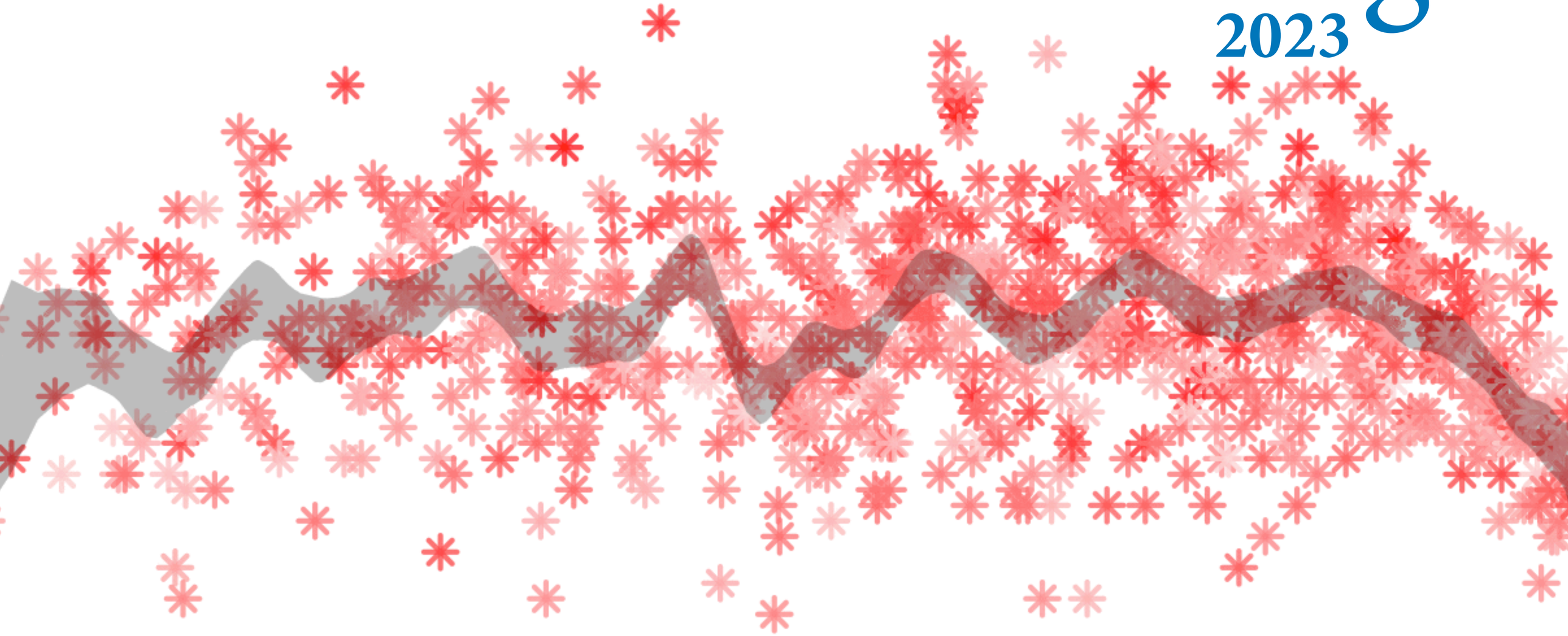
Statistical Rethinking

2023



18. Missing Data

Missing Data, Found

Observed data is special case: We trick ourselves into believing there is no error

Missing Data, Found

Observed data is special case: We trick ourselves into believing there is no error

Most data are missing most of the time

"Missing" data: Some cases unobserved

Not totally "missing": We know
(1) constraints
(2) relationships to other variables

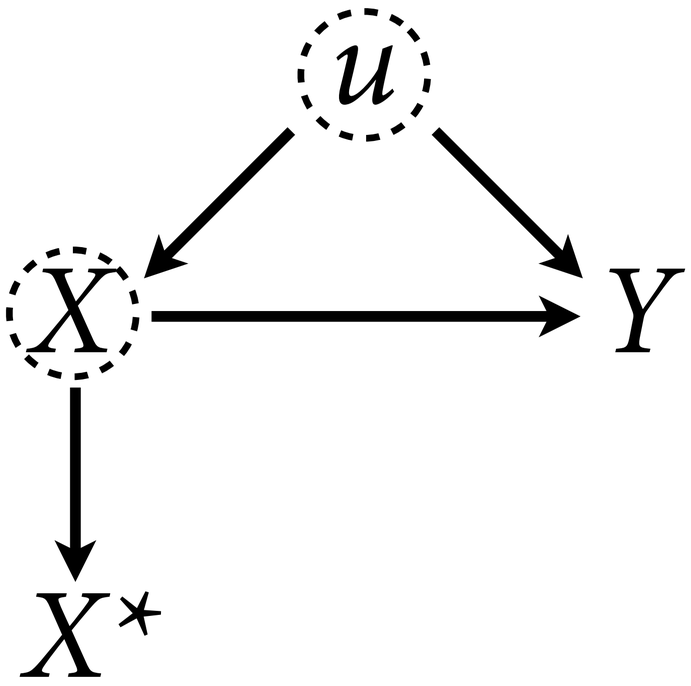
Missing Data is Workflow

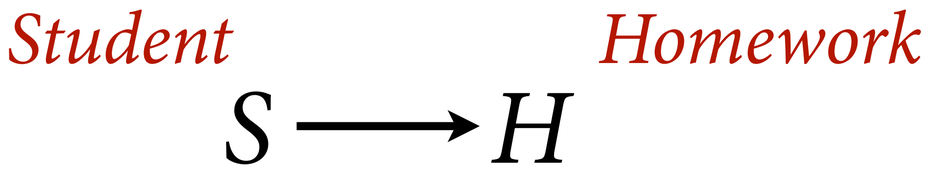
What to do with missing data?

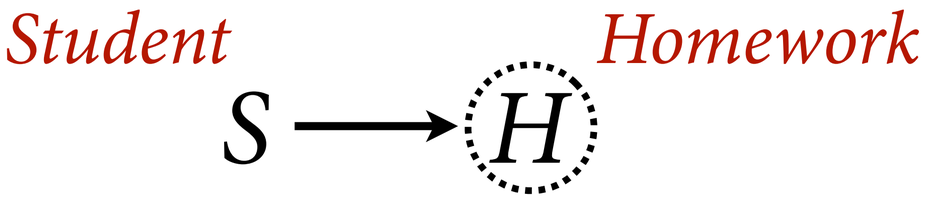
Dropping cases with missing values sometimes justifiable

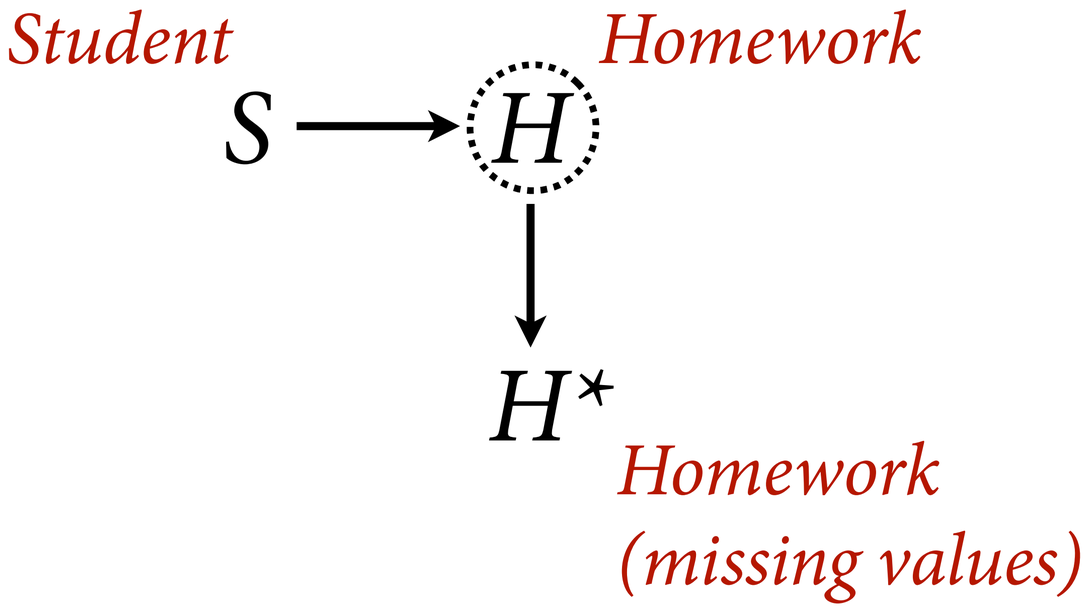
Right thing to do depends upon causal assumptions

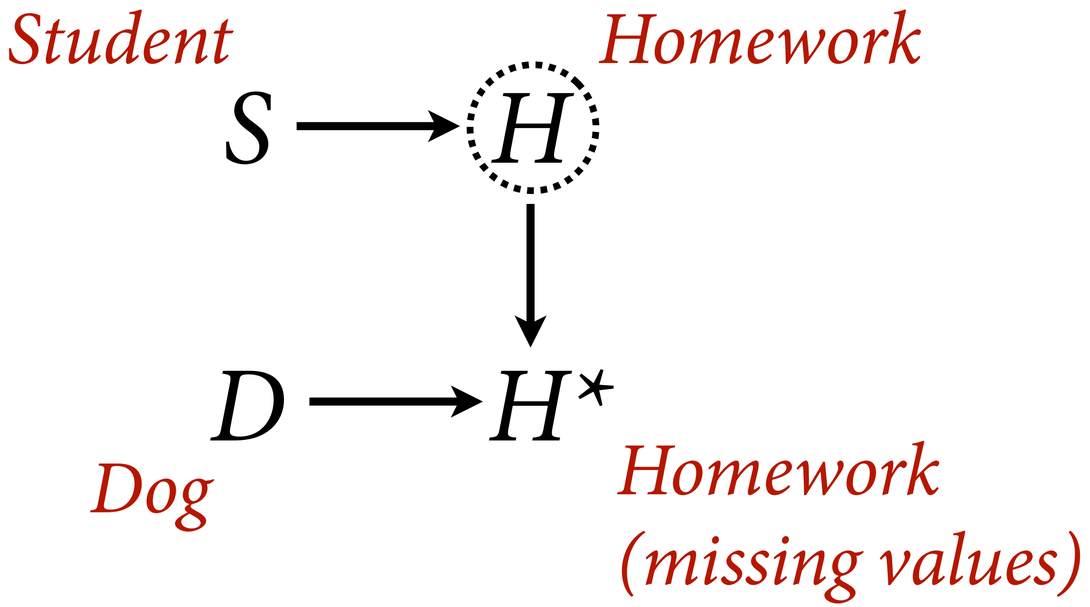
Imputation often beneficial/necessary



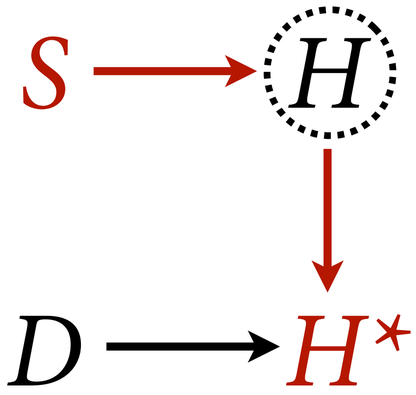

Student
Homework
S
H


Student
Homework
S
H


Student
Homework
S
H
H*
Homework
(missing values)


Student
Homework
S
H
D
H*
Dog
Homework
(missing values)

