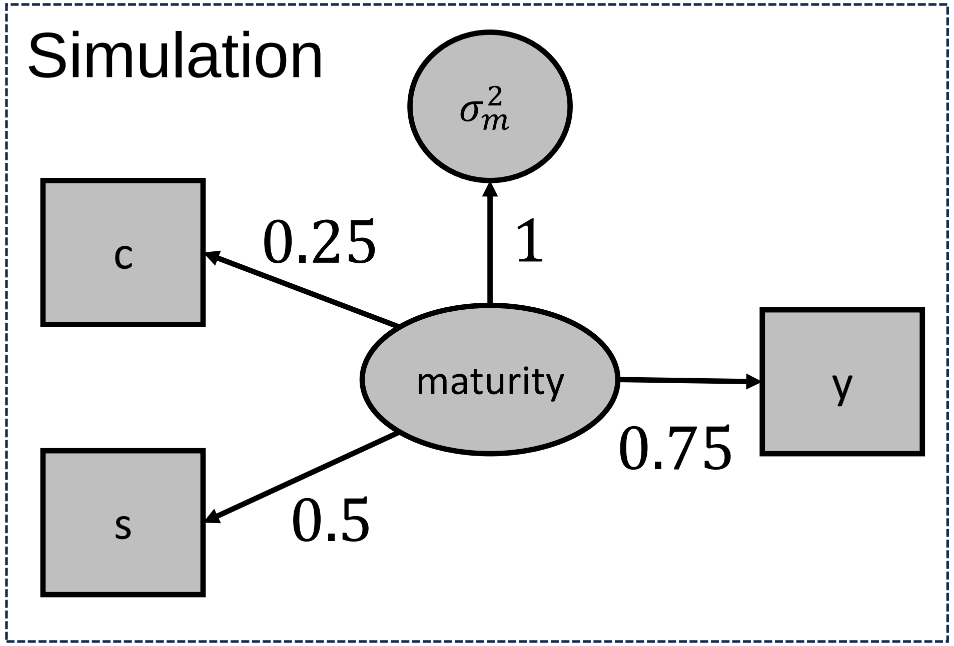
All we're really assuming when we fix that beta is that there is a positive relationship between our latent variable and the measured variable



So, let’s talk about this ‘fixing a loading to 1’ thing

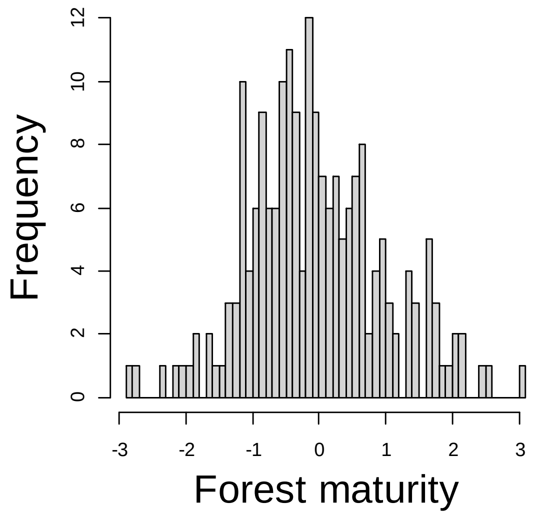
Let’s simulate some data

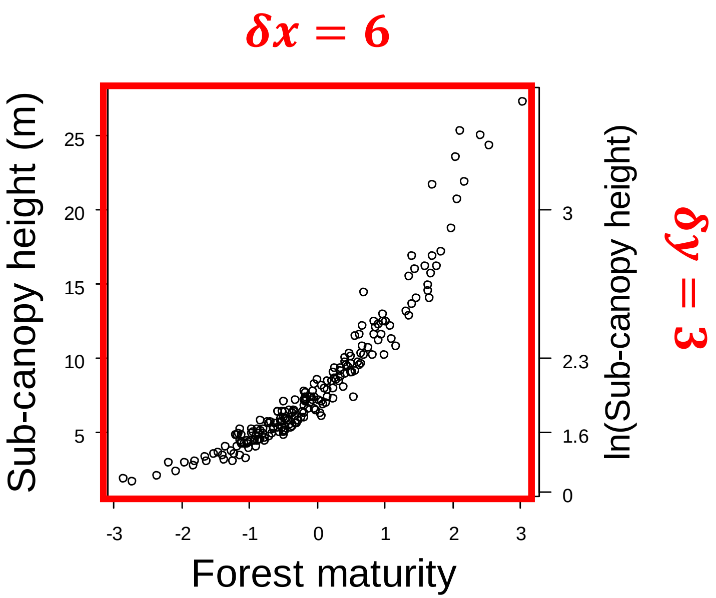


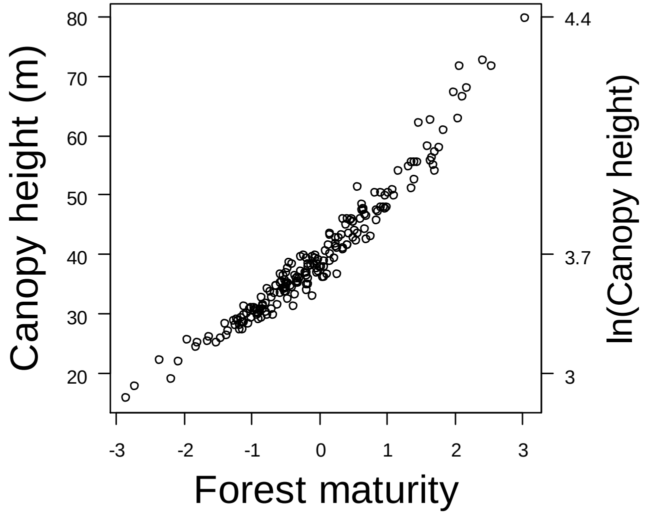


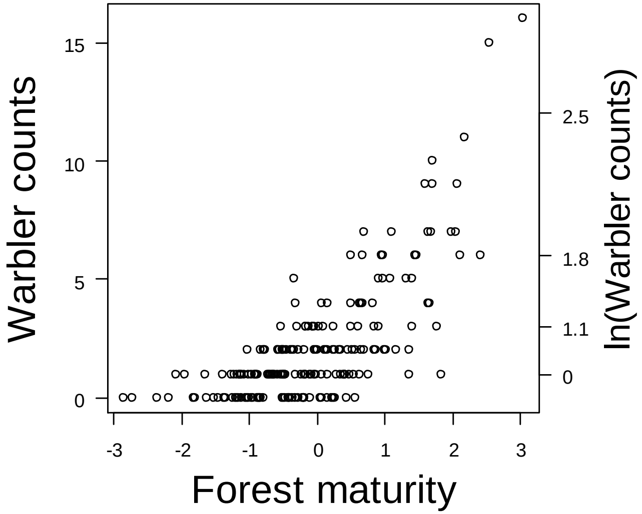
$$\delta y = \delta x \beta$$

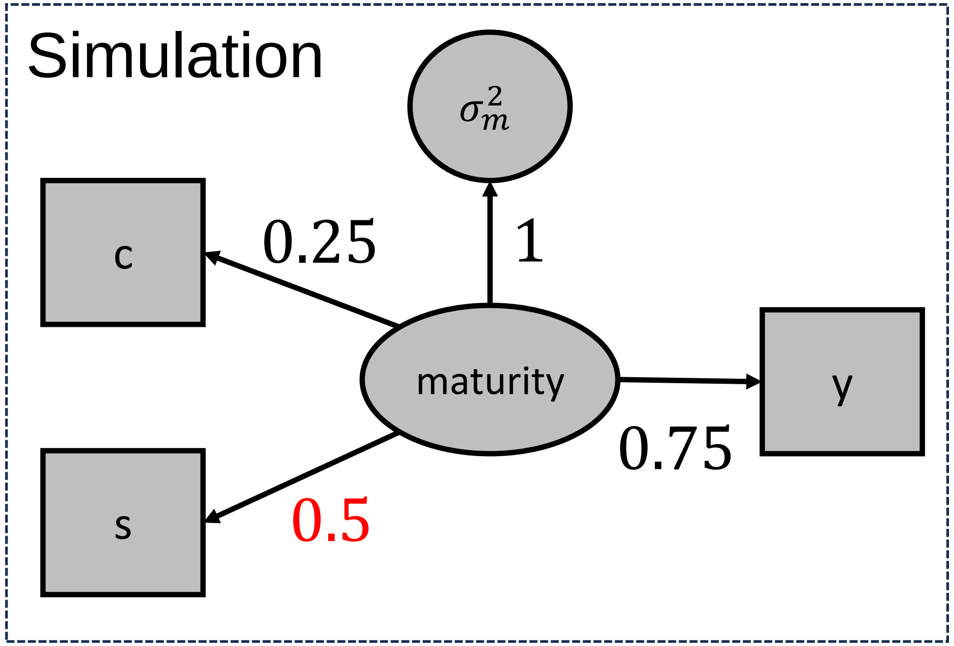
Clarifying a loading…







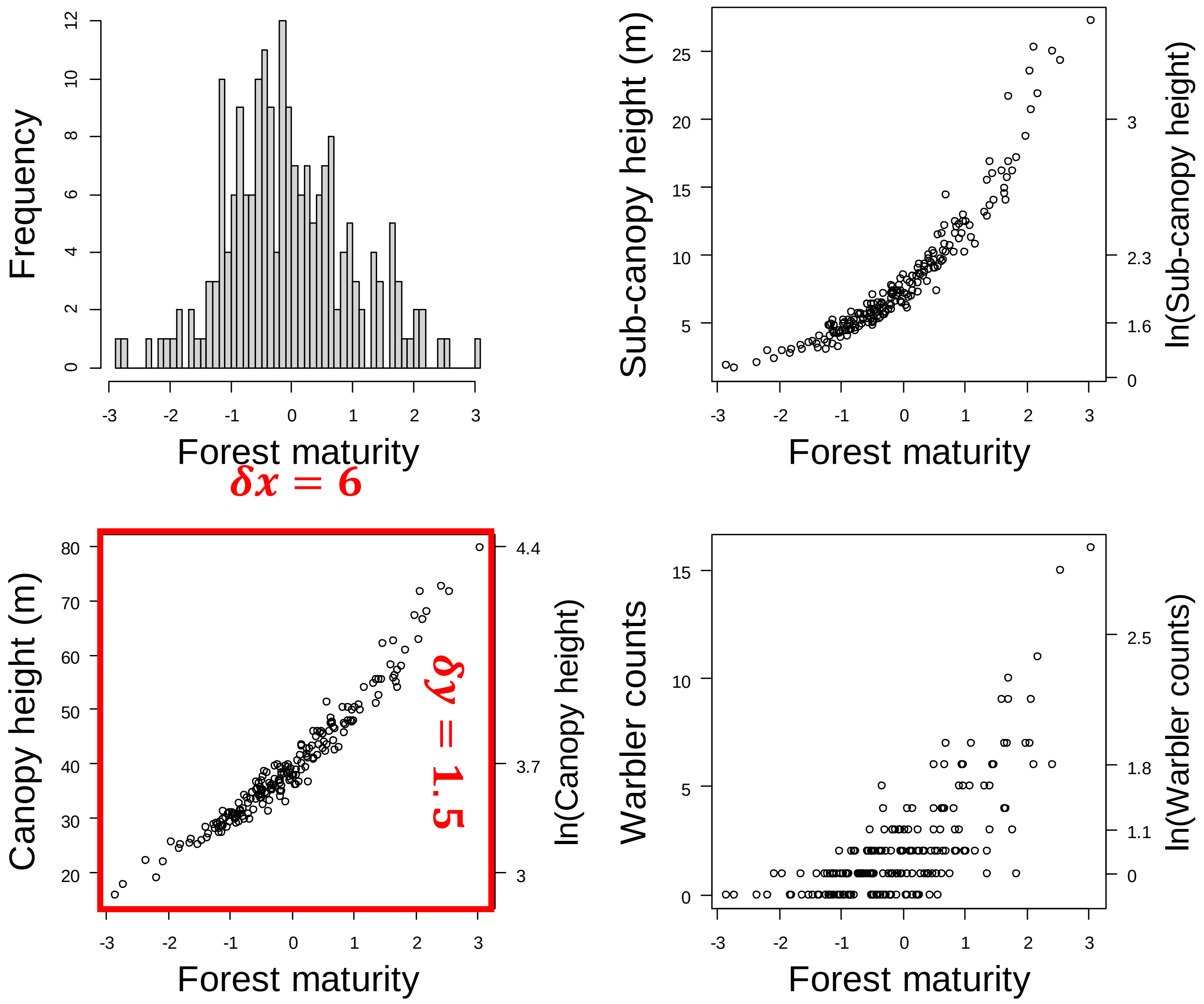


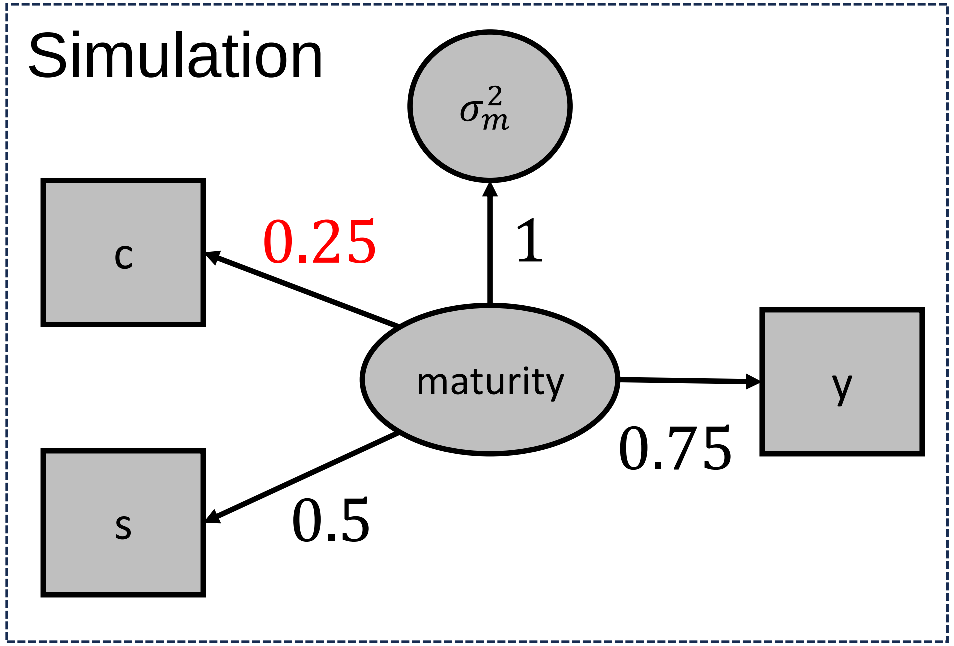


$$\delta y = \delta x \beta$$

$$\beta = 0.5$$

Clarifying a loading…

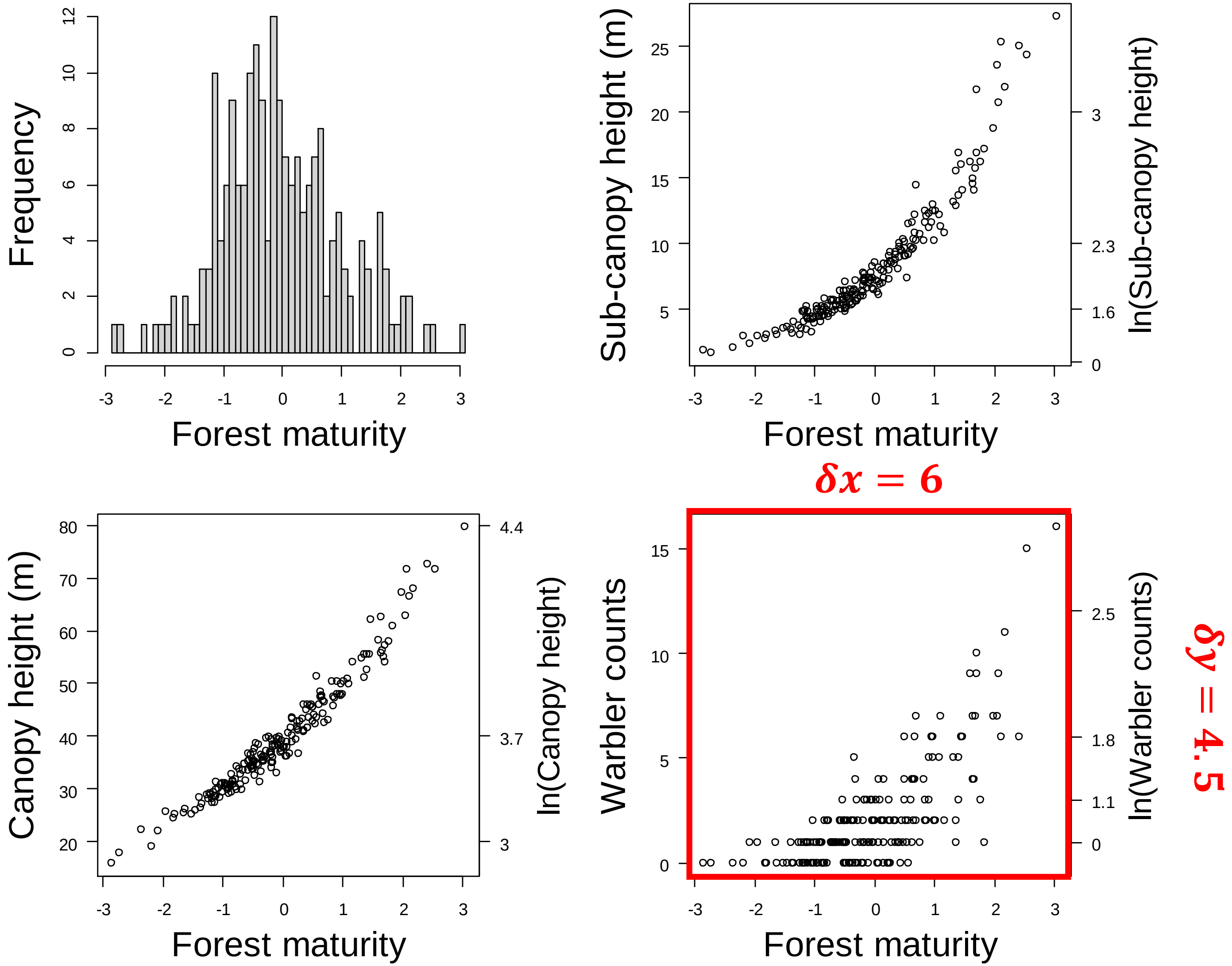


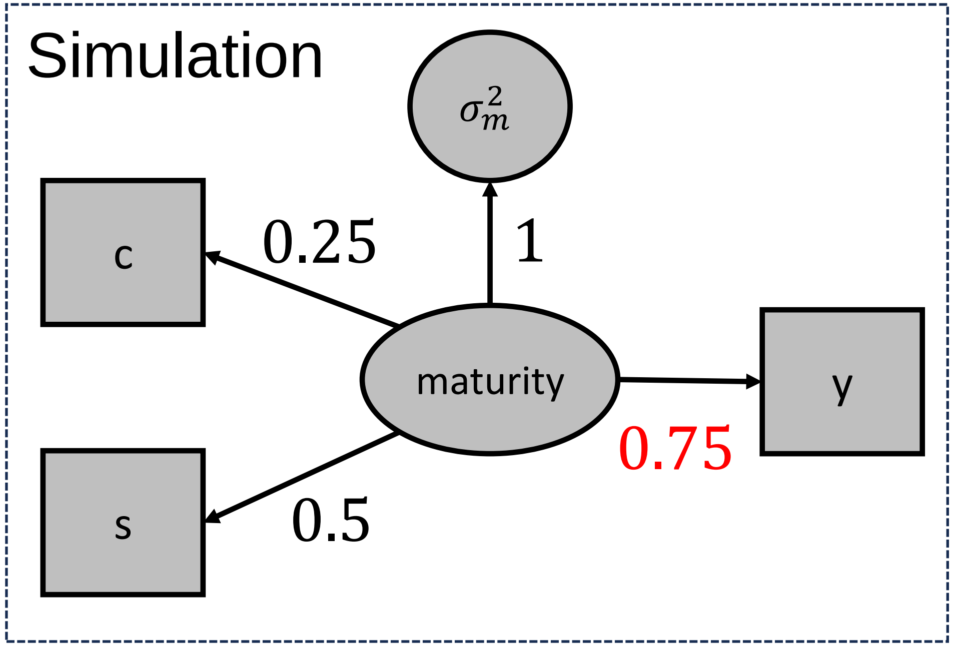


$$\delta y = \delta x \beta$$

$$\beta = 0.25$$

Clarifying a loading…





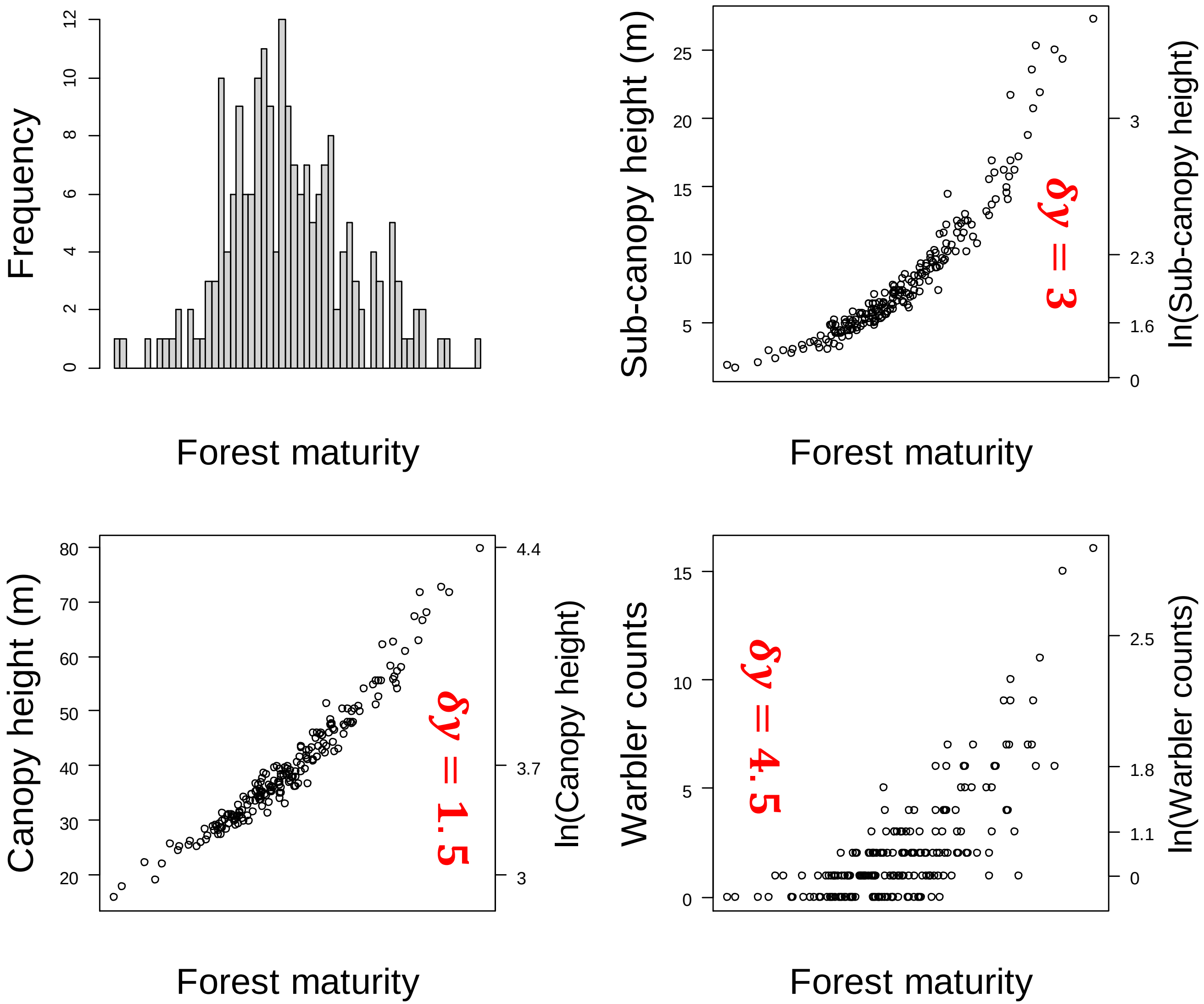
$$\delta y = \delta x \beta$$

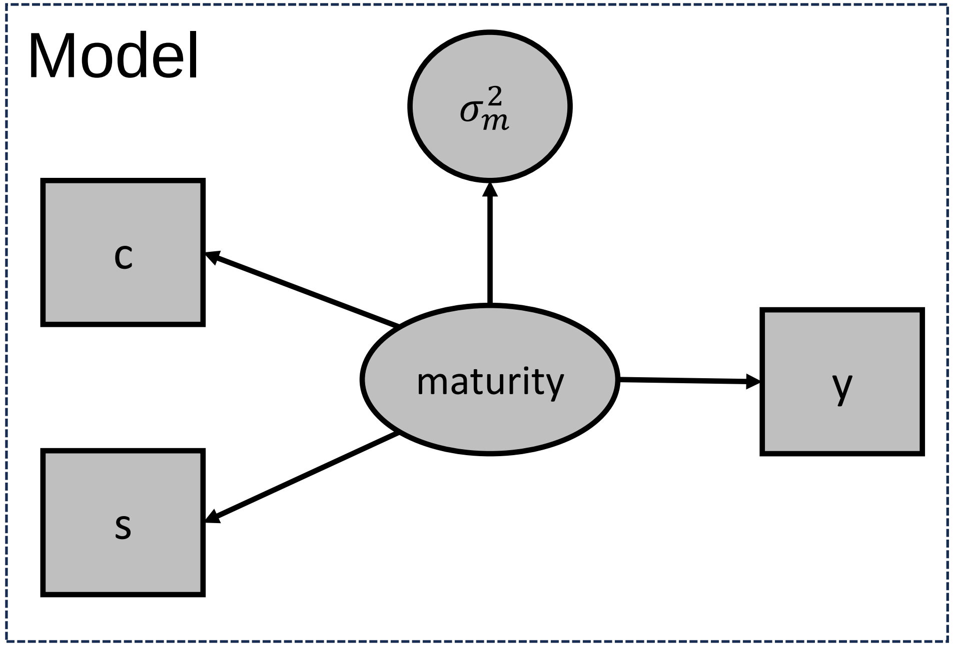
$$\beta = 0.75$$