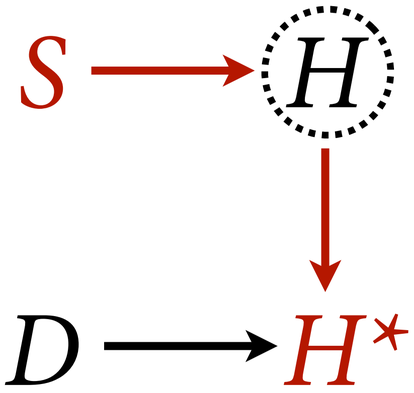
No biasing paths
connecting H to S

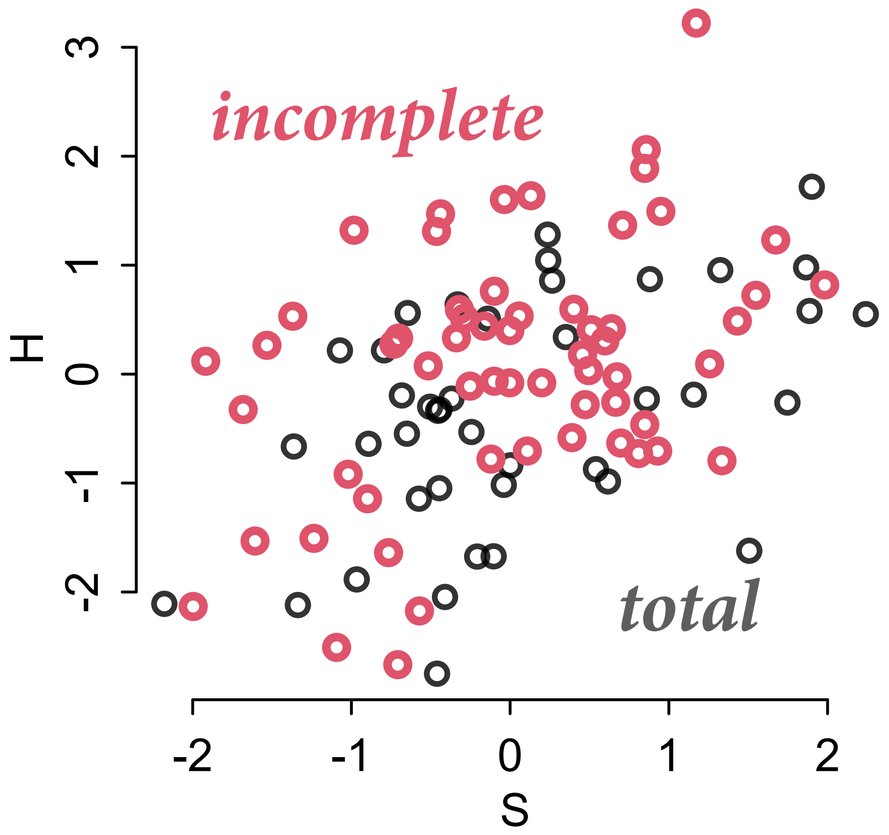


“Dog eats random homework”

Dog usually benign

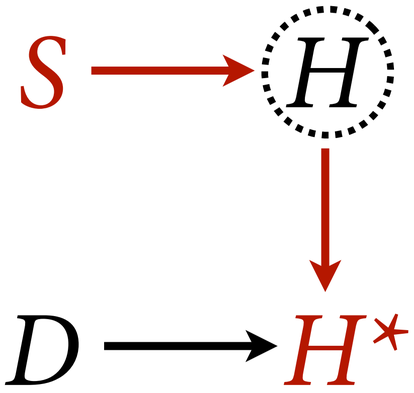
```
# Dog eats random homework
# aka missing completely at random
N <- 100
S <- rnorm(N)
H <- rnorm(N,0.5*S)
# dog eats 50% of homework at random
D <- rbern(N,0.5)
Hstar <- H
Hstar[D==1] <- NA

plot( S , H , col=grau(0.8) , lwd=2 )
points( S , Hstar , col=2 , lwd=3 )
```

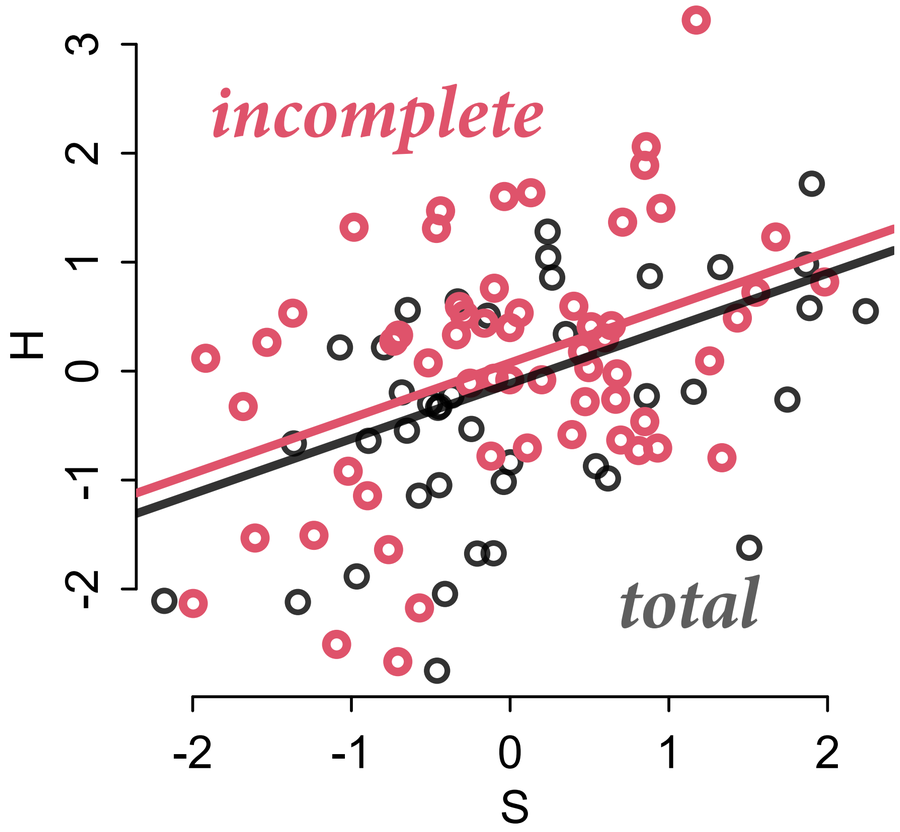




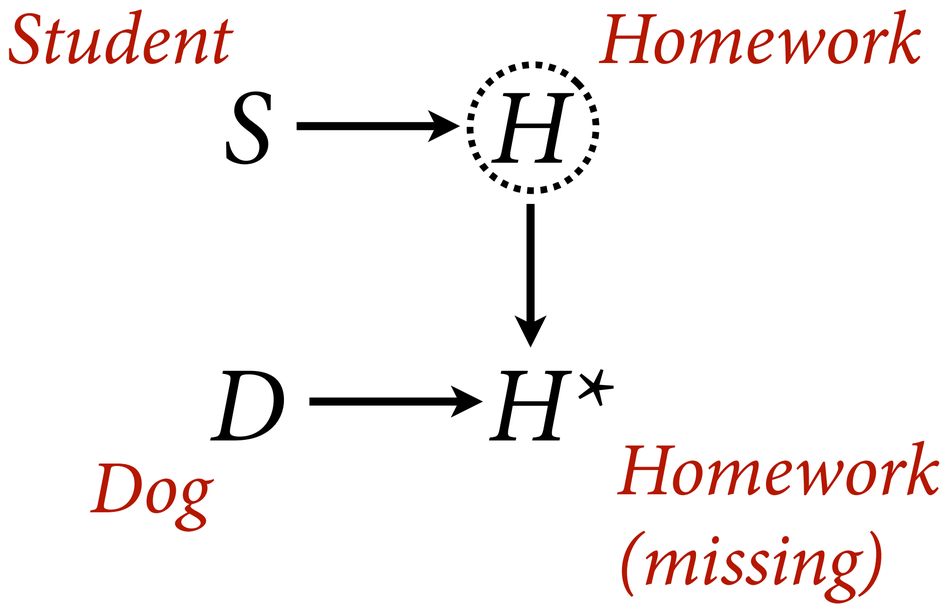
Dog usually benign

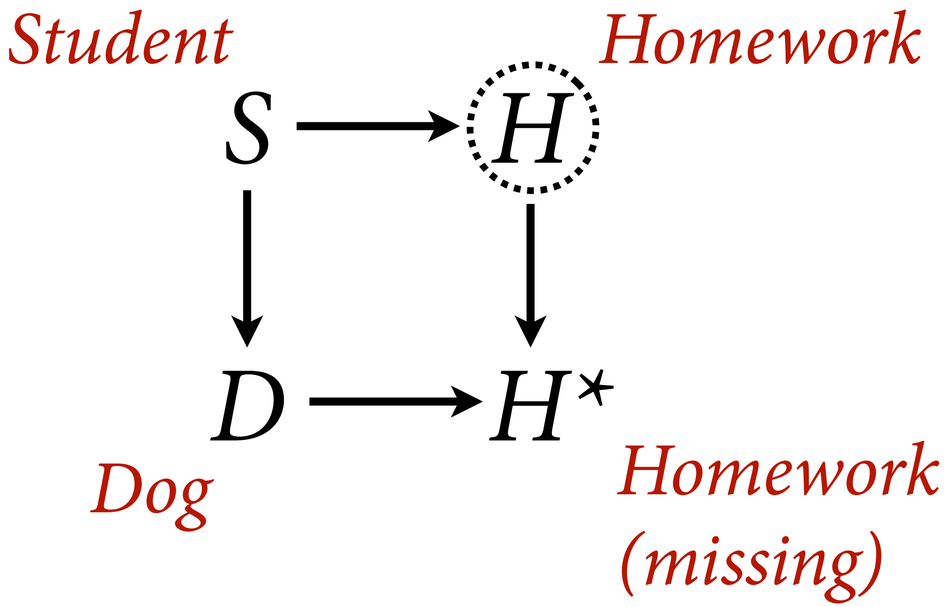
```
# Dog eats random homework
# aka missing completely at random
N <- 100
S <- rnorm(N)
H <- rnorm(N,0.5*S)
# dog eats 50% of homework at random
D <- rbern(N,0.5)
Hstar <- H
Hstar[D==1] <- NA

plot( S , H , col=grau(0.8) , lwd=2 )
points( S , Hstar , col=2 , lwd=3 )
```



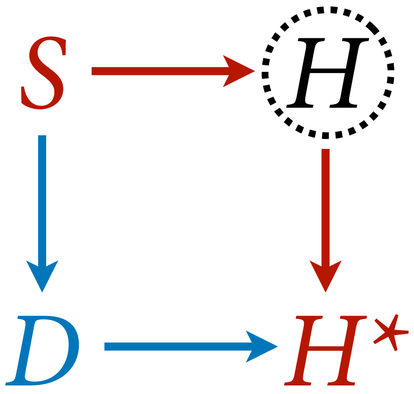


Loss of precision, usually no bias


Student
Homework
S
H
D
H*
Dog
Homework
(missing)


Student
Homework
S
H
D
H*
Dog
Homework
(missing)

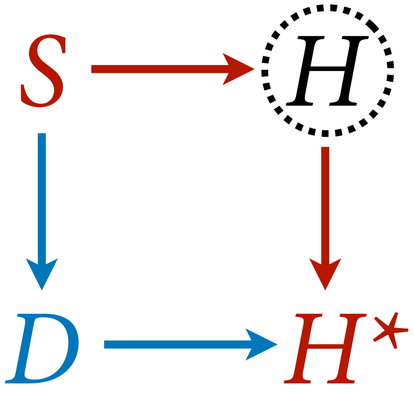
Maybe biasing path

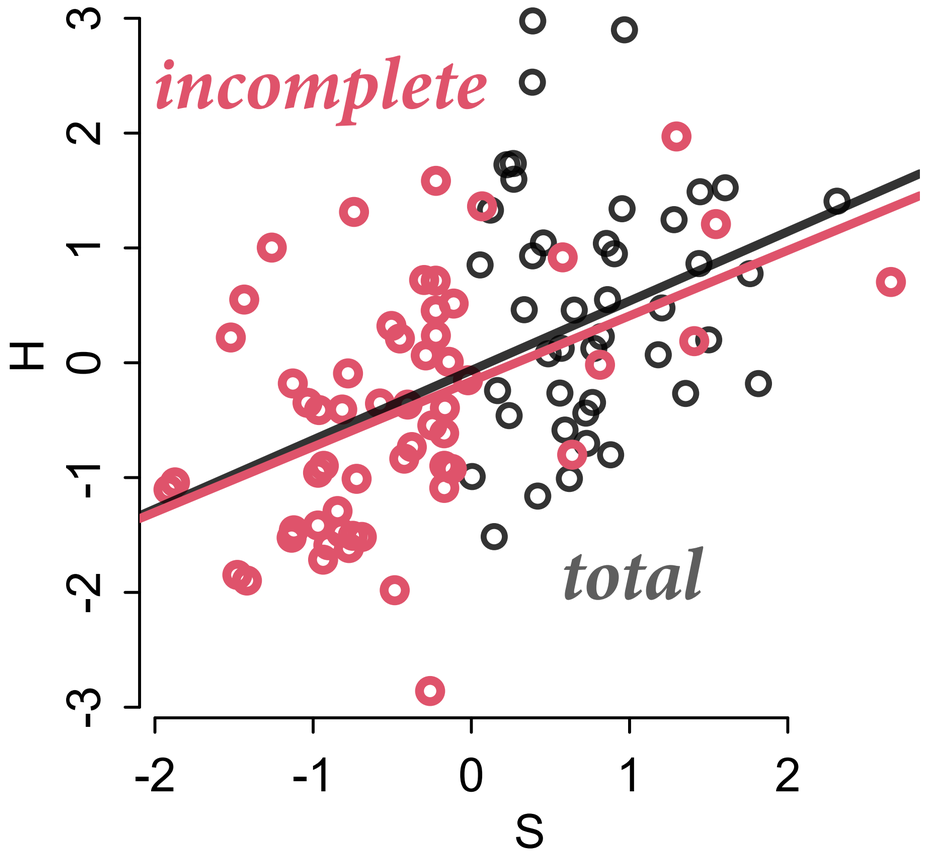


"Dog eats conditional on cause of homework"