Our latent variable is unobservable…

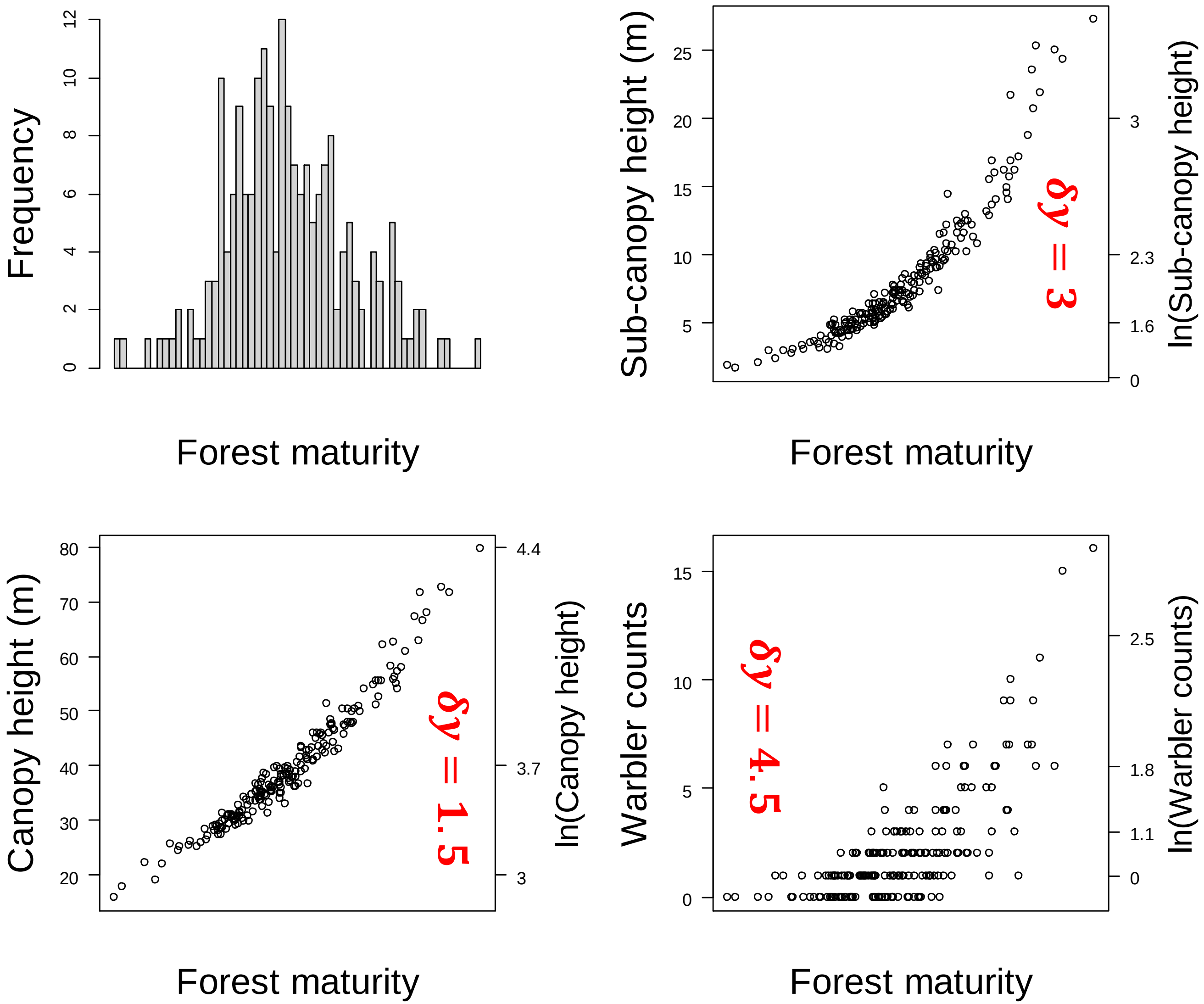
We don't know its scale…

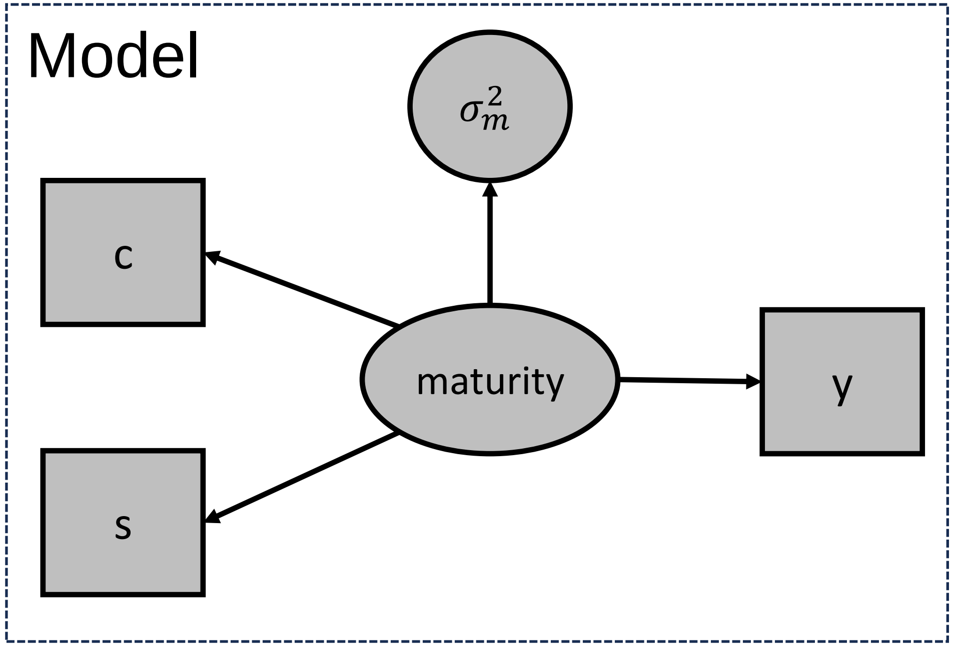
Clarifying a loading…





Shoot…

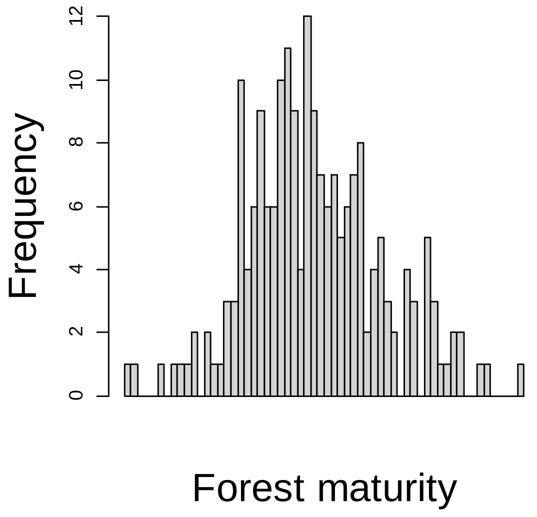


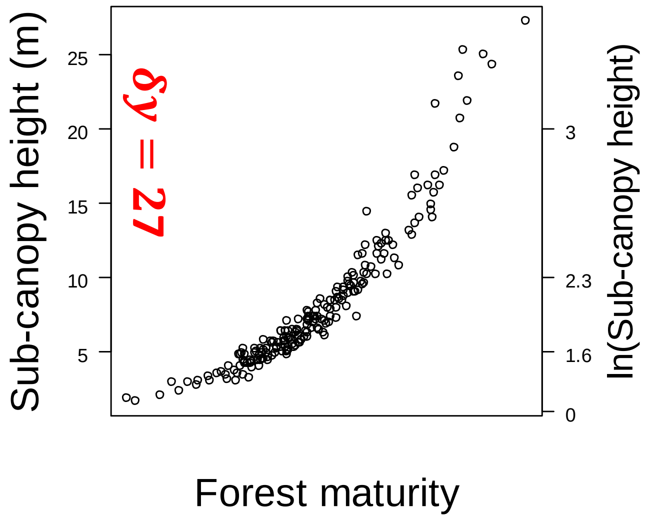


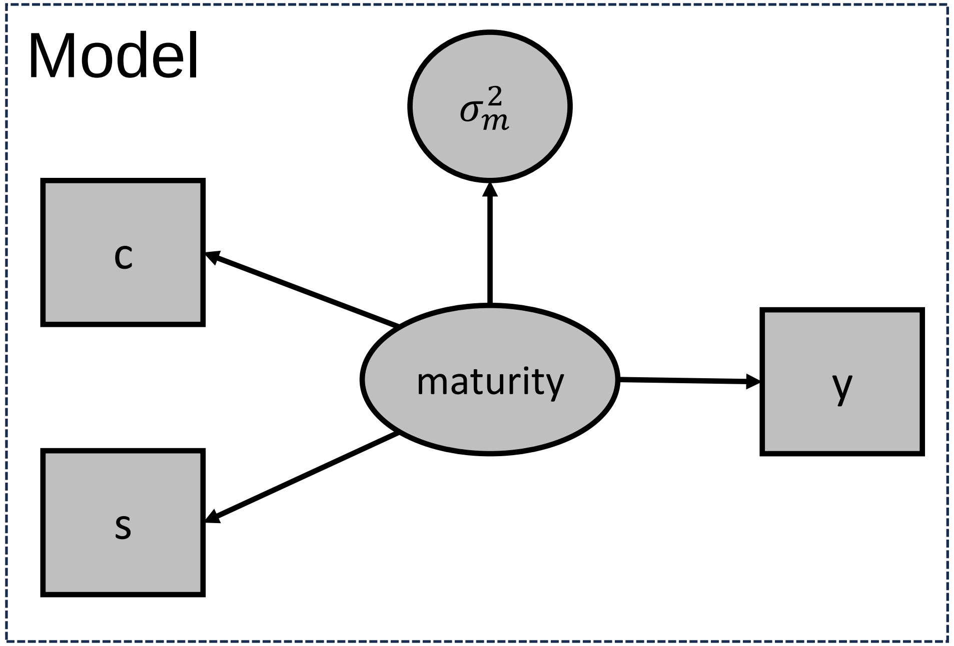
There's a big problem:
We don't know the range of maturity

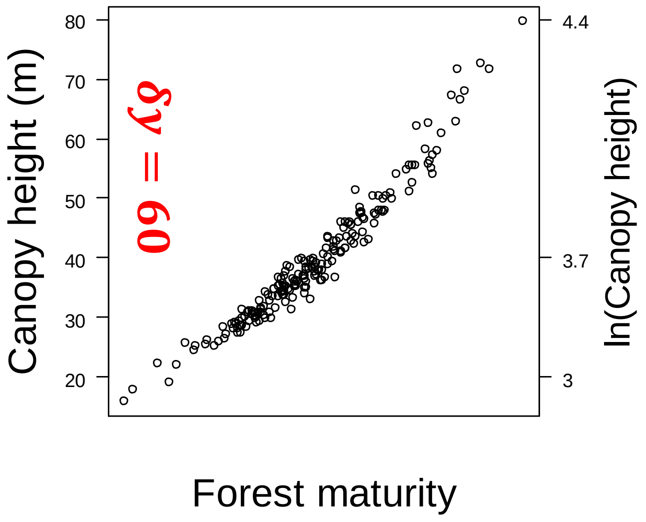
$$\boldsymbol{m} \sim \mathrm{normal}(0, \sigma_m^2)$$

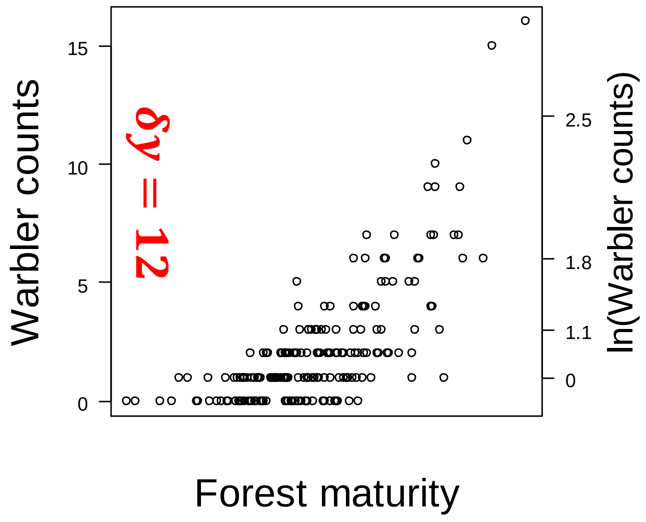
Small groups!







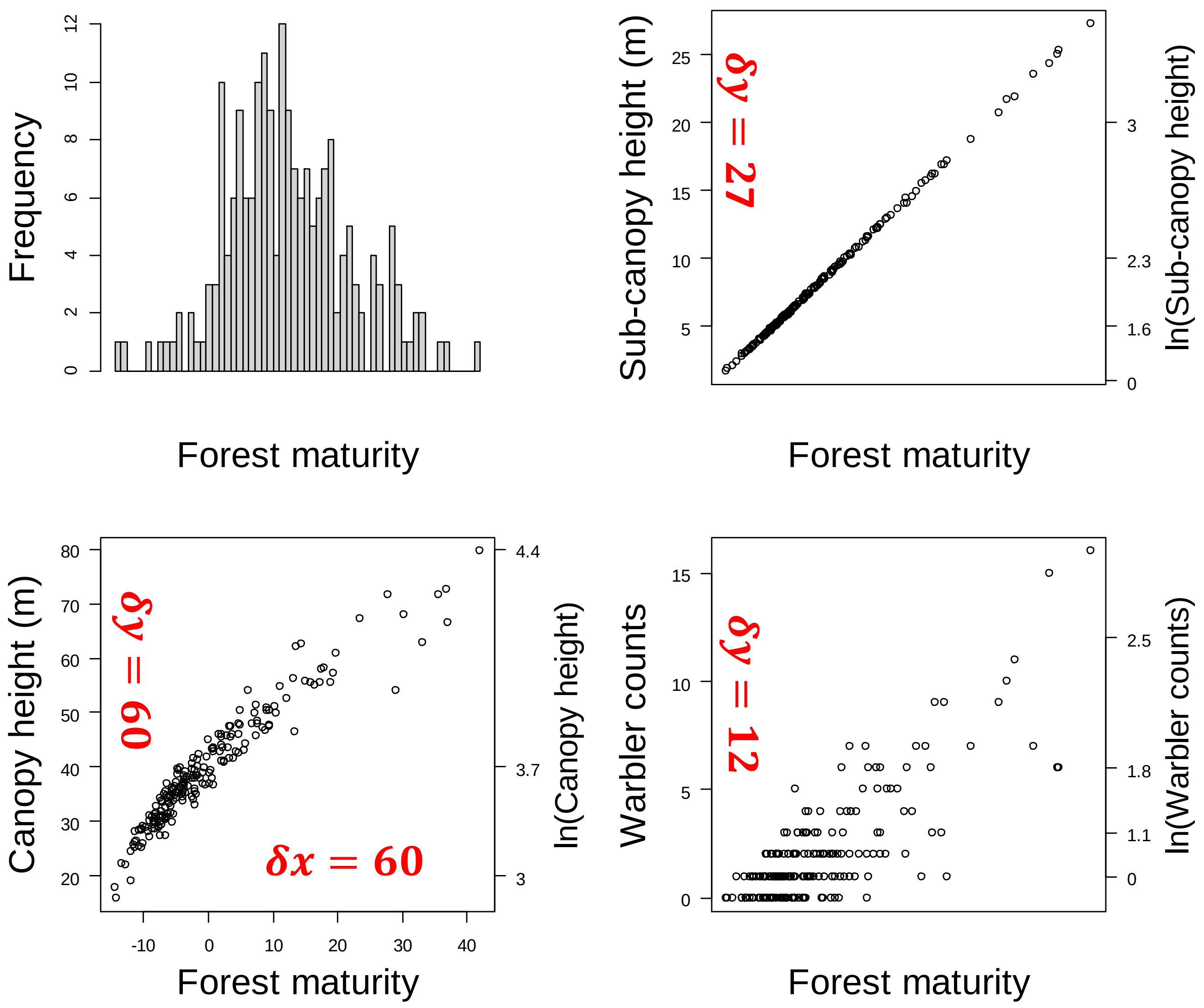


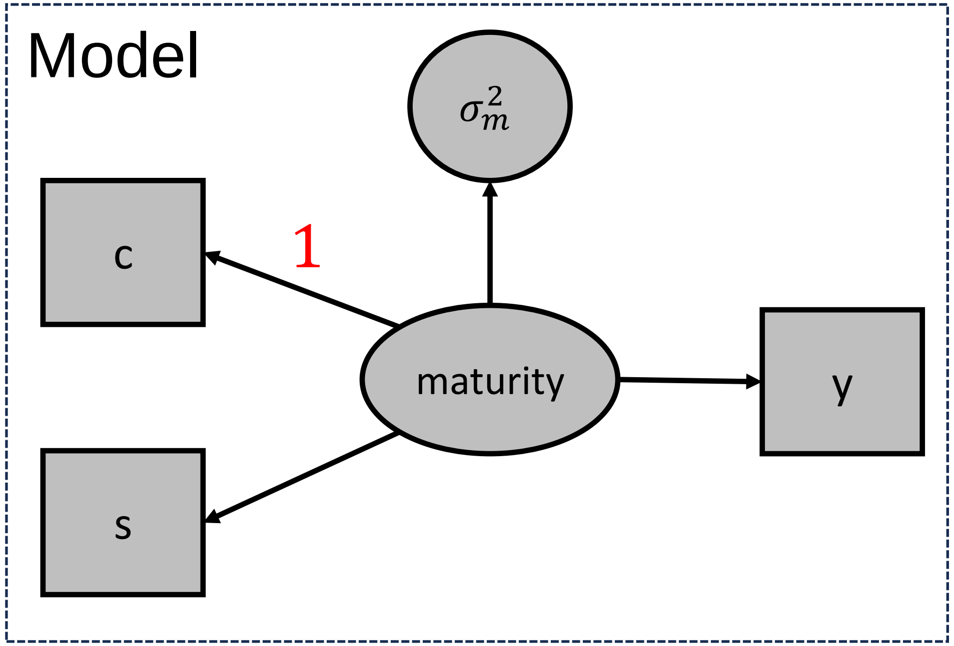


How could we scale maturity, i.e., what should the minimum and maximum values of maturity be?

$$\boldsymbol{m} \sim \text{normal}(0, \sigma_m^2)$$

What if we fix a beta (to give it a scale)?

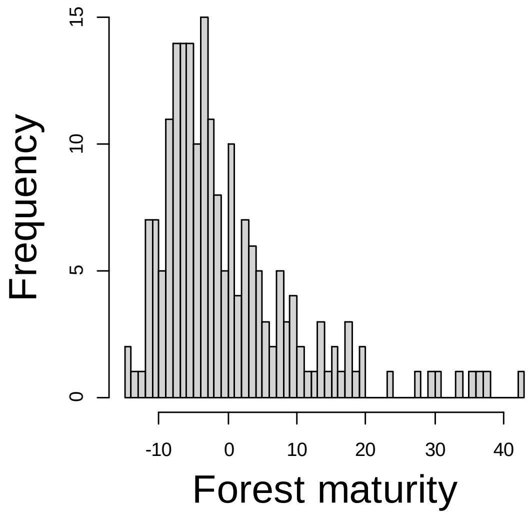


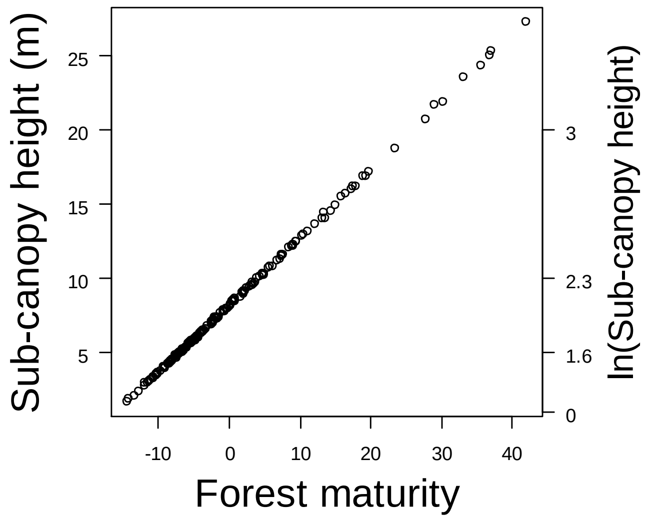


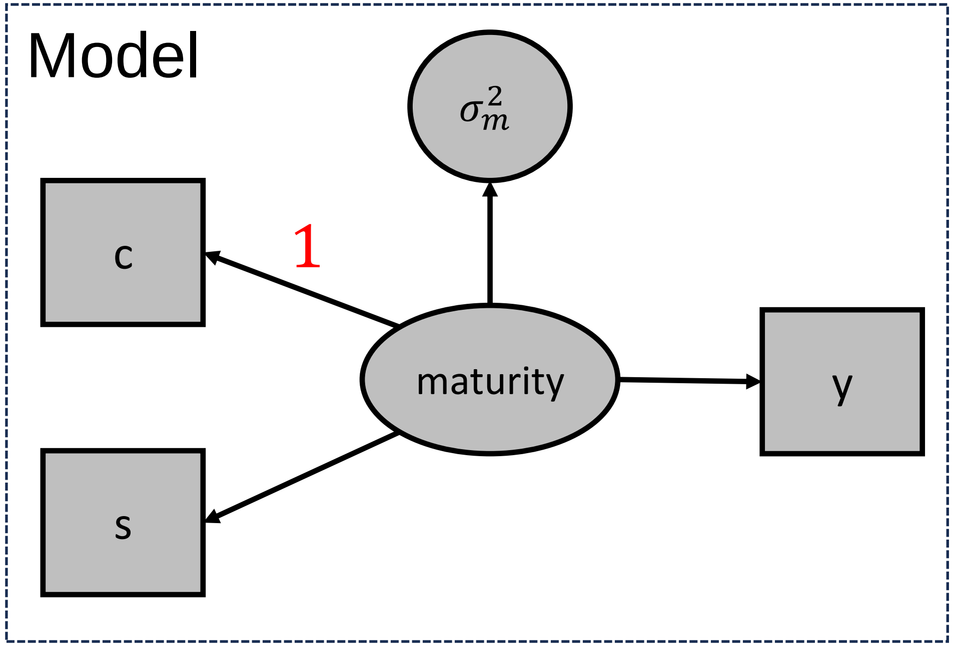
How could we scale maturity, i.e., what should the minimum and maximum values of maturity be?

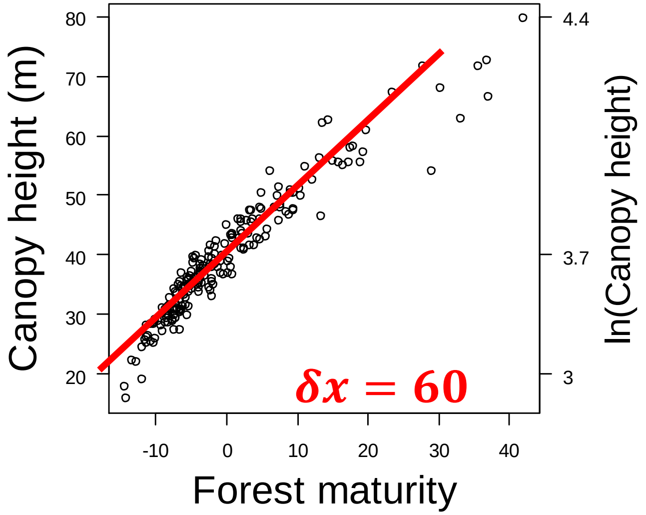
$$\boldsymbol{m} \sim \text{normal}(0, \sigma_m^2)$$

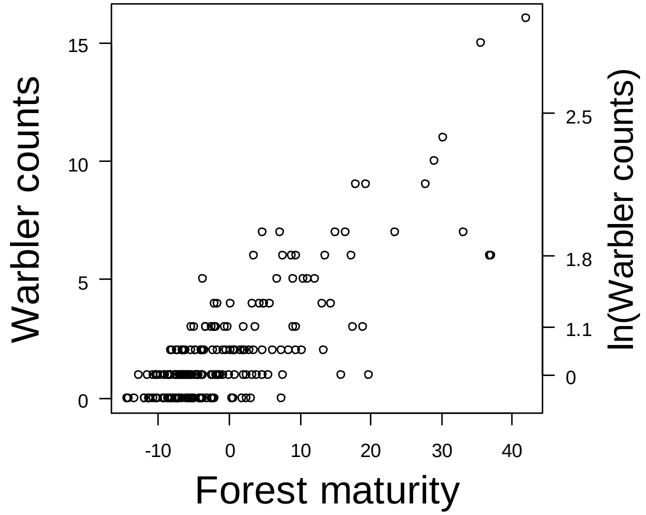
Now we have a scale!!



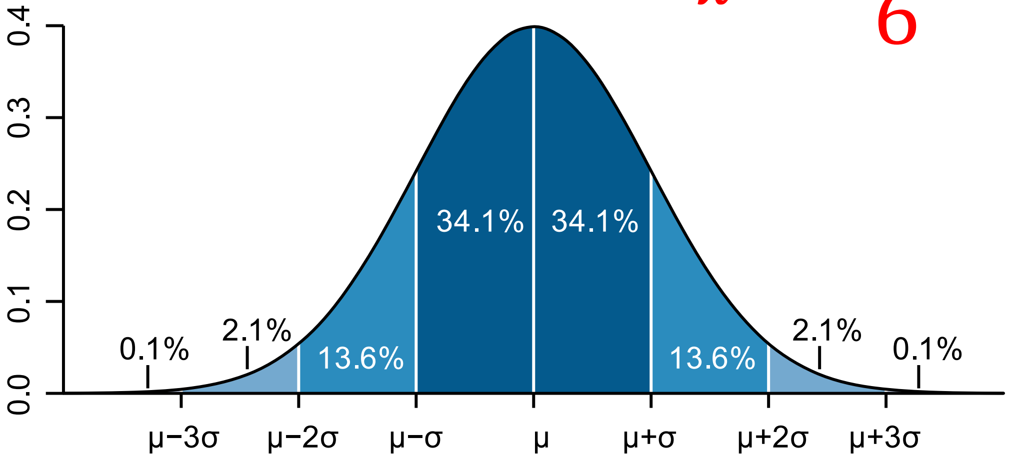




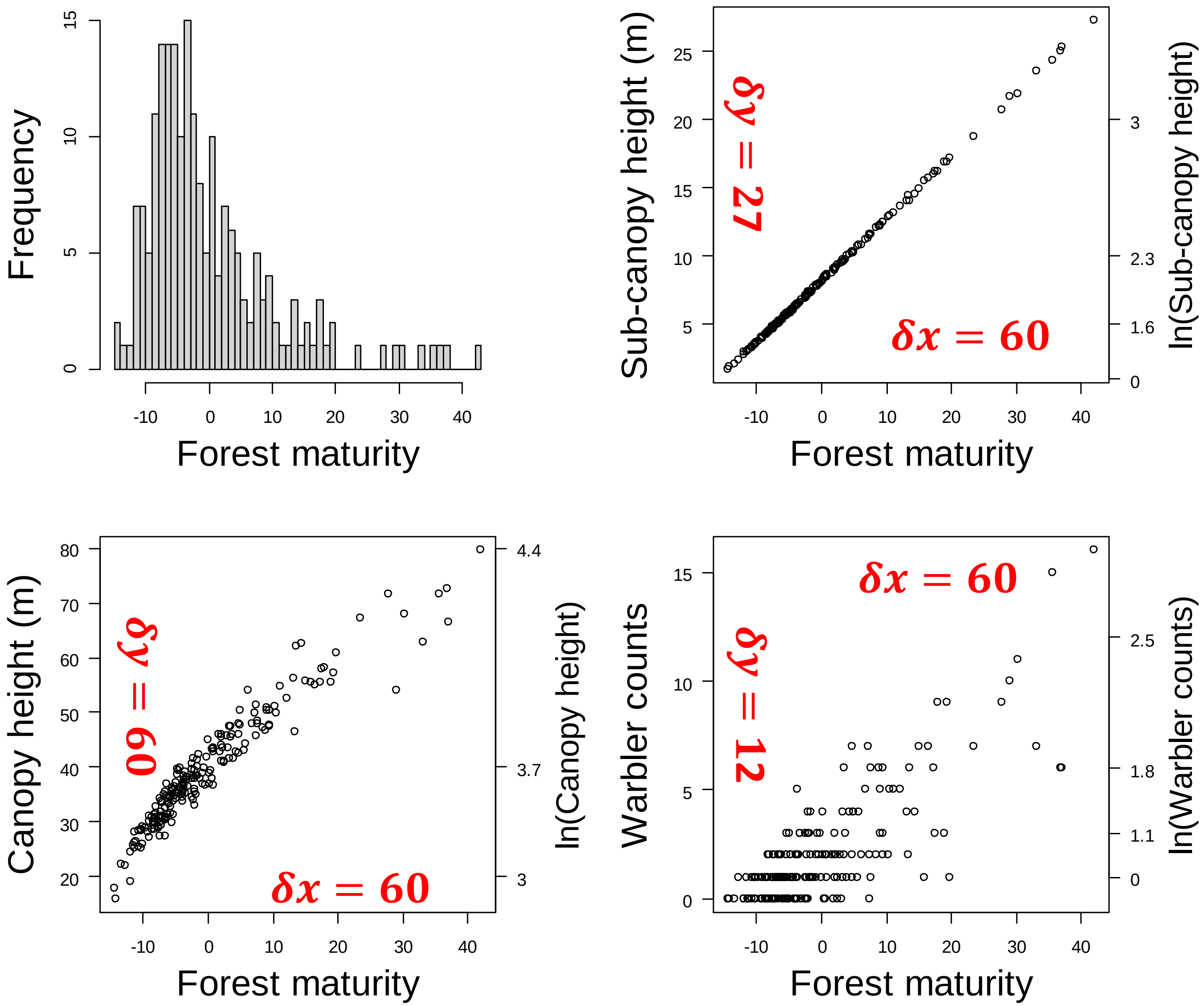


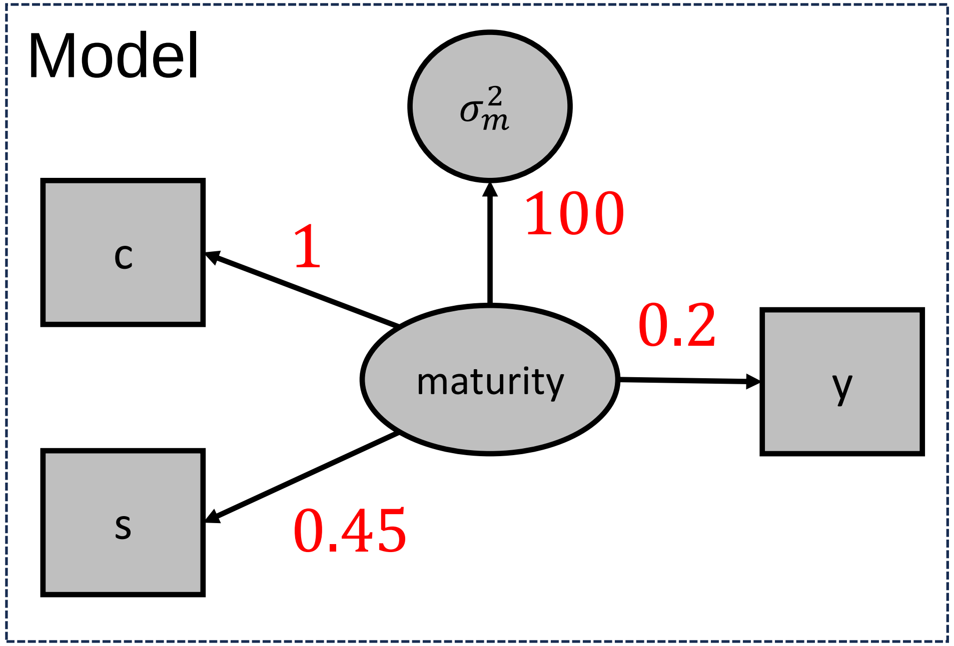


$$\sigma_x \cong \frac{\delta x}{6}$$



We can estimate all the betas

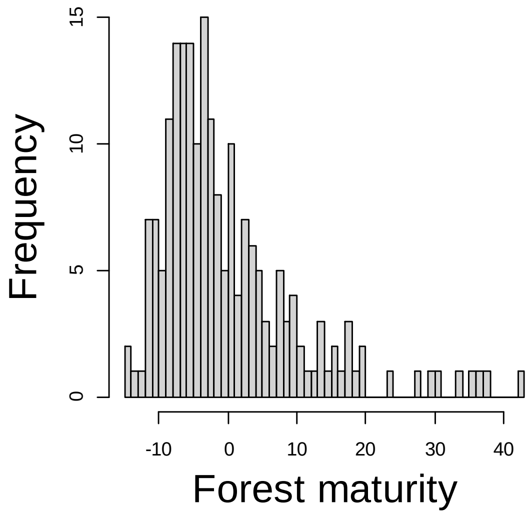


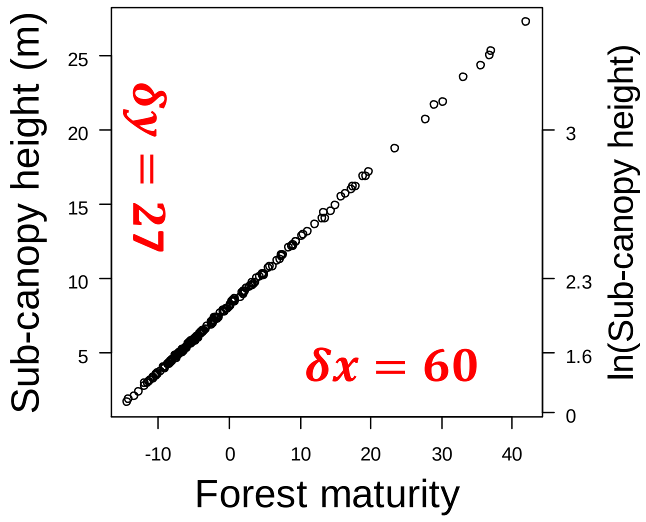


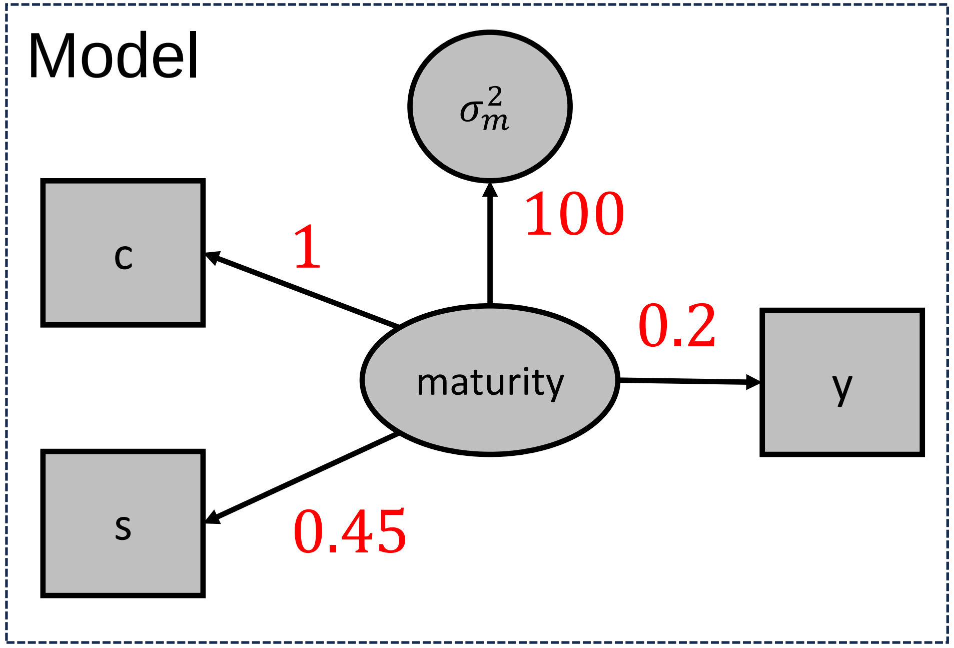
How could we scale maturity, i.e., what should the minimum and maximum values of maturity be?