Dog path can be benign

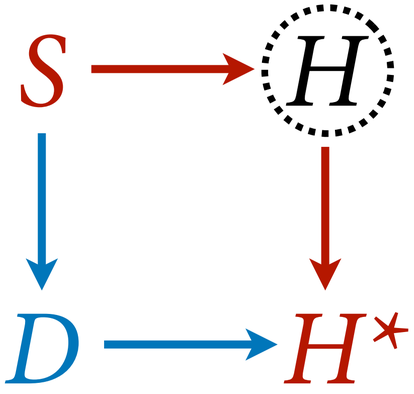
```
# Dog eats homework of students who study too much
# aka missing at random
N <- 100
S <- rnorm(N)
H <- rnorm(N,0.5*S)
# dog eats 80% of homework where S>0
D <- rbern(N, ifelse(S>0,0.8,0) )
Hstar <- H
Hstar[D==1] <- NA

plot( S , H , col=grau(0.8) , lwd=2 )
points( S , Hstar , col=2 , lwd=3 )
```

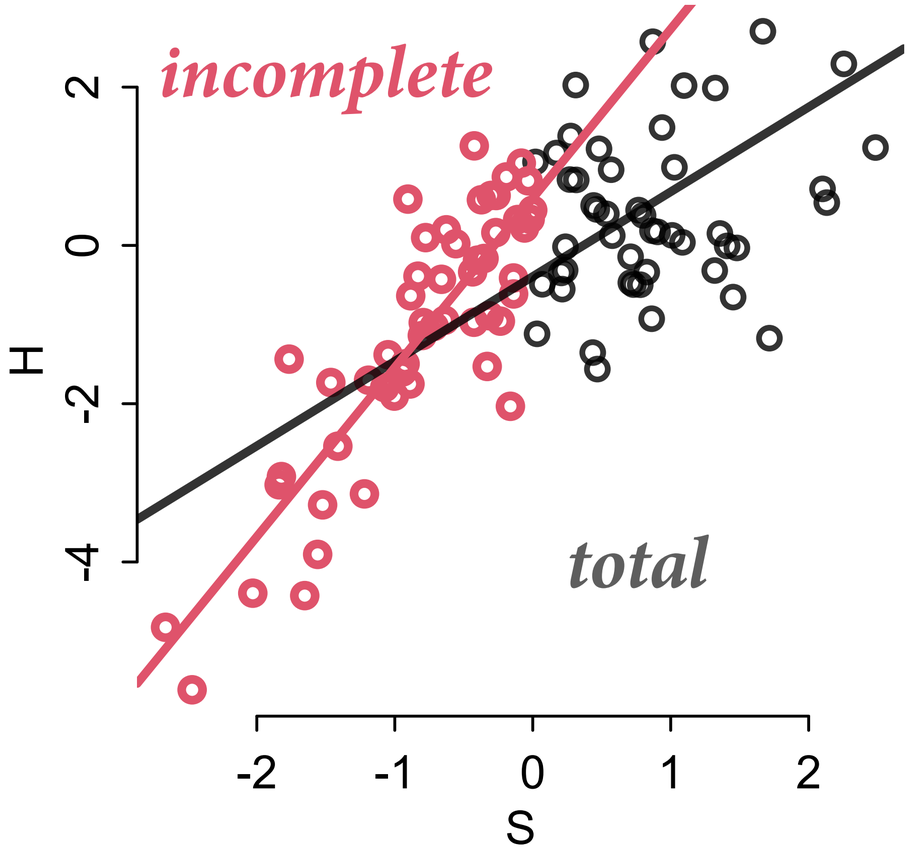


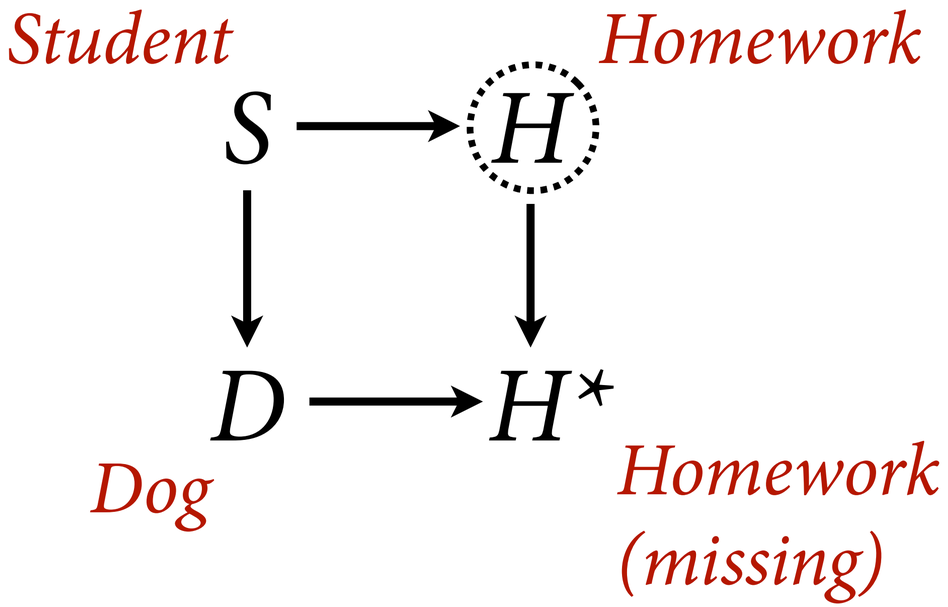


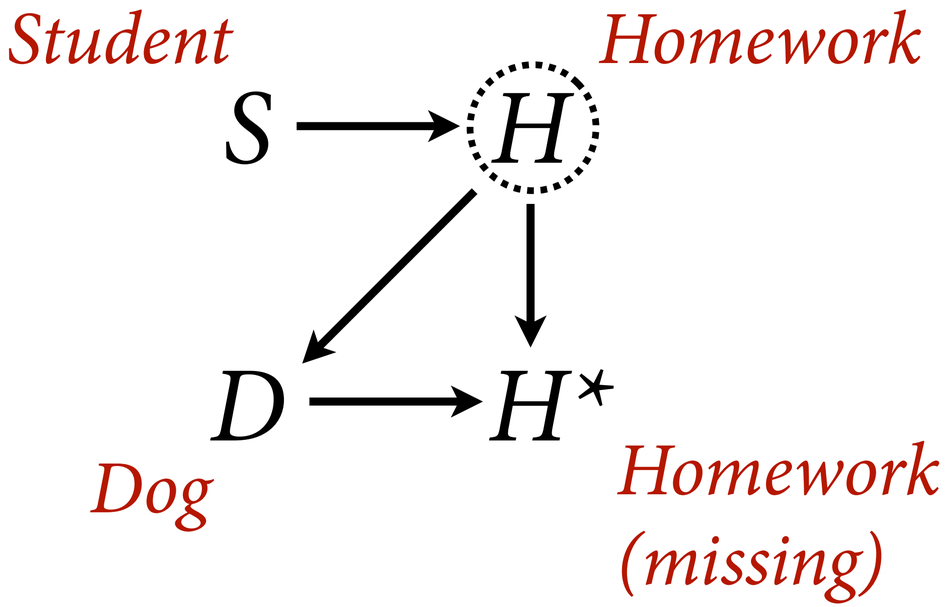
Non-linear relationships and poor modeling, less benign

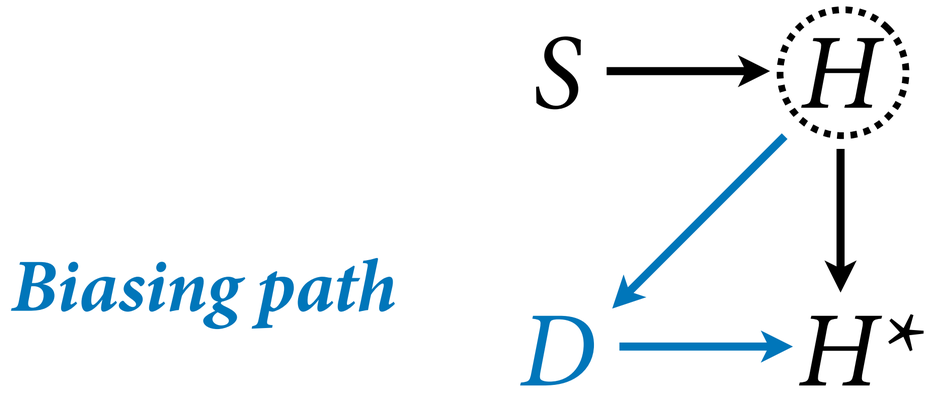
```
# Dog eats homework of students who study too much
# BUT NOW NONLINEAR WITH CEILING EFFECT
N <- 100
S <- rnorm(N)
H <- rnorm(N,(1-exp(-0.7*S)))
# dog eats 100% of homework where S>0
D <- rbern(N, ifelse(S>0,1,0) )
Hstar <- H
Hstar[D==1] <- NA

plot( S , H , col=grau(0.8) , lwd=2 )
points( S , Hstar , col=2 , lwd=3 )
```






Student
Homework
S
H
D
H*
Dog
Homework
(missing)

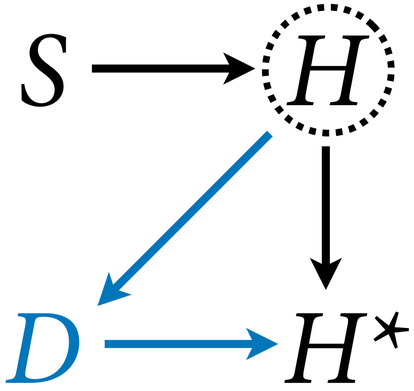

Student
Homework
S
H
D
H*
Dog
Homework
(missing)