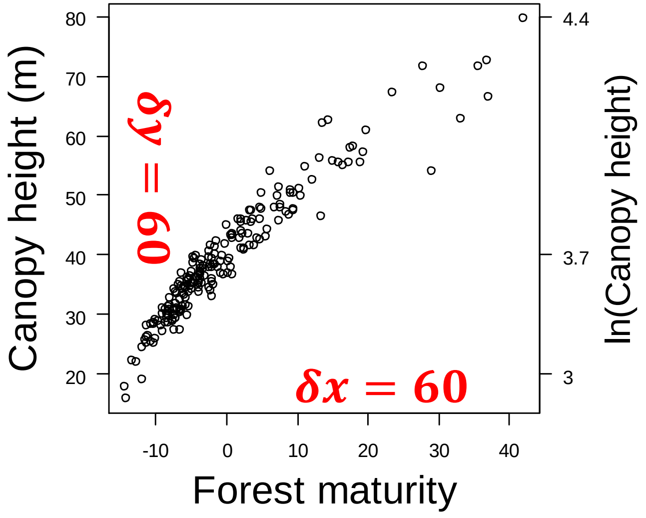
$$\boldsymbol{m} \sim \mathrm{normal}(0, \sigma_m^2)$$

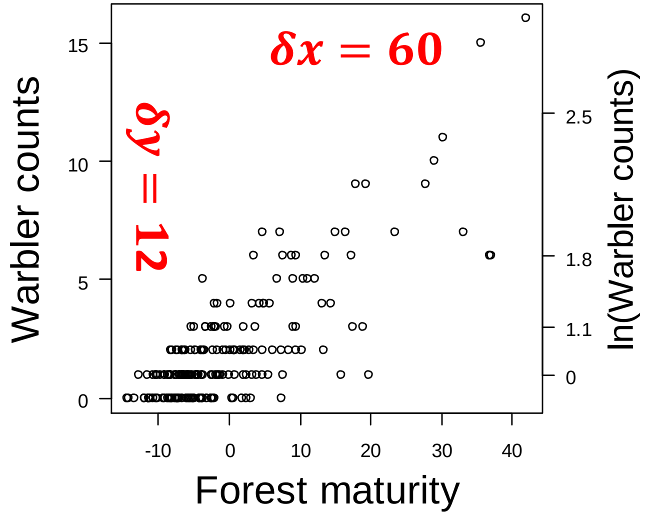
The scale of our latent variable is arbitrary







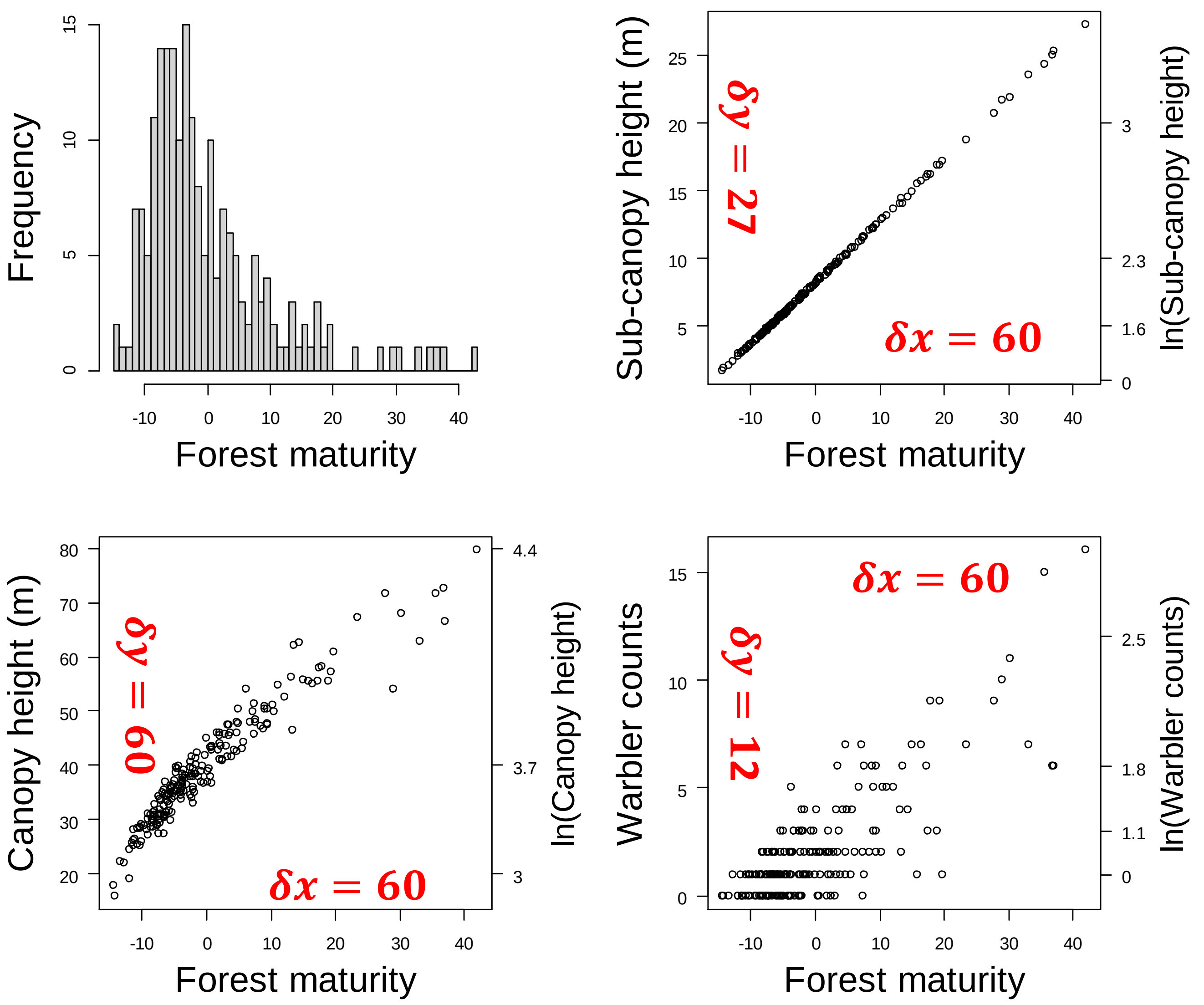


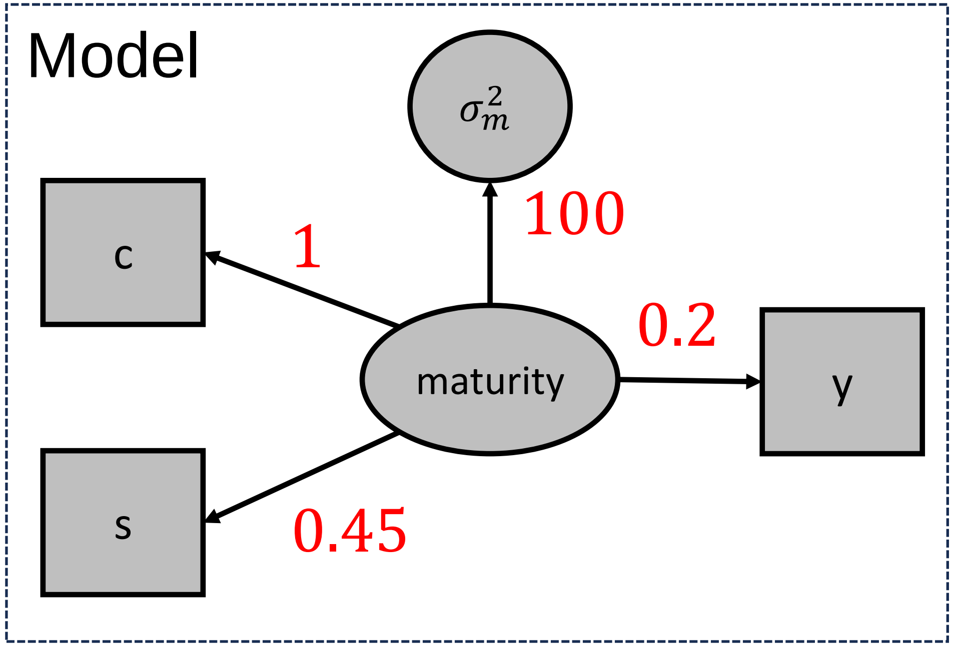


How could we scale maturity, i.e., what should the minimum and maximum values of maturity be?

$$\boldsymbol{m} \sim \mathrm{normal}(0, \sigma_m^2)$$

The scale of our latent variable is arbitrary

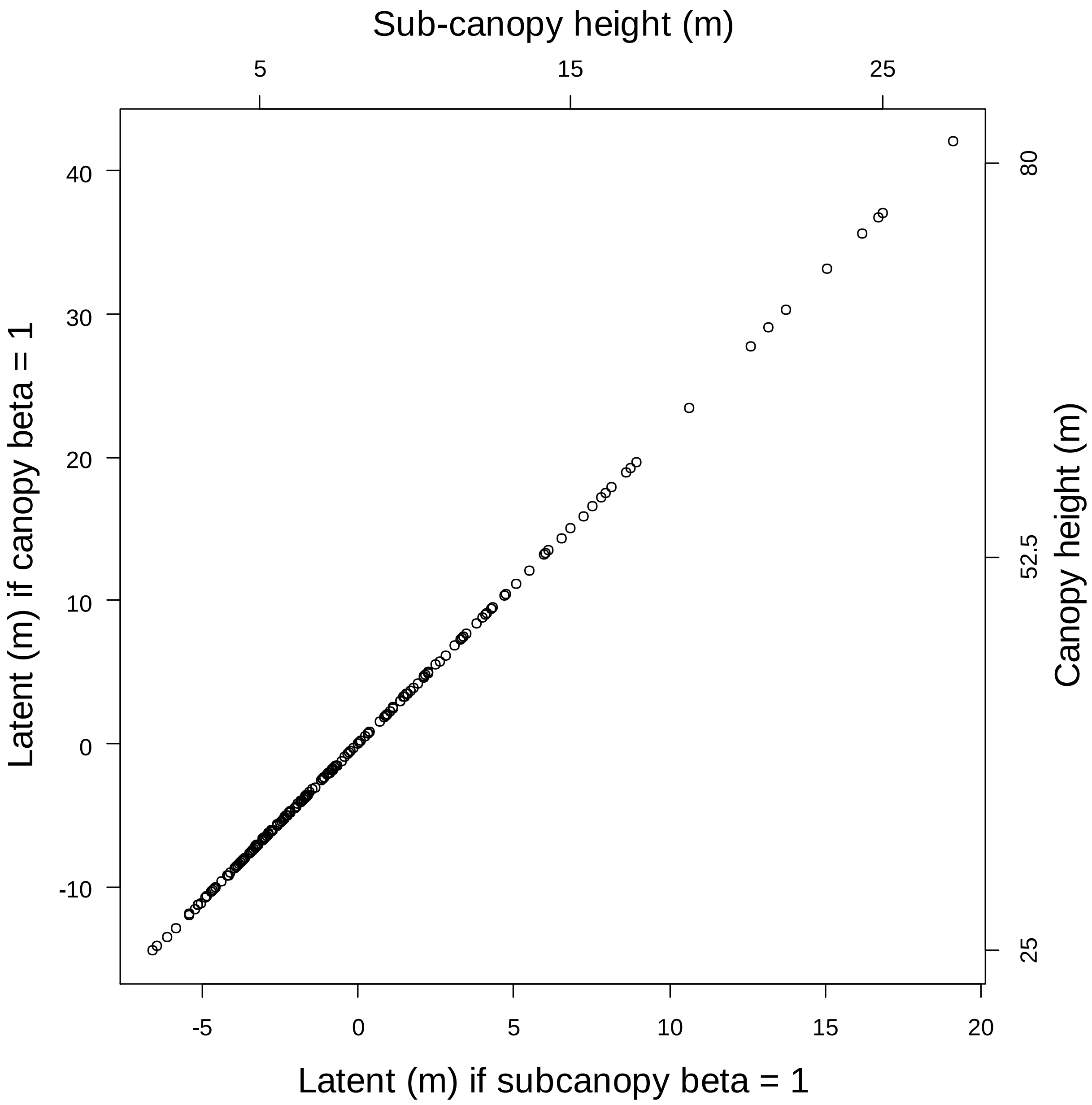


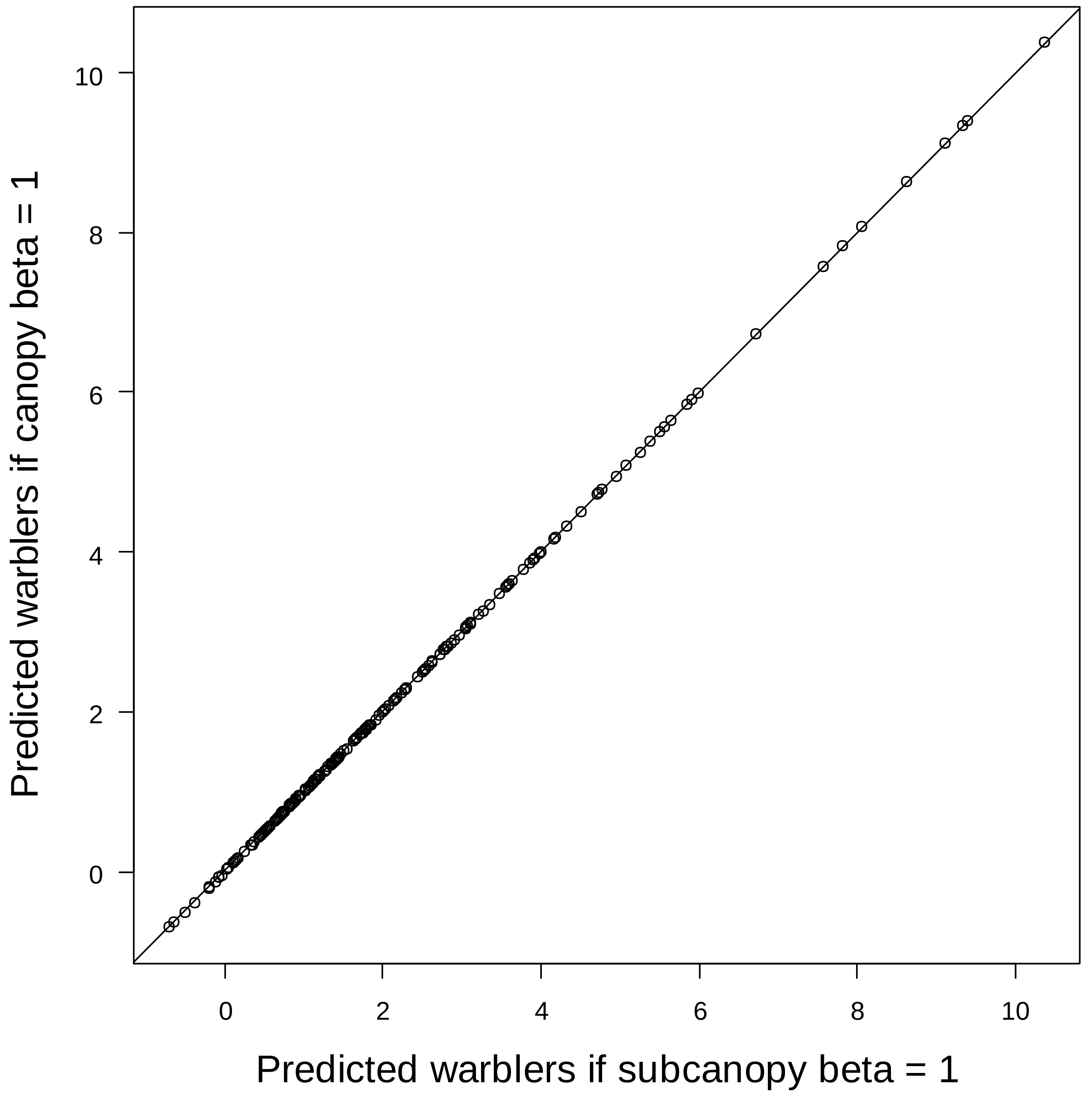


How could we scale maturity, i.e., what should the minimum and maximum values of maturity be?