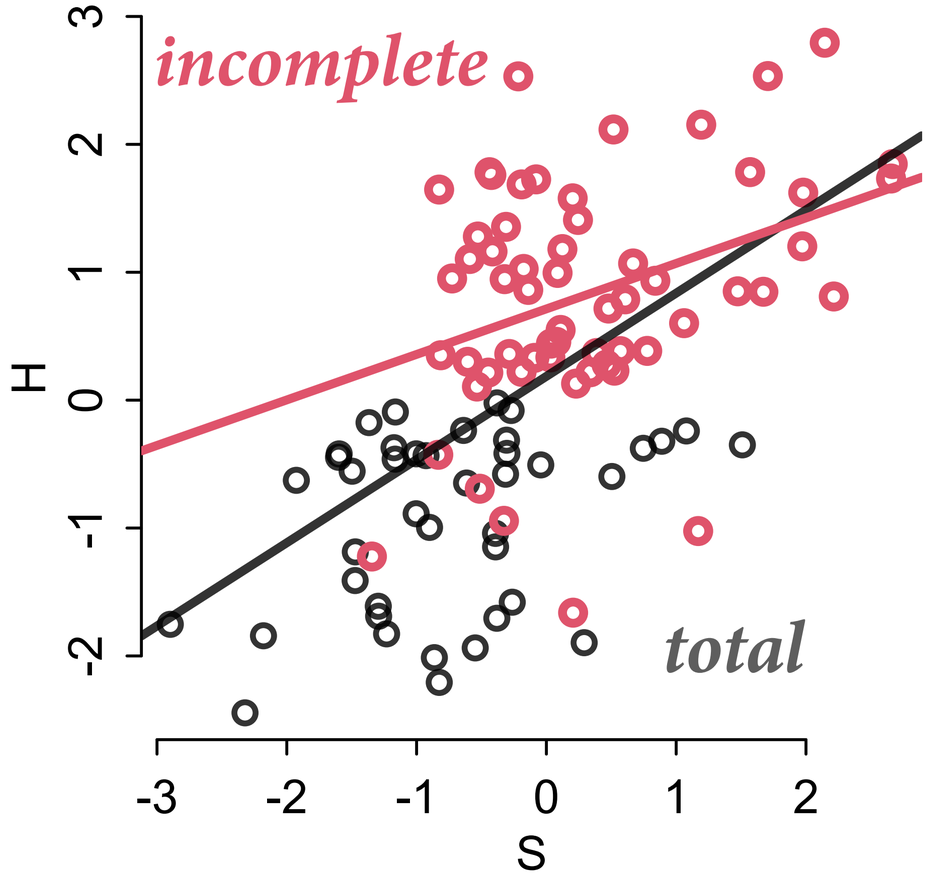


"Dog eats conditional on homework itself"

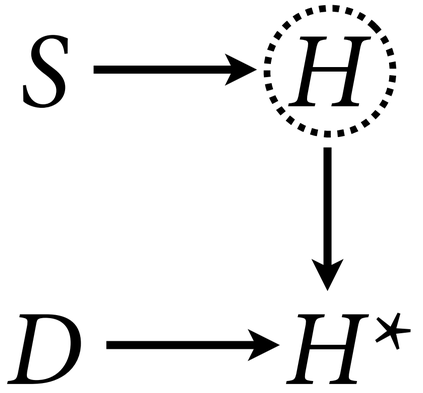
Usually not benign

```
# Dog eats bad homework
# aka missing not at random
N <- 100
S <- rnorm(N)
H <- rnorm(N,0.5*S)
# dog eats 90% of homework where H<0
D <- rbern(N, ifelse(H<0,0.9,0) )
Hstar <- H
Hstar[D==1] <- NA

plot( S , H , col=grau(0.8) , lwd=2 )
points( S , Hstar , col=2 , lwd=3 )
```

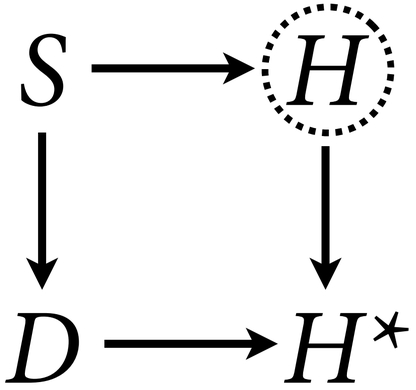




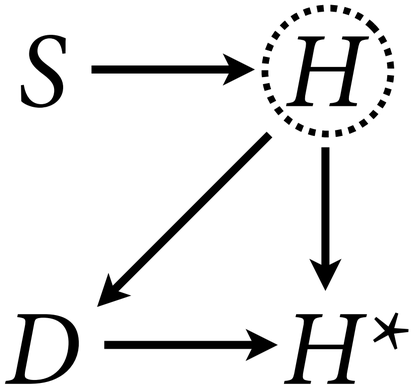
(1) Dog eats random homework:
Dropping incomplete cases okay, but loss of efficiency



(2) Dog eats conditional on cause:
Correctly condition on cause



(3) Dog eats homework itself:
Usually hopeless unless we can model the dog (e.g. survival analysis)



Bayesian Imputation

(1) Dog eats random homework

(2) Dog eats conditional on cause

Both imply need to impute or marginalize over missing values

Bayesian Imputation: Compute posterior probability distribution of missing values

Marginalizing unknowns: Averaging over distribution of missing values

Bayesian Imputation

Causal model of all variables implies strategy for imputation

Technical obstacles exist!

Sometimes imputation is unnecessary, e.g. discrete parameters

Sometimes imputation is easier, e.g. censored observations

Complete marginalization example in book, page 517

15.3.1. Discrete cats. Imagine a neighborhood in which every house contains a songbird. Suppose we survey the neighborhood and sample one minute of song from each house, recording the number of notes. You notice that some houses also have house cats, and wonder if the presence of a cat changes the amount that each bird sings. So you try to also figure out which houses have cats. You can do this easily in some cases, either by seeing the cat or by asking a human resident. But in about 20% of houses, you can't determine whether or not a cat lives there.

This very silly example sets us a very practical working example of how to cope with discrete missing data. We will translate this story into a generative model, simulate data from it, and then build a statistical model that copes with the missing values. Let's consider the story above first as a DAG:

The presence/absence of a cat C influences the number of sung notes N. Because of missing values R_C however, we only observe C^*. To make this into a fully generative model, we must now pick functions for each arrow above. Here are my choices, in statistical notation:

$$N_i \sim \text{Poisson}(\lambda_i) \quad \text{[Probability of notes sung]}$$
$$\log \lambda_i = \alpha + \beta C_i \quad \text{[Rate of notes as function of cat]}$$
$$C_i \sim \text{Bernoulli}(k) \quad \text{[Probability cat is present]}$$
$$R_{C,i} \sim \text{Bernoulli}(r) \quad \text{[Probability of not knowing } C_i\text{]}$$

And then to actually simulate some demonstration data, we'll have to pick values for α, β, k, and r. Here's a working simulation.

R code 15.29

```
set.seed(9)
N_houses <- 100L
alpha <- 5
beta <- (-3)
k <- 0.5
r <- 0.2
```

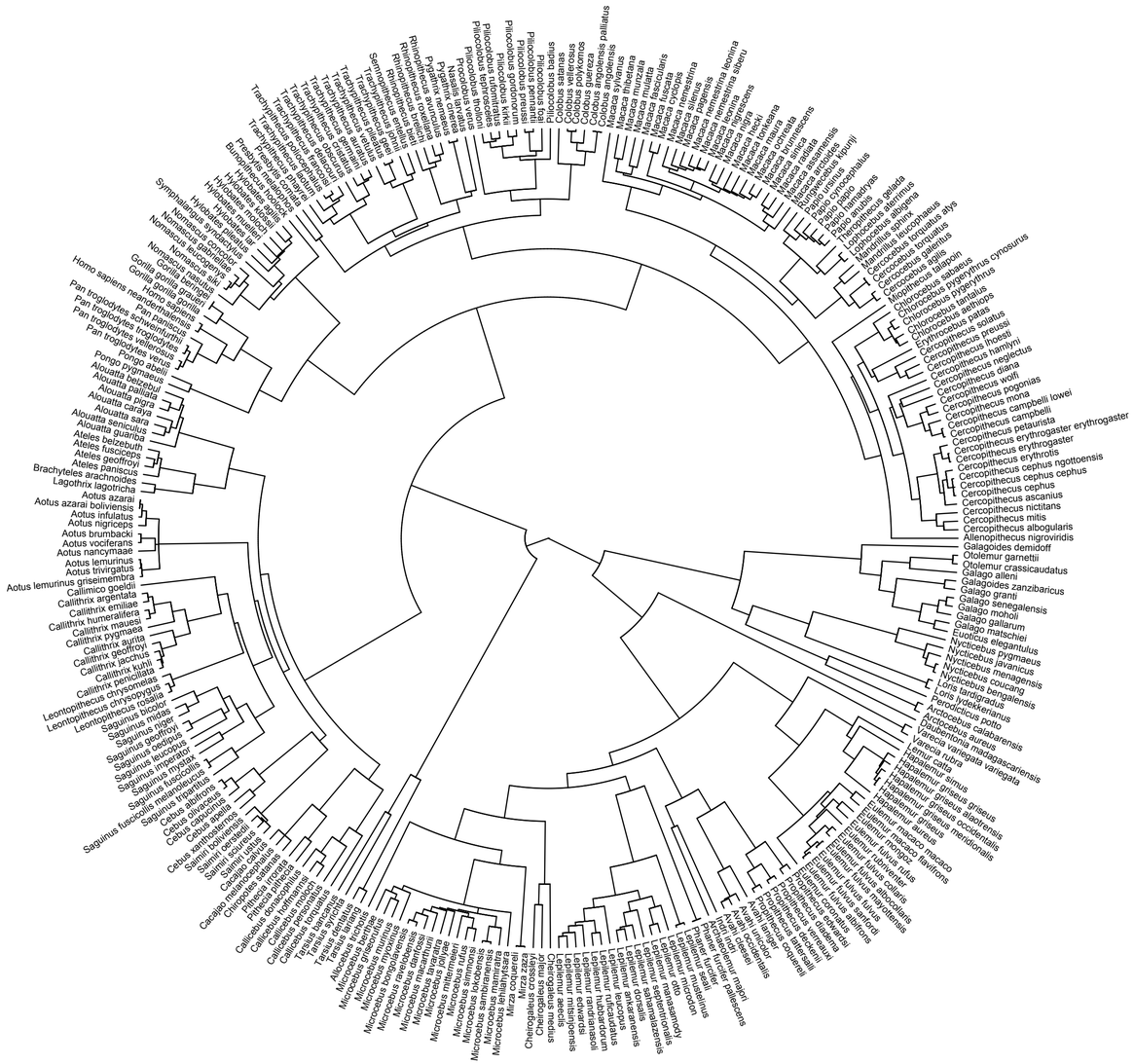
Phylogenetic regression

data(Primates301)

Life history traits

Mass g, brain cc, group size

Much missing data,
measurement error, unobserved
confounding

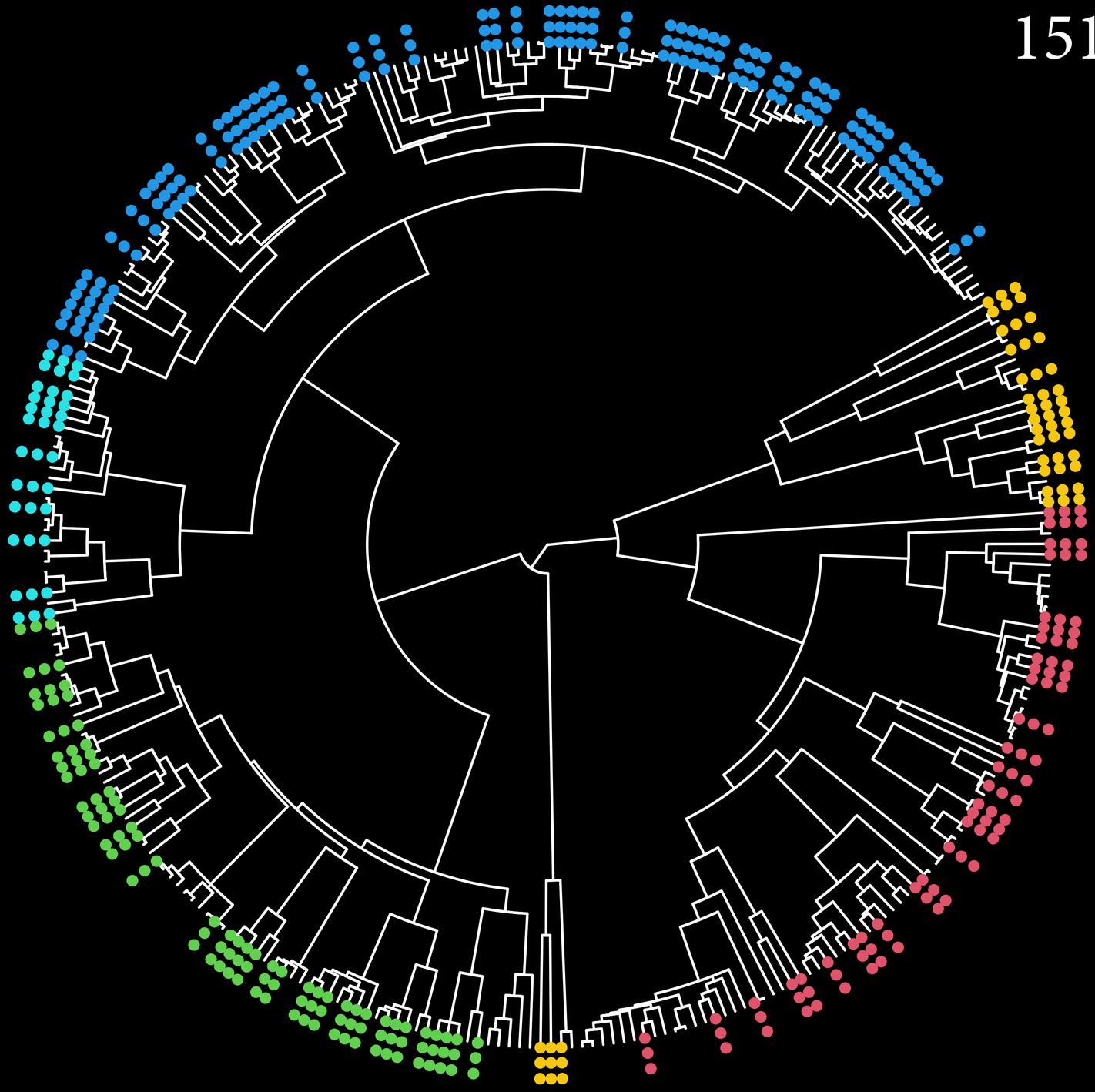


Street et al 2017 Coevolution of cultural intelligence, extended life history, sociality, and brain size in primates

301 species

301 species

151 complete cases



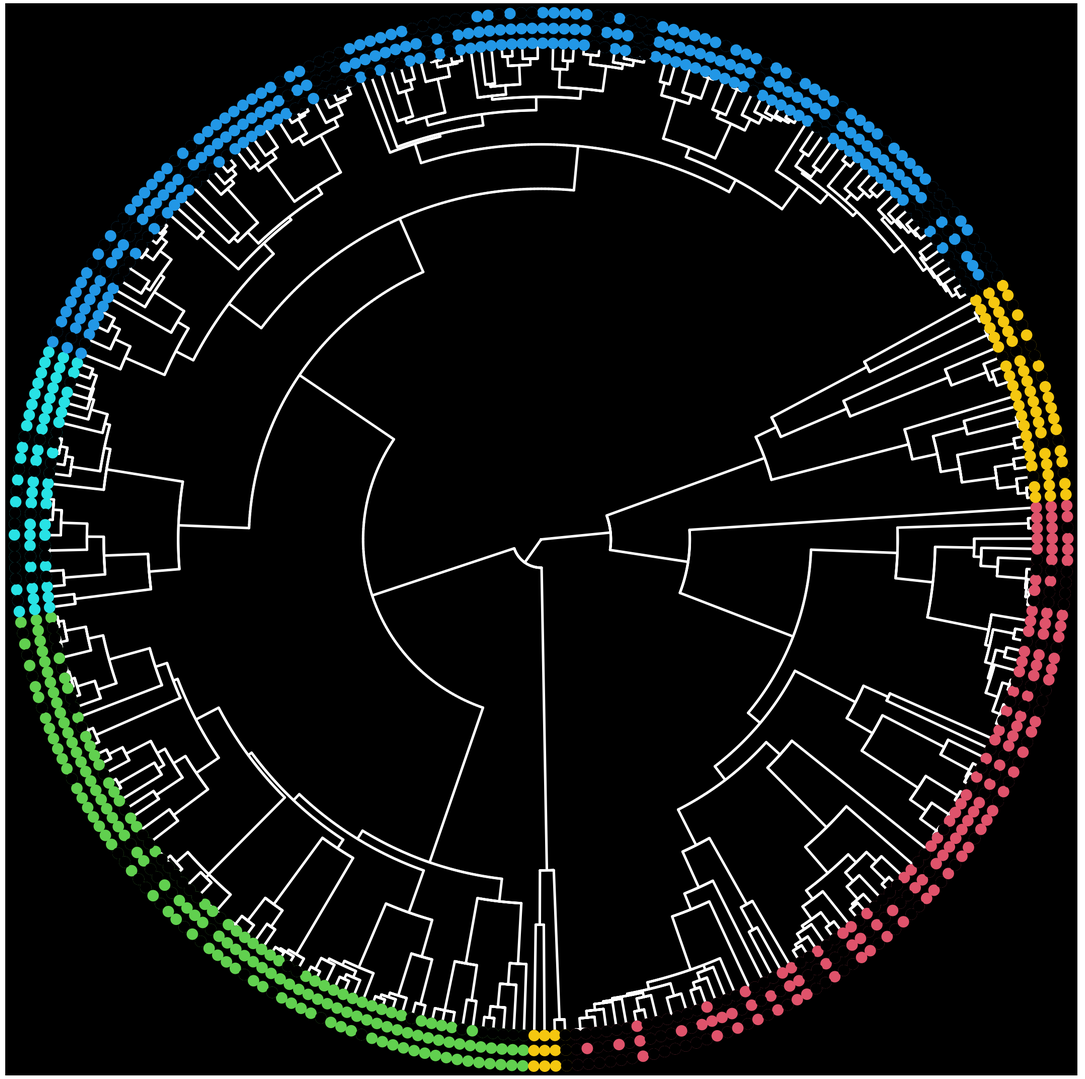
Imputing Primates

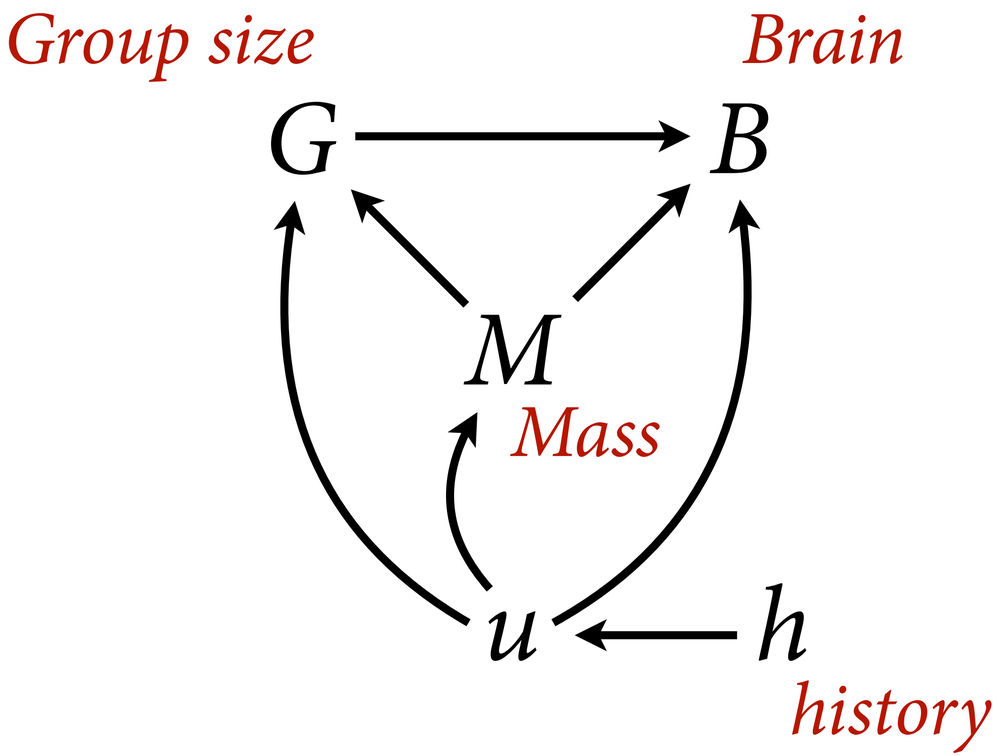
Key idea: Missing values already have probability distributions

Express causal model for each partially-observed variable

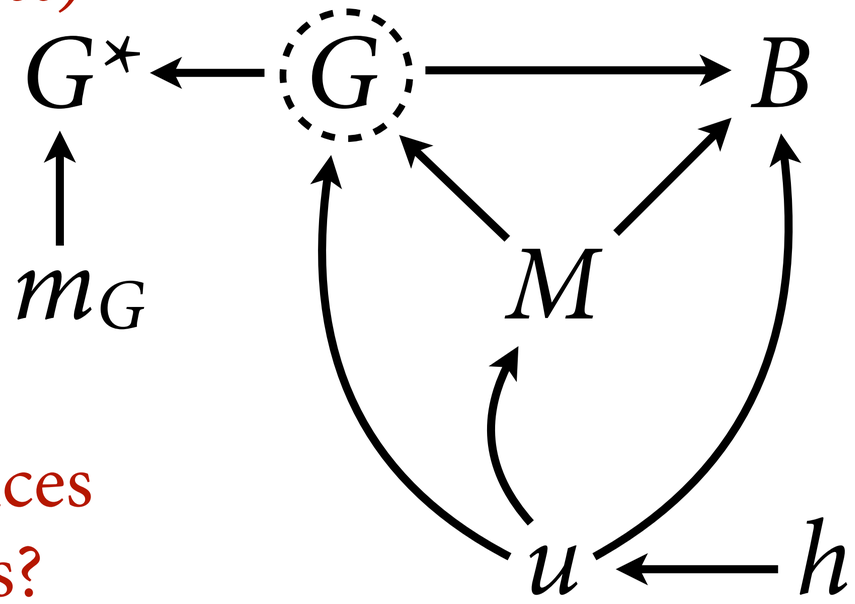
Replace each missing value with a parameter, let model do the rest

Conceptually weird, technically awkward

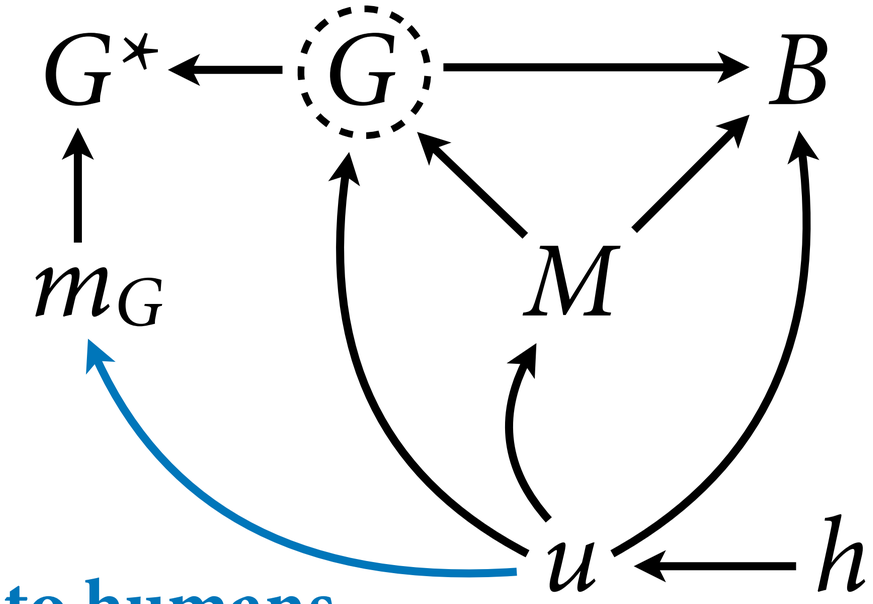



Group size
Brain
G
B
M
Mass
u
h
history

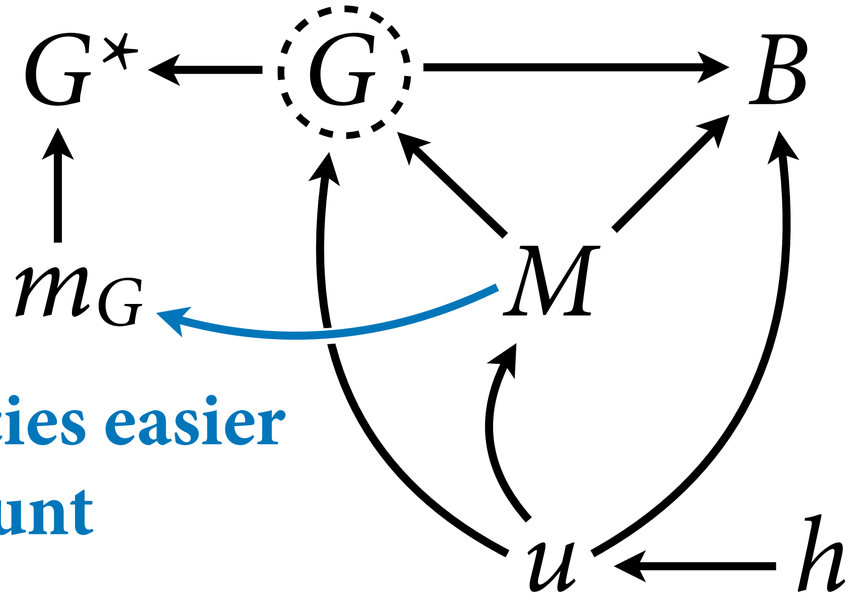
Group size
(missing values)