$$\boldsymbol{m} \sim \text{normal}(0, \sigma_m^2)$$

Scale of latent variable can change, predictions don't



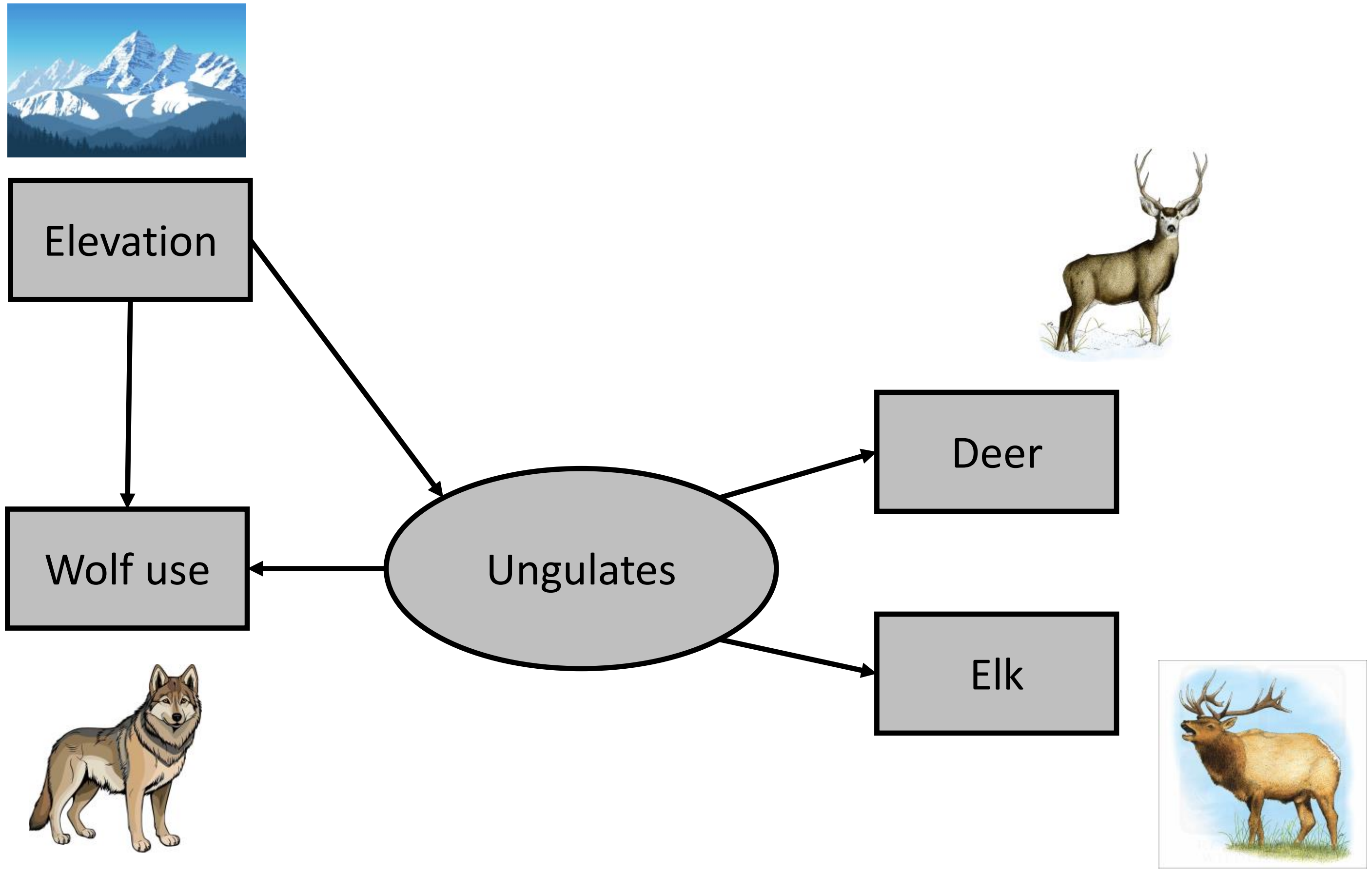


`lavaan` syntax

```
d <- data.frame(c = canopy, s = subcan, y = warblers)
sem1 <- sem('m =~ c + s
            y ~ m
            c ~ 1
            s ~ 1',
            data = d)
summary(sem1)
m.pred <- as.numeric(predict(sem1))
```

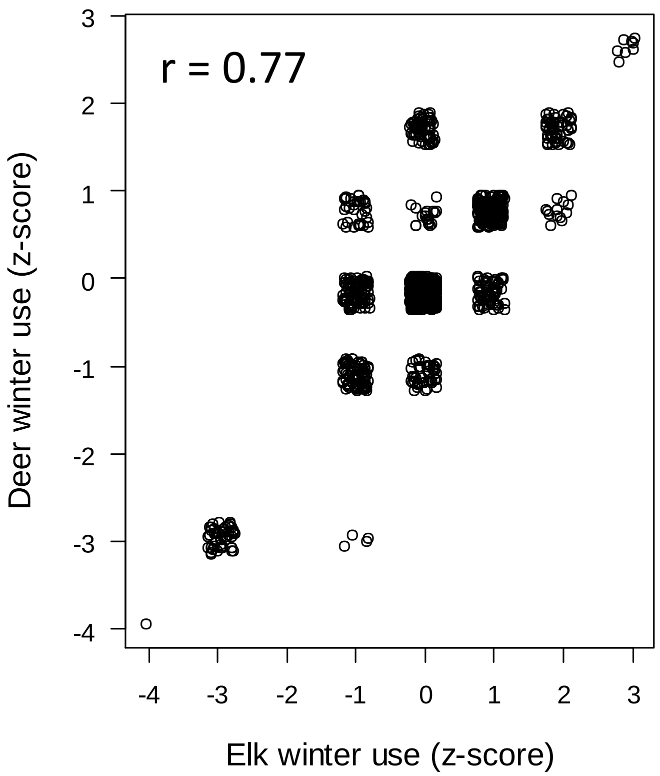
Expanding the Bow Valley wolf analysis with latent variables

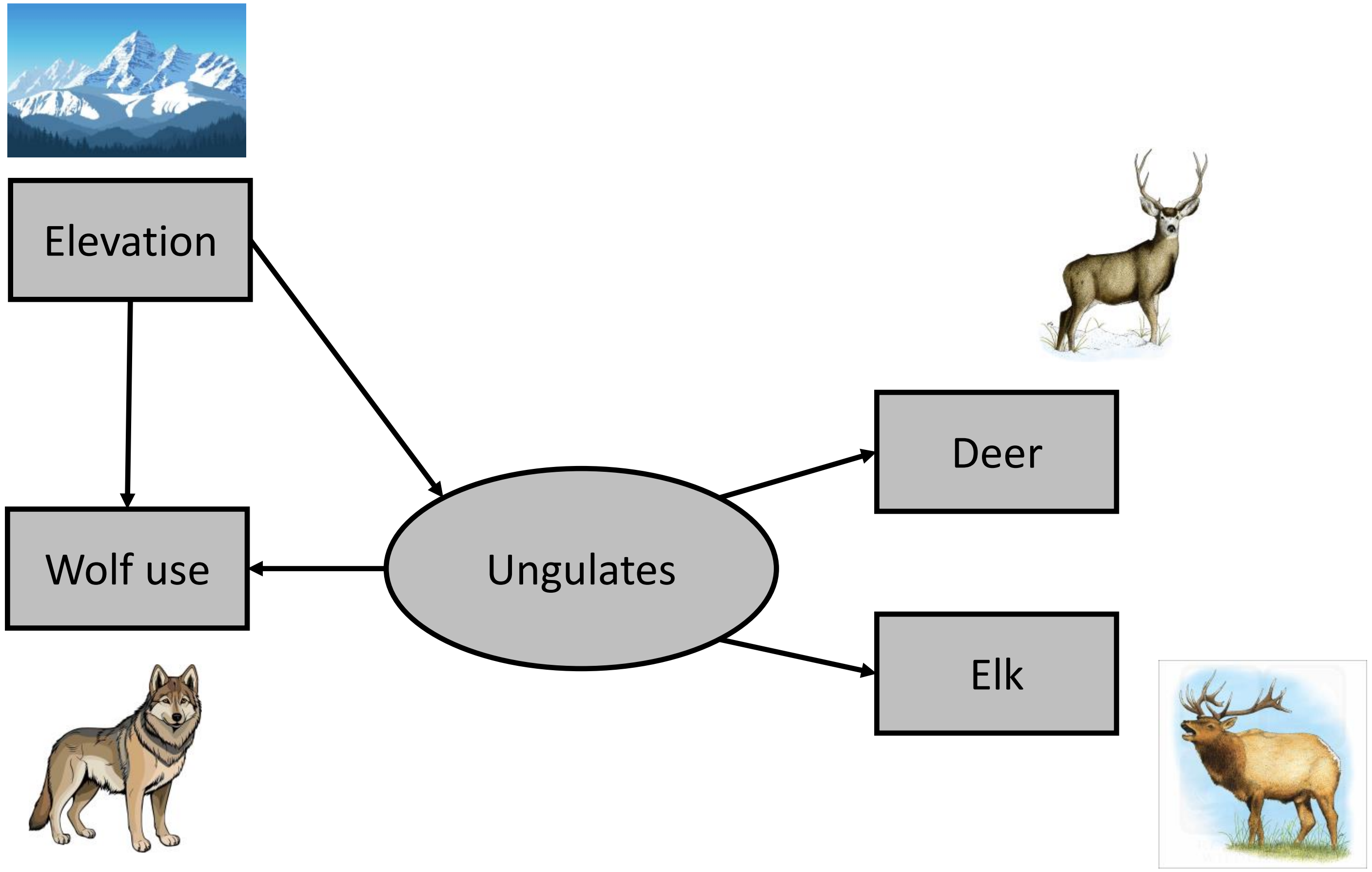
- Imagine that we're not just interested in deer, but in deer and elk abundance



Expanding the Bow Valley wolf analysis with latent variables

- Imagine that we're not just interested in deer, but in modeling deer and elk abundance simultaneously

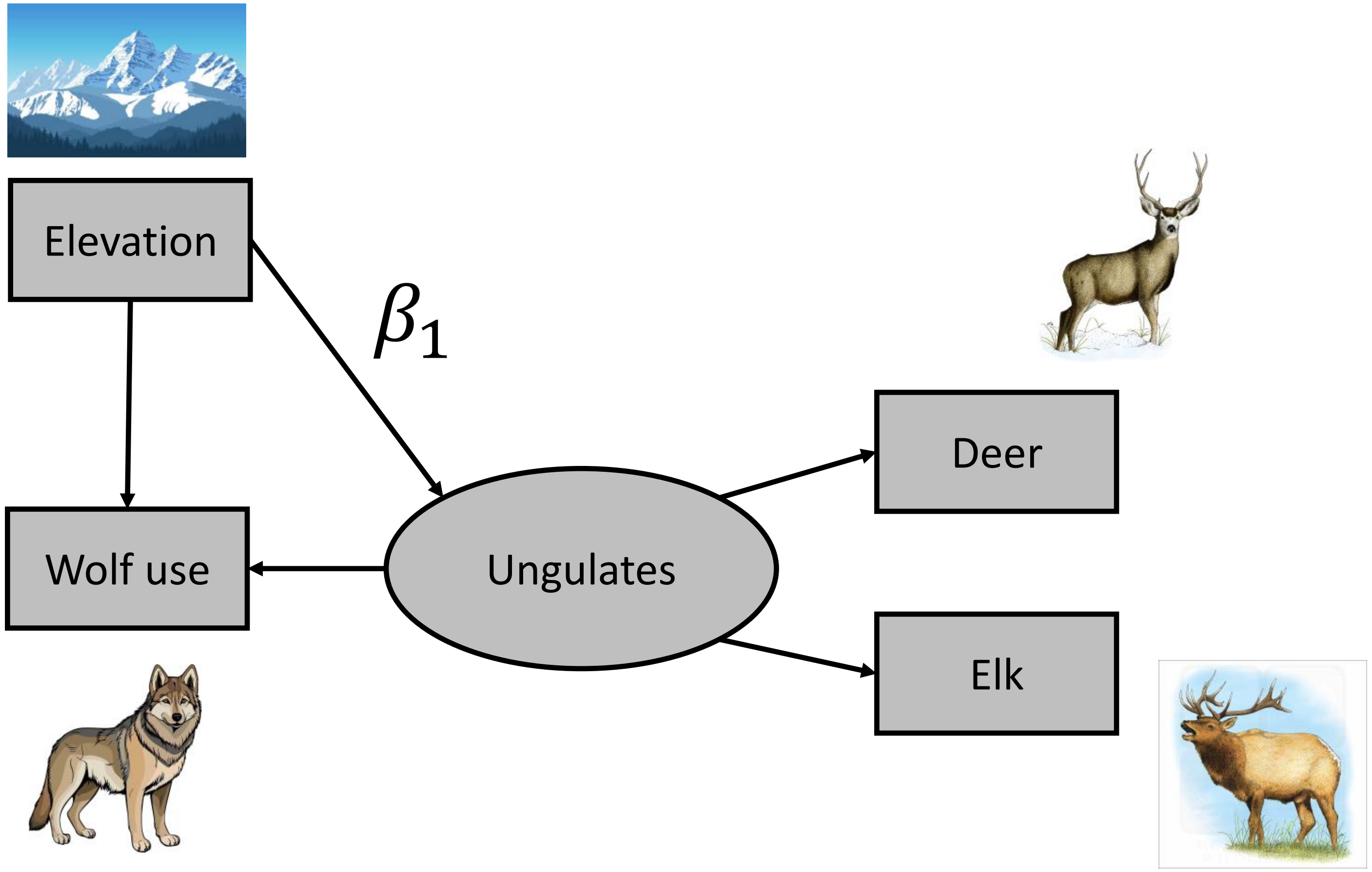




Expanding the Bow Valley wolf analysis with latent variables

- We can construct an ungulate latent variable modeled as a function of elevation (e), and measured via deer and elk winter habitat use…