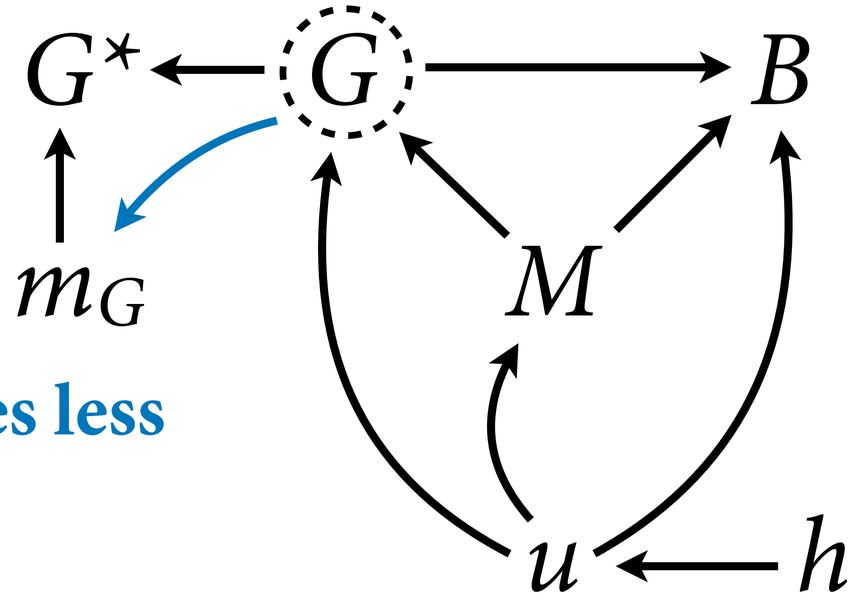
What influences missingness?



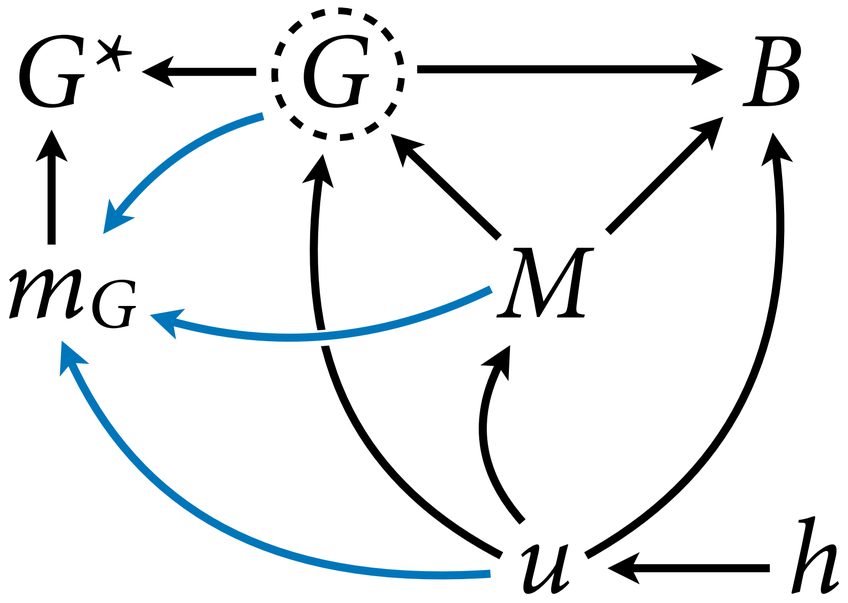
Species close to humans
better studied



Larger species easier to count



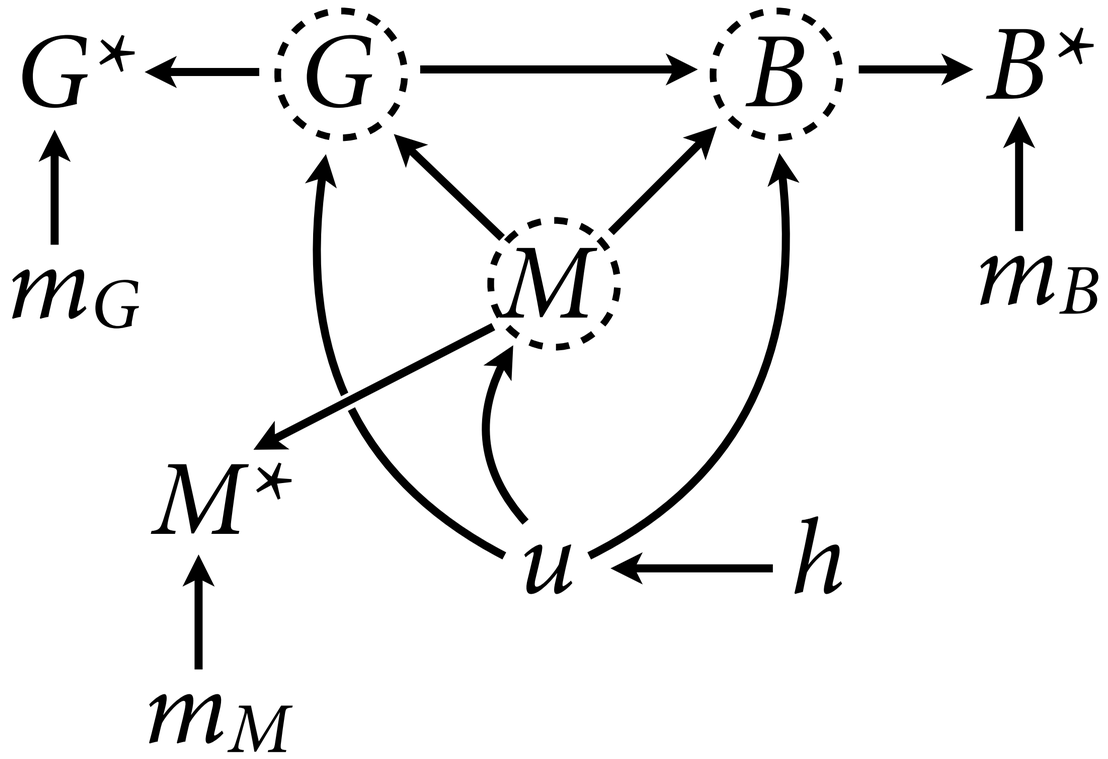
Solitary species less studied



Whatever the assumption, our goal is to use the causal model to infer probability distribution of each missing value.

Uncertainty in each missing value cascades through the entire model.

PAUSE


$G^\star$
G
B
$B^\star$
m_G
M
m_B
$M^\star$
u
h
m_M

$$B \sim \text{MVNormal}(\mu, \mathbf{K})$$

$$\mu_i = \alpha + \beta_G G_i + \beta_M M_i$$

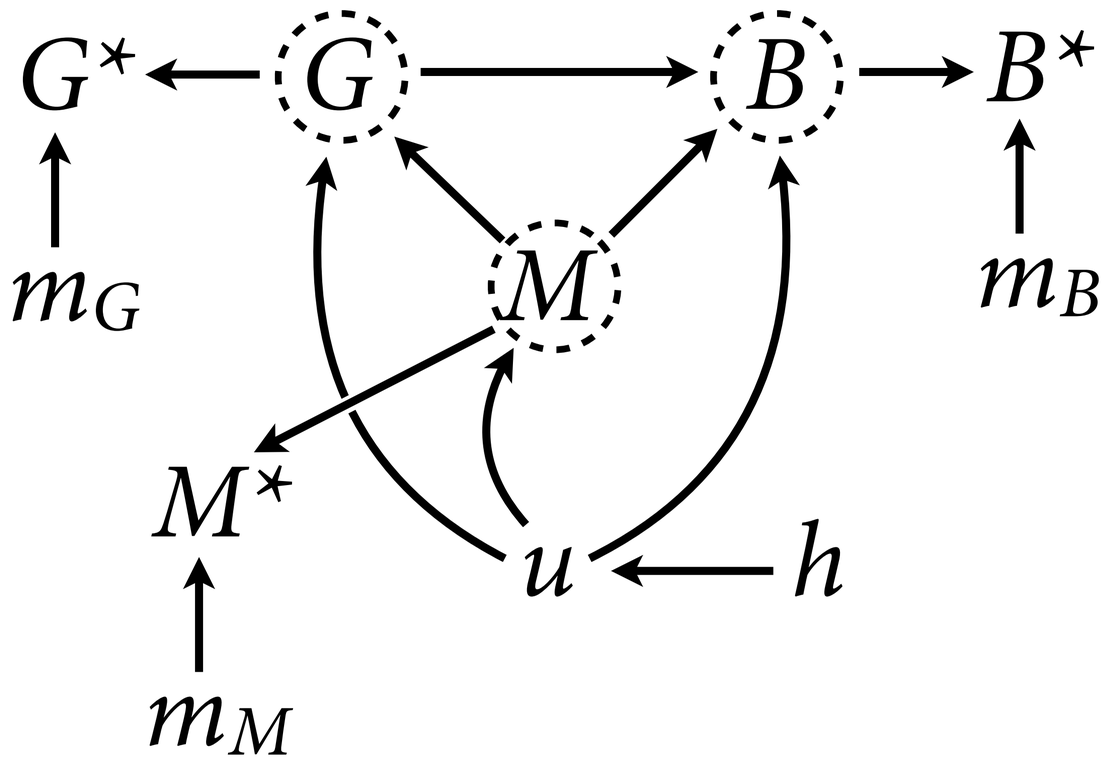
$$\mathbf{K} = \eta^2 \exp(-\rho d_{i,j})$$

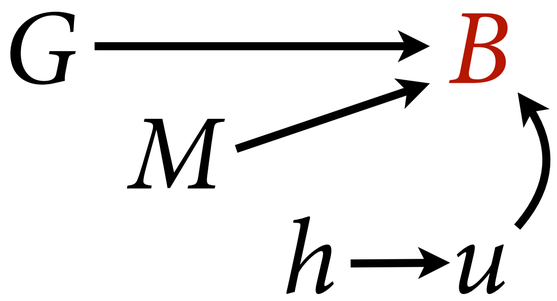
$$\alpha \sim \text{Normal}(0,1)$$

$$\beta_G, \beta_M \sim \text{Normal}(0,0.5)$$

$$\eta^2 \sim \text{HalfNormal}(1,0.25)$$

$$\rho \sim \text{HalfNormal}(3,0.25)$$





$$B \sim \text{MVNormal}(\mu, \mathbf{K})$$

$$\mu_i = \alpha + \beta_G G_i + \beta_M M_i$$

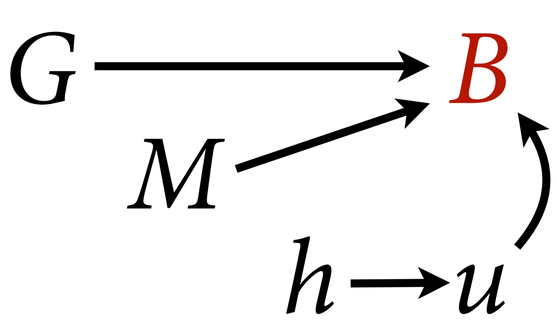
$$\mathbf{K} = \eta^2 \exp(-\rho d_{i,j})$$