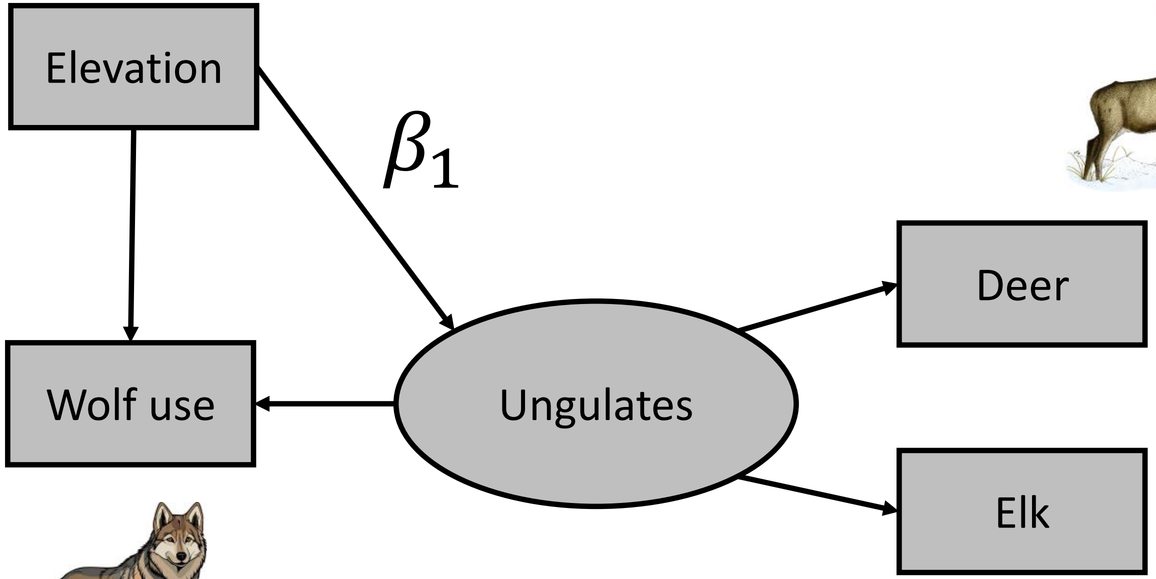
$$u_i \sim \mathrm{normal}(\beta_1 \times elev_i, \sigma_u^2)$$



Expanding the Bow Valley wolf analysis with latent variables

- We can construct an ungulate latent variable modeled as a function of elevation (e), and measured via deer and elk winter habitat use…

$$\boxed{u_i \sim \text{normal}(\beta_1 \times elev_i, \sigma_u^2)}$$



```
model <- "ung =~ z.deer + z.elk
          ung ~ z.elev
          used ~ ung + z.elev
          z.deer ~ 0
          z.elk ~ 0"
```

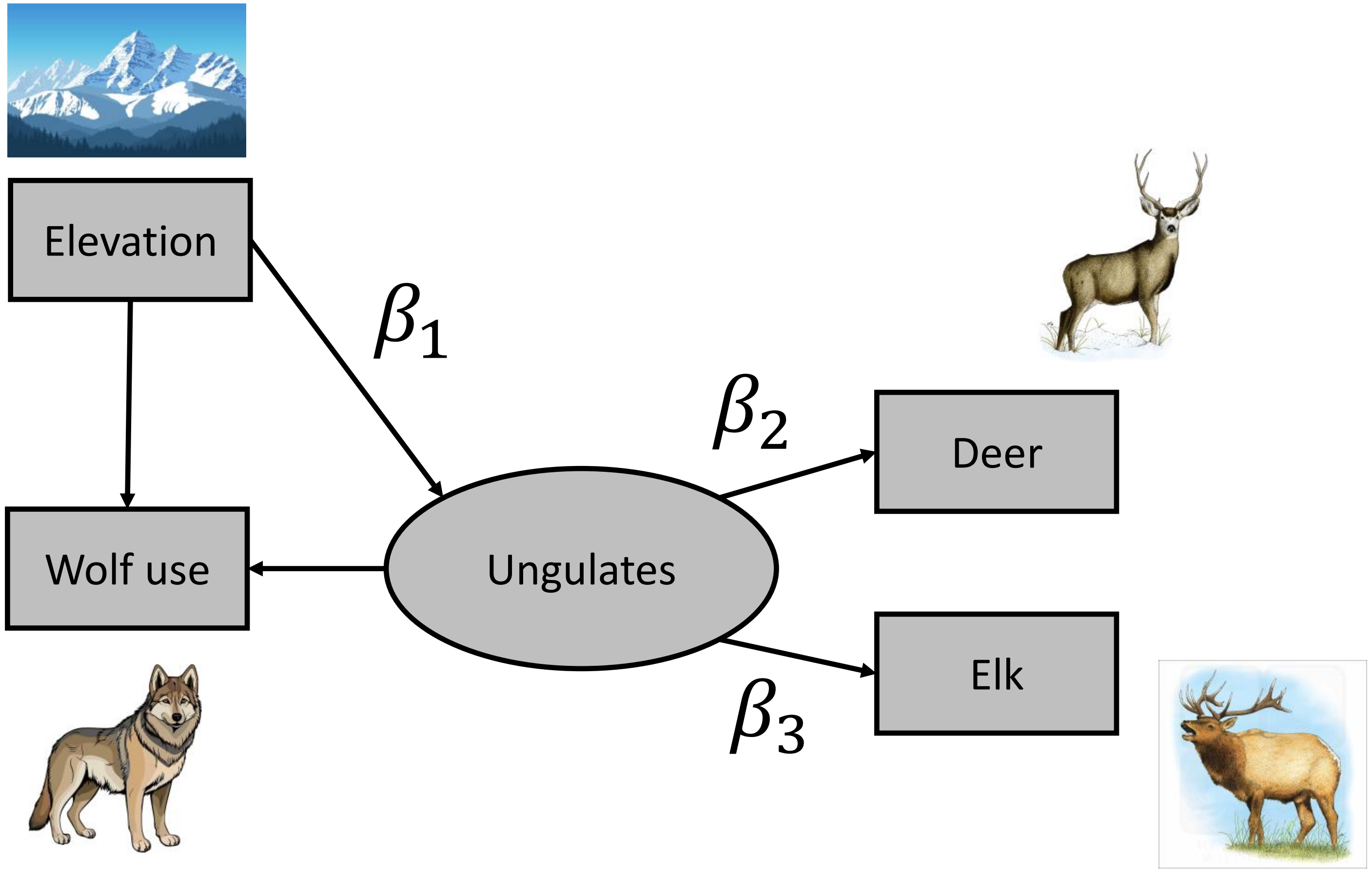
Expanding the Bow Valley wolf analysis with latent variables

- We can construct an ungulate latent variable modeled as a function of elevation (e), and measured via deer and elk winter habitat use…

$$u_i \sim \mathrm{normal}(\beta_1 \times elev_i, \sigma_u^2)$$

$$deer_i \sim \mathrm{normal}(\beta_2 \times u_i, \sigma_{deer}^2)$$

$$elk_i \sim \mathrm{normal}(\beta_3 \times u_i, \sigma_{elk}^2)$$



Expanding the Bow Valley wolf analysis with latent variables

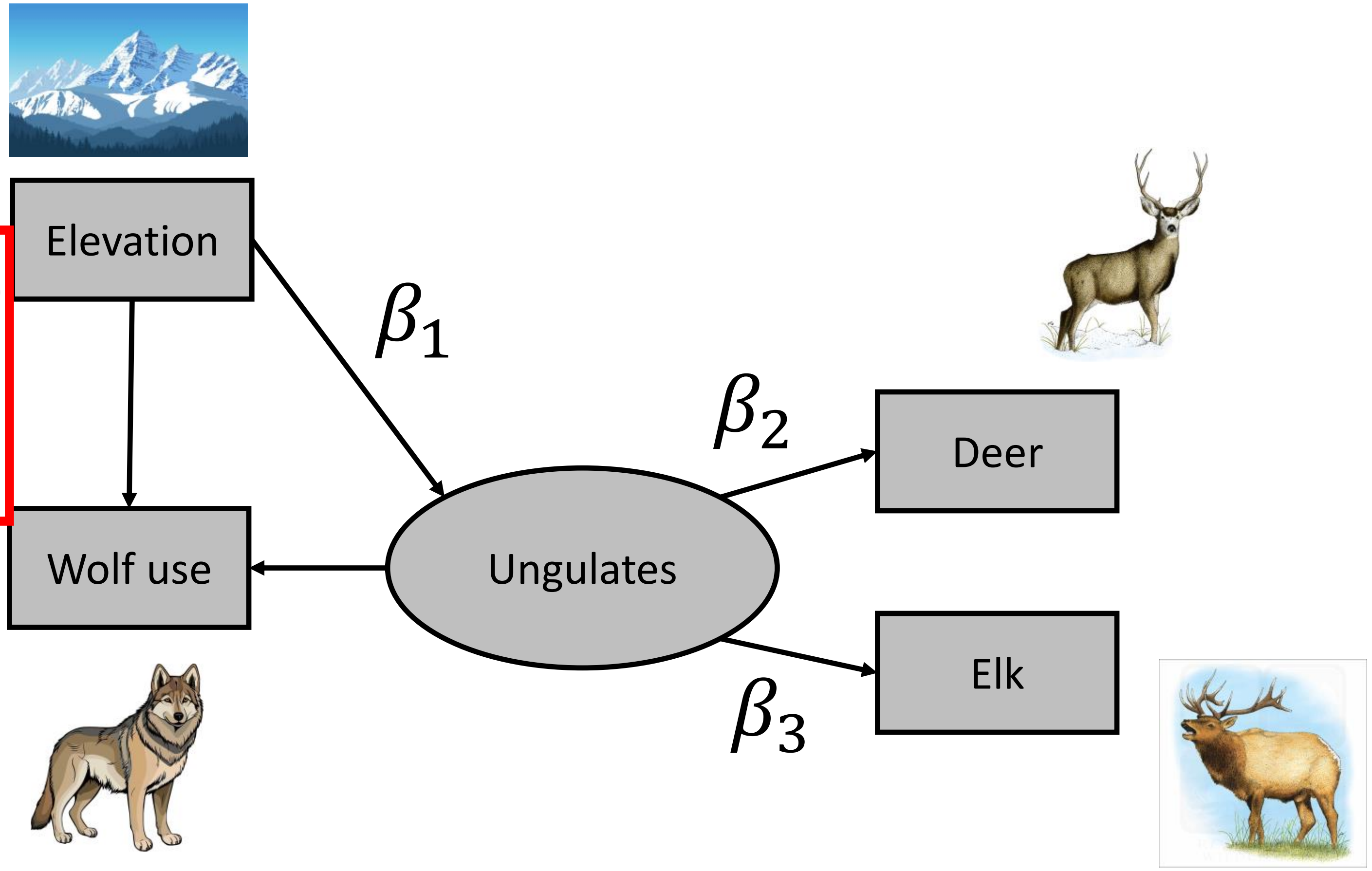
- We can construct an ungulate latent variable modeled as a function of elevation (e), and measured via deer and elk winter habitat use…

$$u_i \sim \text{normal}(\beta_1 \times elev_i, \sigma_u^2)$$

$$deer_i \sim \text{normal}(\beta_2 \times u_i, \sigma_{deer}^2)$$

$$elk_i \sim \text{normal}(\beta_3 \times u_i, \sigma_{elk}^2)$$

```
model <- "ung =~ z.deer + z.elk
          ung ~ z.elev
          used ~ ung + z.elev
          z.deer ~ 0
          z.elk ~ 0"
```



Expanding the Bow Valley wolf analysis with latent variables

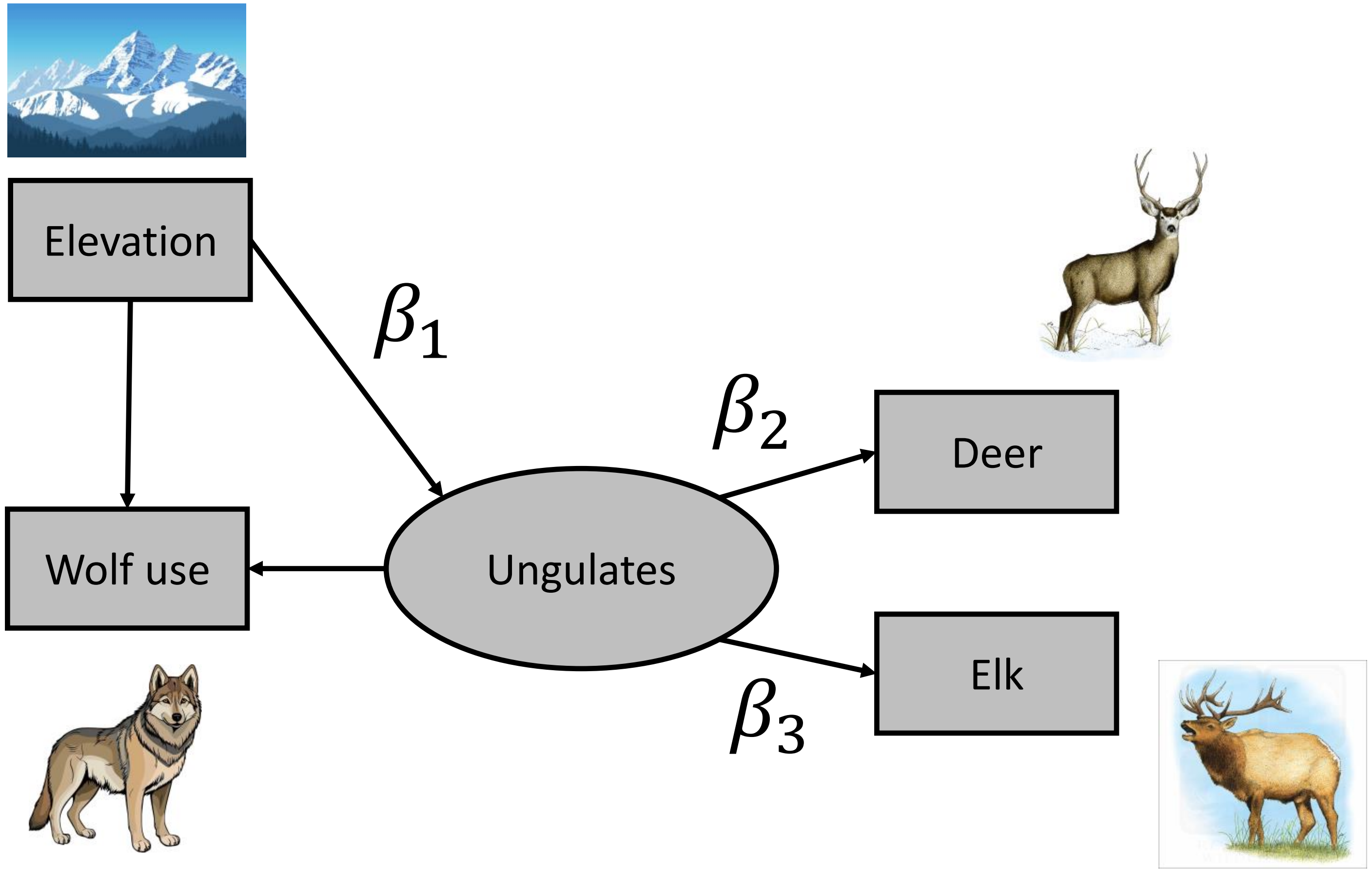
- We can construct an ungulate latent variable modeled as a function of elevation (e), and measured via deer and elk winter habitat use…

$$u_i \sim \text{normal}(\beta_1 \times elev_i, \sigma_u^2)$$

$$deer_i \sim \text{normal}(\beta_2 \times u_i, \sigma_{deer}^2)$$

$$elk_i \sim \text{normal}(\beta_3 \times u_i, \sigma_{elk}^2)$$

```
model <- "ung =~ z.deer + z.elk
          ung ~ z.elev
          used ~ ung + z.elev
          z.deer ~ 0
          z.elk ~ 0"
```