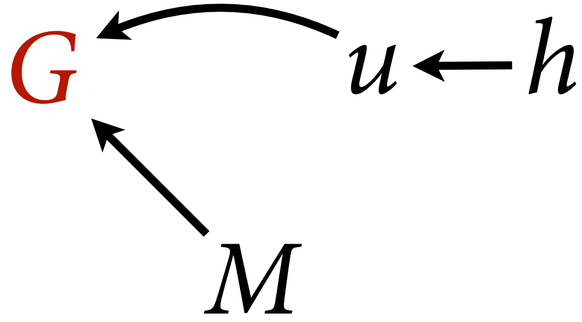
$$\alpha \sim \text{Normal}(0,1)$$

$$\beta_G, \beta_M \sim \text{Normal}(0,0.5)$$

$$\eta^2 \sim \text{HalfNormal}(1,0.25)$$

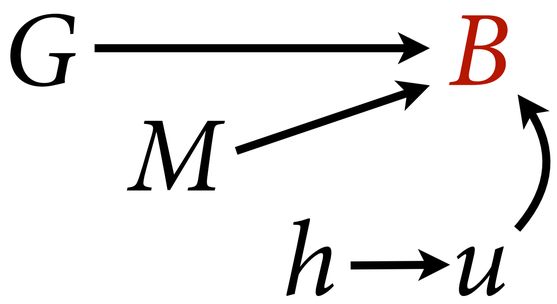
$$\rho \sim \text{HalfNormal}(3,0.25)$$

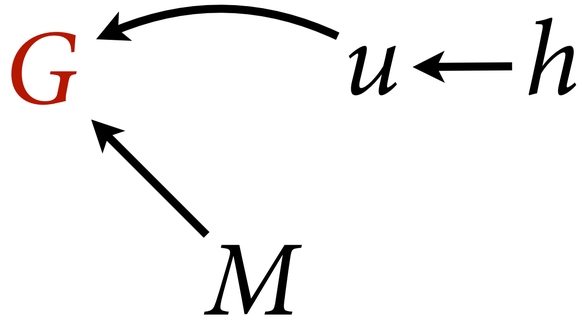


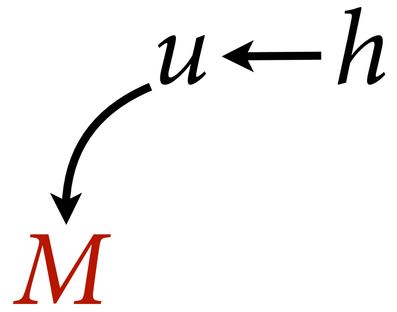


$$B \sim \text{MVNormal}(\mu, \mathbf{K})$$
$$\mu_i = \alpha + \beta_G G_i + \beta_M M_i$$
$$\mathbf{K} = \eta^2 \exp(-\rho d_{i,j})$$
$$\alpha \sim \text{Normal}(0,1)$$
$$\beta_G, \beta_M \sim \text{Normal}(0,0.5)$$
$$\eta^2 \sim \text{HalfNormal}(1,0.25)$$
$$\rho \sim \text{HalfNormal}(3,0.25)$$

$$G \sim \text{MVNormal}(\nu, \mathbf{K}_G)$$
$$\nu_i = \alpha_G + \beta_{MG} M_i$$
$$\mathbf{K}_G = \eta_G^2 \exp(-\rho_G d_{i,j})$$
$$\alpha_G \sim \text{Normal}(0,1)$$
$$\beta_{MG} \sim \text{Normal}(0,0.5)$$
$$\eta_G^2 \sim \text{HalfNormal}(1,0.25)$$
$$\rho_G \sim \text{HalfNormal}(3,0.25)$$







$$B \sim \text{MVNormal}(\mu, \mathbf{K})$$
$$\mu_i = \alpha + \beta_G G_i + \beta_M M_i$$
$$\mathbf{K} = \eta^2 \exp(-\rho d_{i,j})$$
$$\alpha \sim \text{Normal}(0,1)$$
$$\beta_G, \beta_M \sim \text{Normal}(0,0.5)$$
$$\eta^2 \sim \text{HalfNormal}(1,0.25)$$
$$\rho \sim \text{HalfNormal}(3,0.25)$$

$$G \sim \text{MVNormal}(\nu, \mathbf{K}_G)$$
$$\nu_i = \alpha_G + \beta_{MG} M_i$$
$$\mathbf{K}_G = \eta_G^2 \exp(-\rho_G d_{i,j})$$
$$\alpha_G \sim \text{Normal}(0,1)$$
$$\beta_{MG} \sim \text{Normal}(0,0.5)$$
$$\eta_G^2 \sim \text{HalfNormal}(1,0.25)$$
$$\rho_G \sim \text{HalfNormal}(3,0.25)$$

$$M \sim \text{MVNormal}(0, \mathbf{K}_M)$$
$$\mathbf{K}_M = \eta_M^2 \exp(-\rho_M d_{i,j})$$
$$\eta_M^2 \sim \text{HalfNormal}(1,0.25)$$
$$\rho_M \sim \text{HalfNormal}(3,0.25)$$

Draw the Missing Owl

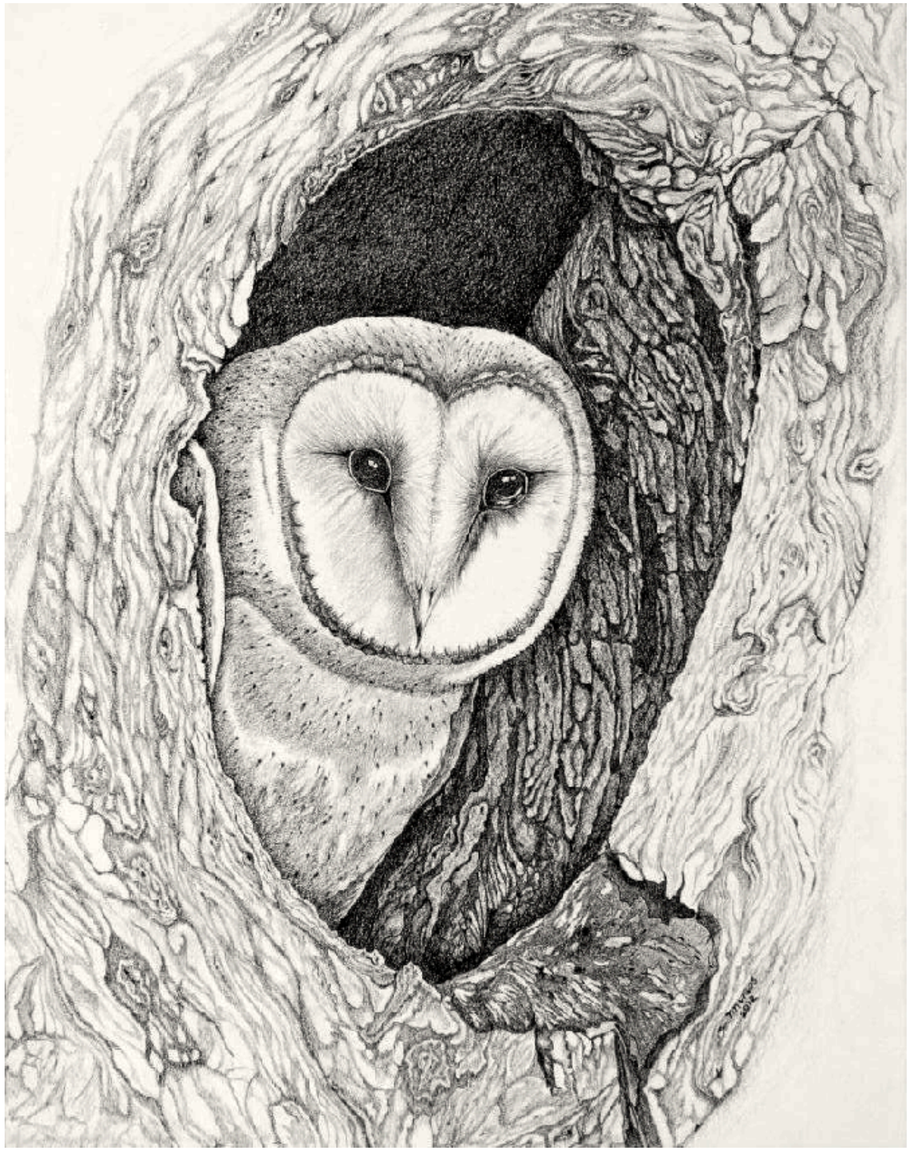
Let's take it slow…

(1) Ignore cases with missing *B* values (for now)

(2) Impute *G* and *M* ignoring models for each

(3) Impute *G* using model

(4) Impute *B*, *G*, *M* using model



Draw the Missing Owl

Let's take it slow…

(1) Ignore cases with missing *B* values
(for now)

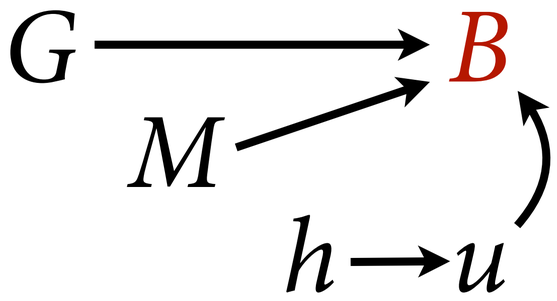
(2) Impute *G* and *M* ig
for each

(3) Impute *G* using model

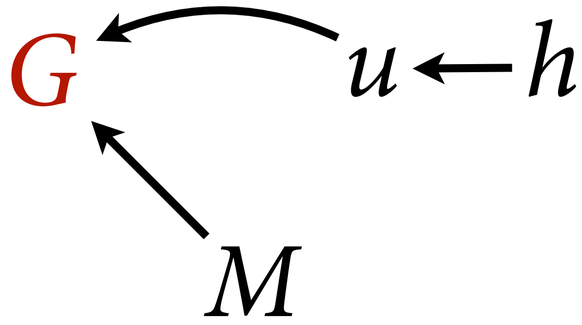
(4) Impute *B*, *G*, *M* using model

```
> dd <- d[complete.cases(d$brain),]
> table( M=!is.na(dd$body) , G=!is.na(dd$group_size) )
       G
M        FALSE TRUE
  FALSE      2    0
  TRUE      31  151
```

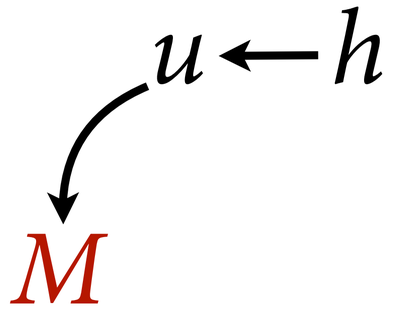
(2) Impute G and M ignoring models for each



$$B \sim \text{MVNormal}(\mu, \mathbf{K})$$
$$\mu_i = \alpha + \beta_G G_i + \beta_M M_i$$
$$\mathbf{K} = \eta^2 \exp(-\rho d_{i,j})$$
$$\alpha \sim \text{Normal}(0,1)$$
$$\beta_G, \beta_M \sim \text{Normal}(0,0.5)$$
$$\eta^2 \sim \text{HalfNormal}(1,0.25)$$
$$\rho \sim \text{HalfNormal}(3,0.25)$$

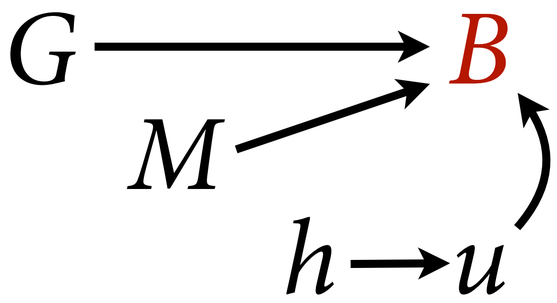


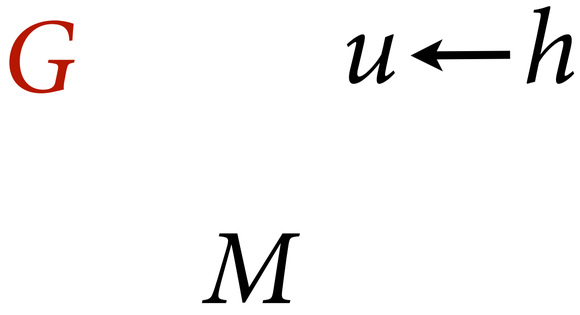
$$G \sim \text{MVNormal}(\nu, \mathbf{K}_G)$$
$$\nu_i = \alpha_G + \beta_{MG} M_i$$
$$\mathbf{K}_G = \eta_G^2 \exp(-\rho_G d_{i,j})$$
$$\alpha_G \sim \text{Normal}(0,1)$$
$$\beta_{MG} \sim \text{Normal}(0,0.5)$$
$$\eta_G^2 \sim \text{HalfNormal}(1,0.25)$$
$$\rho_G \sim \text{HalfNormal}(3,0.25)$$