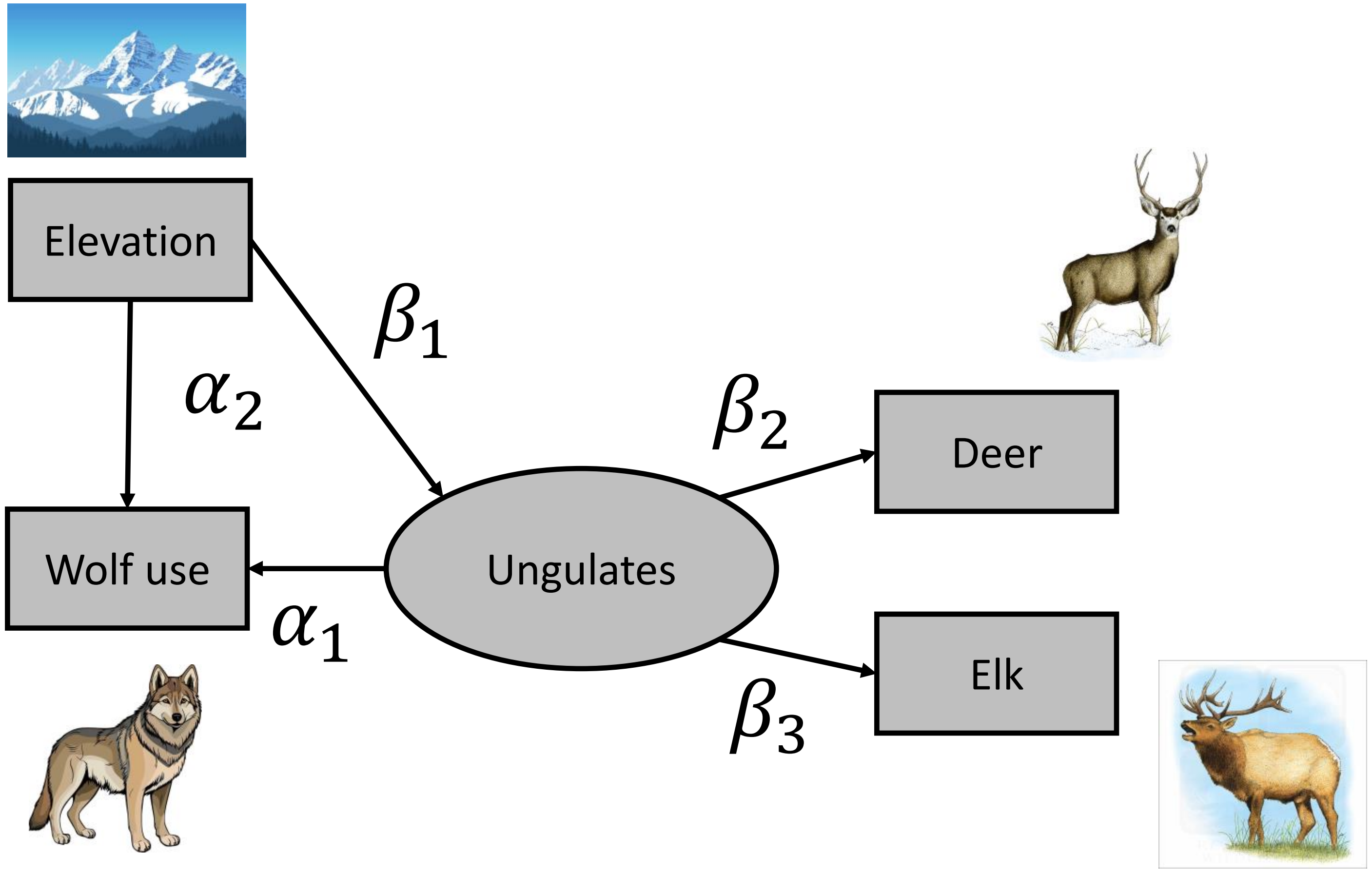


Expanding the Bow Valley wolf analysis with latent variables

- We can construct an ungulate latent variable modeled as a function of elevation (e), and measured via deer and elk winter habitat use…

$$w_i \sim \text{Bernoulli}(\varphi_i)$$

$$\text{logit}(\varphi_i) = \alpha_0 + \alpha_1 u_i + \alpha_2 e_i$$



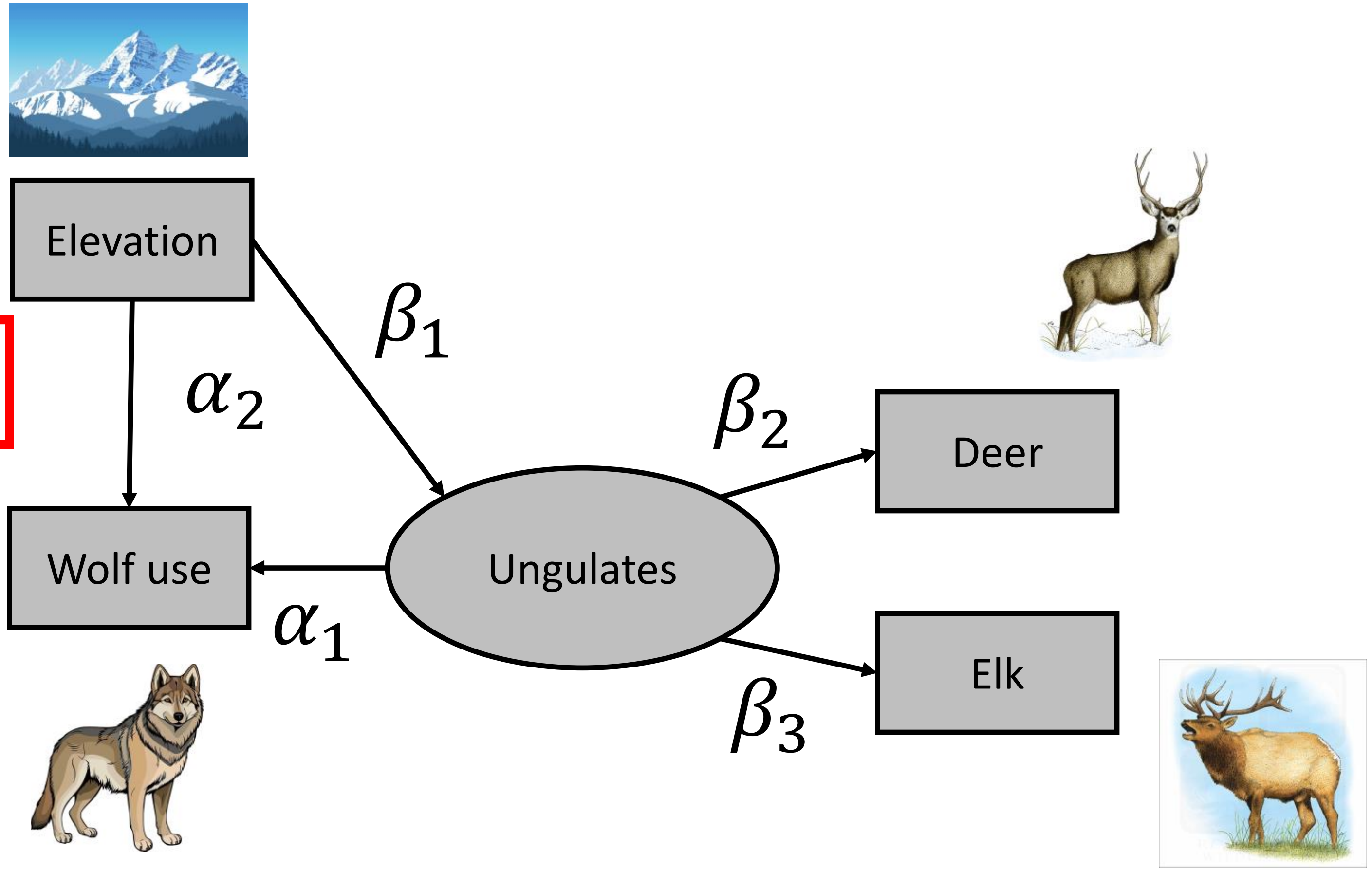
Expanding the Bow Valley wolf analysis with latent variables

- We can construct an ungulate latent variable modeled as a function of elevation (e), and measured via deer and elk winter habitat use…

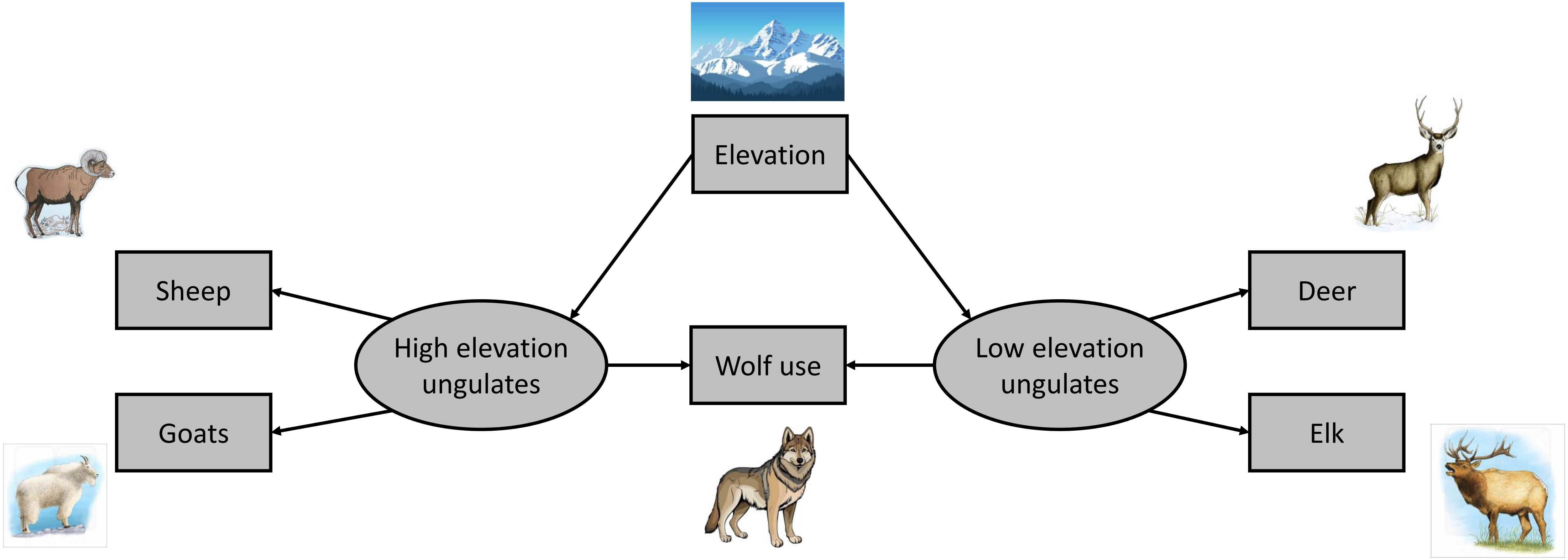
$$w_i \sim \text{Bernoulli}(\varphi_i)$$

$$\boxed{\text{logit}(\varphi_i) = \alpha_0 + \alpha_1 u_i + \alpha_2 e_i}$$

```
model <- "ung =~ z.deer + z.elk
          ung ~ z.elev
          used ~ ung + z.elev
          z.deer ~ 0
          z.elk ~ 0"
```



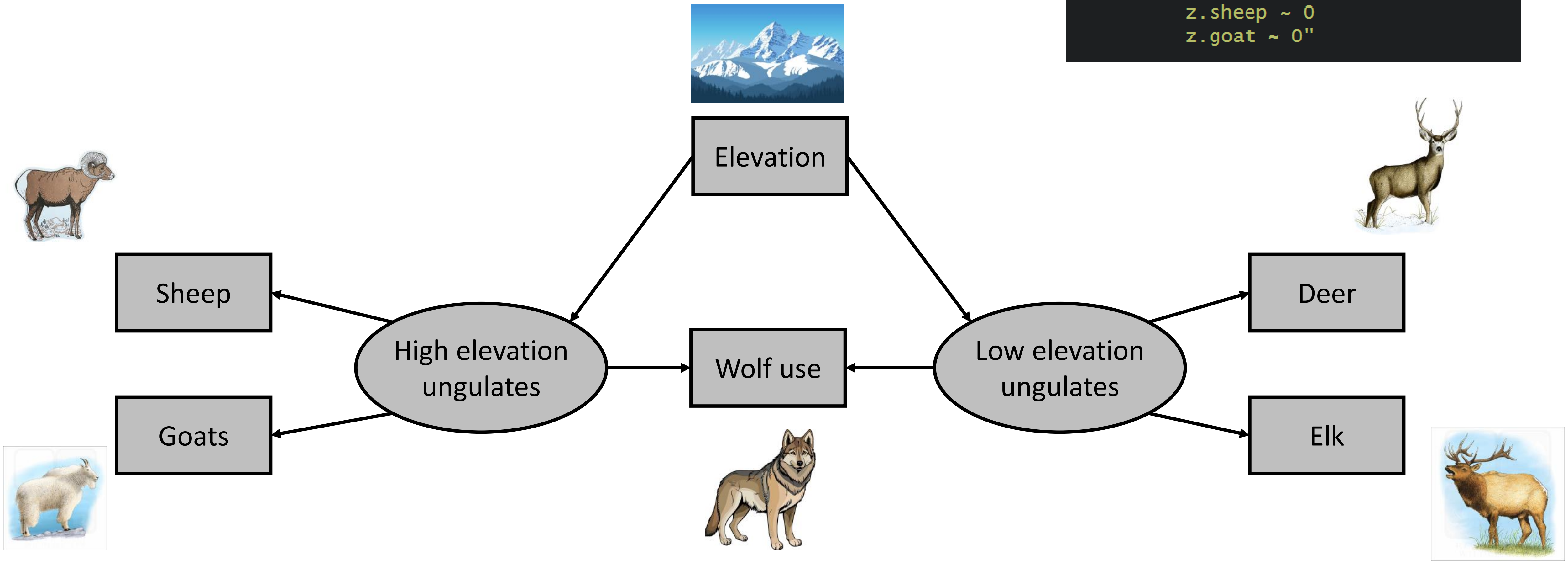
Do wolves prefer high or low elevation ungulates?



Do wolves prefer high or low elevation ungulates?

- Imagine that we're not just interested in wolves, but in how ungulates respond to the landscape, and which ungulates wolves prefer…

```
model <- "lo.ung =~ z.deer + z.elk
          hi.ung =~ z.sheep + z.goat
          lo.ung ~ z.elev
          hi.ung ~ z.elev
          used ~ hi.ung + lo.ung
          z.deer ~ 0
          z.elk ~ 0
          z.sheep ~ 0
          z.goat ~ 0"
```



Do wolves prefer high or low elevation ungulates?

```
model <- "lo.ung =~ z.deer + z.elk
          hi.ung =~ z.sheep + z.goat
          lo.ung ~ z.elev
          hi.ung ~ z.elev
          used ~ hi.ung + lo.ung
          z.deer ~ 0
          z.elk ~ 0
          z.sheep ~ 0
          z.goat ~ 0"
```