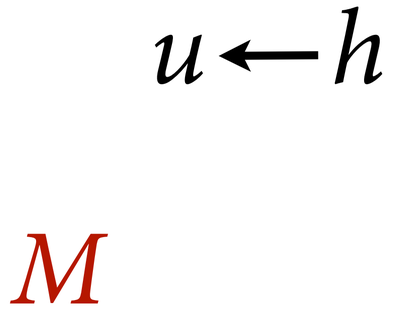


$$M \sim \text{MVNormal}(0, \mathbf{K}_M)$$
$$\mathbf{K}_M = \eta_M^2 \exp(-\rho_M d_{i,j})$$
$$\eta_M^2 \sim \text{HalfNormal}(1,0.25)$$
$$\rho_M \sim \text{HalfNormal}(3,0.25)$$

(2) Impute G and M ignoring models for each







$$B \sim \text{MVNormal}(\mu, \mathbf{K})$$

$$\mu_i = \alpha + \beta_G G_i + \beta_M M_i$$

$$\mathbf{K} = \eta^2 \exp(-\rho d_{i,j})$$

$$\alpha \sim \text{Normal}(0,1)$$

$$\beta_G, \beta_M \sim \text{Normal}(0,0.5)$$

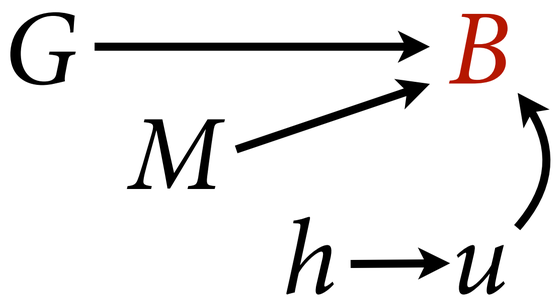
$$\eta^2 \sim \text{HalfNormal}(1,0.25)$$

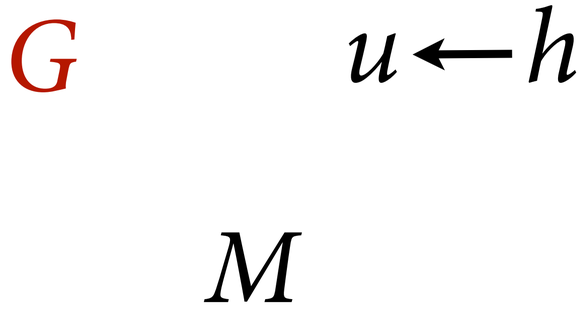
$$\rho \sim \text{HalfNormal}(3,0.25)$$

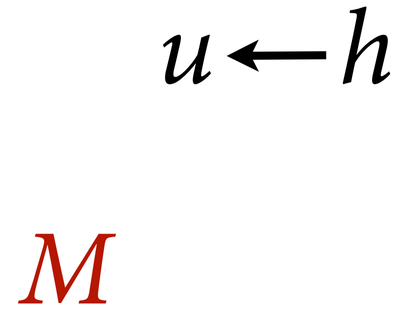
$$G_i \sim \text{Normal}(0,1)$$

$$M_i \sim \text{Normal}(0,1)$$

(2) Impute G and M ignoring models for each







$$B \sim \text{MVNormal}(\mu, \mathbf{K})$$

$$\mu_i = \alpha + \beta_G G_i + \beta_M M_i$$

$$\mathbf{K} = \eta^2 \exp(-\rho d_{i,j})$$

$$\alpha \sim \text{Normal}(0,1)$$

$$\beta_G, \beta_M \sim \text{Normal}(0,0.5)$$

$$\eta^2 \sim \text{HalfNormal}(1,0.25)$$

$$\rho \sim \text{HalfNormal}(3,0.25)$$

$$G_i \sim \text{Normal}(0,1)$$

$$M_i \sim \text{Normal}(0,1)$$

When G_i observed, likelihood for standardized variable

When G_i missing, prior

(2) Impute G and M ignoring models for each

```
# imputation ignoring models of M and G
fMBG_OU <- alist(
    B ~ multi_normal( mu , K ),
    mu <- a + bM*M + bG*G,
    matrix[N_spp,N_spp]:K <- cov_GPL1(Dmat,etasq,rho,0.01),
    M ~ normal(0,1),
    G ~ normal(0,1),
    a ~ normal( 0 , 1 ),
    c(bM,bG) ~ normal( 0 , 0.5 ),
    etasq ~ half_normal(1,0.25),
    rho ~ half_normal(3,0.25)
)
mBMG_OU <- ulam( fMBG_OU , data=dat_all,chains=4,cores=4 )
```

$$B \sim \text{MVNormal}(\mu, \mathbf{K})$$

$$\mu_i = \alpha + \beta_G G_i + \beta_M M_i$$

$$\mathbf{K} = \eta^2 \exp(-\rho d_{i,j})$$

$$G_i \sim \text{Normal}(0,1)$$

$$M_i \sim \text{Normal}(0,1)$$

$$\alpha \sim \text{Normal}(0,1)$$

$$\beta_G, \beta_M \sim \text{Normal}(0,0.5)$$

$$\eta^2 \sim \text{HalfNormal}(1,0.25)$$

$$\rho \sim \text{HalfNormal}(3,0.25)$$

(2) Impute G and M ignoring models for each

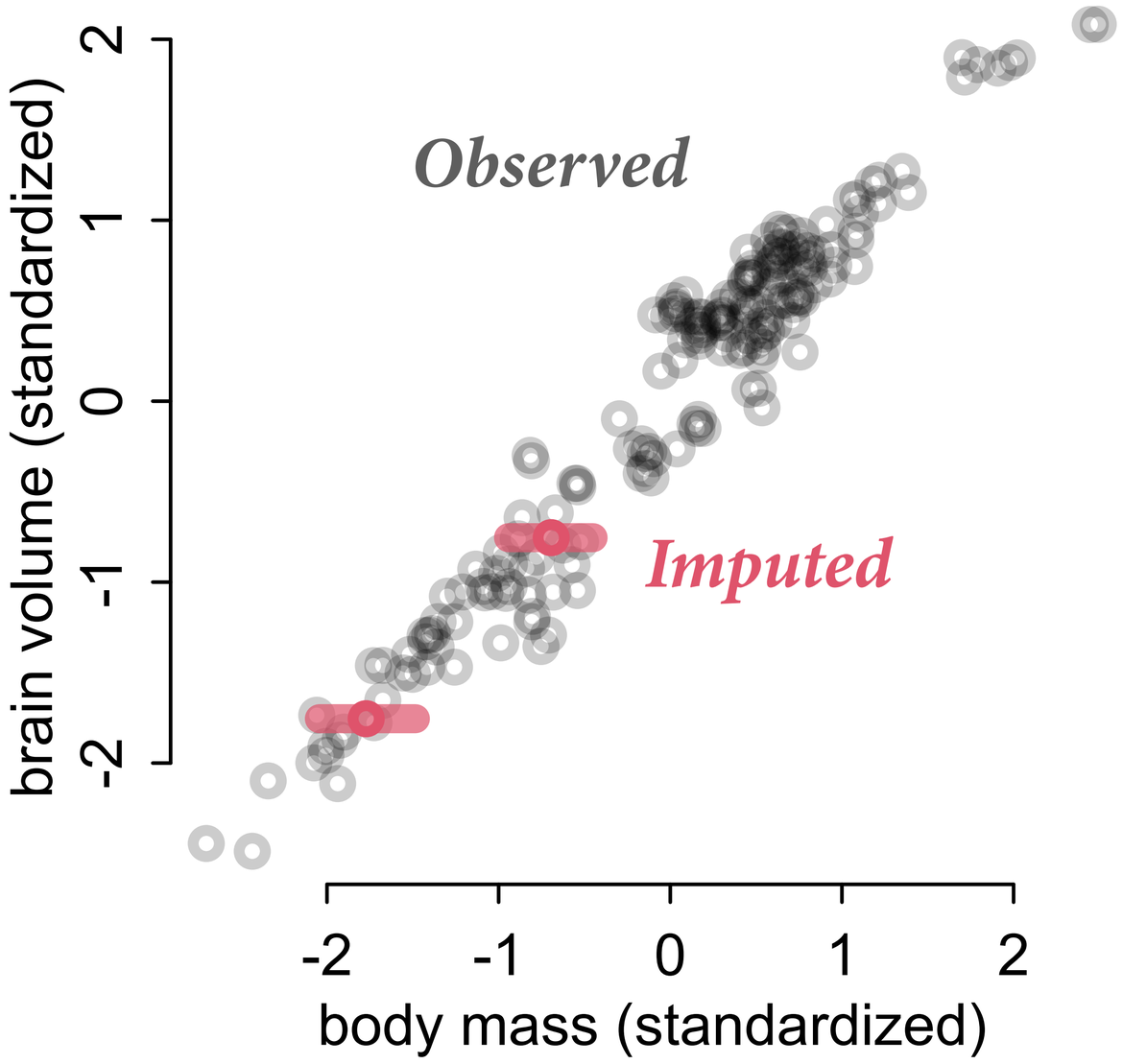
```
# imputation ignoring models of M and G
fMBG_OU <- alist(
    B ~ multi_normal( mu , K ),
    mu <- a + bM*M + bG*G,
    matrix[N_spp,N_spp]:K <- cov_GPL1(Dmat,etasq,rho,0.01),
    M ~ normal(0,1),
    G ~ normal(0,1),
    a ~ normal( 0 , 1 ),
    c(bM,bG) ~ normal( 0 , 0.5 ),
    etasq ~ half_normal(1,0.25),
    rho ~ half_normal(3,0.25)
)
mBMG_OU <- ulam( fMBG_OU , data=dat_all,chains=4,cores=4 )
```

```
> precis(mBMG_OU,2)
                mean   sd  5.5% 94.5% n_eff Rhat4
a              -0.10 0.08 -0.22  0.02  4569     1
bG              0.01 0.02 -0.01  0.04  2816     1
bM              0.82 0.03  0.78  0.86  3949     1
etasq           0.04 0.01  0.03  0.05  3733     1
rho             2.75 0.26  2.36  3.17  3931     1
M_impute[1]    -1.77 0.18 -2.04 -1.48  5087     1
M_impute[2]    -0.69 0.15 -0.94 -0.45  5065     1
G_impute[1]     0.01 1.01 -1.61  1.63  6311     1
G_impute[2]     0.05 1.02 -1.55  1.67  5616     1
G_impute[3]     0.03 1.04 -1.64  1.71  4876     1
G_impute[4]     0.05 0.95 -1.45  1.56  4978     1
G_impute[5]     0.09 1.03 -1.53  1.69  5244     1
G_impute[6]     0.18 1.02 -1.48  1.79  5414     1
G_impute[7]     0.06 1.00 -1.57  1.65  6207     1
G_impute[8]     0.13 1.00 -1.47  1.72  6329     1
G_impute[9]     0.01 0.98 -1.62  1.62  5942     1
G_impute[10]    0.03 1.00 -1.56  1.67  5197     1
G_impute[11]    0.06 0.95 -1.47  1.60  5057     1
G_impute[12]    0.08 0.97 -1.49  1.63  4907     1
G_impute[13]   -0.03 1.01 -1.65  1.58  5958     1
G_impute[14]    0.04 0.99 -1.52  1.65  5538     1
G_impute[15]   -0.18 1.00 -1.77  1.37  5705     1
G_impute[16]   -0.09 0.97 -1.63  1.45  4026     1
G_impute[17]   -0.04 1.02 -1.74  1.58  6465     1
G_impute[18]   -0.09 1.03 -1.73  1.54  5632     1
G_impute[19]   -0.19 1.01 -1.78  1.45  4251     1
G_impute[20]   -0.01 1.02 -1.62  1.65  6602     1
G_impute[21]   -0.07 0.95 -1.56  1.47  5640     1
G_impute[22]   -0.04 0.99 -1.60  1.55  6602     1
G_impute[23]    0.14 1.03 -1.45  1.82  5644     1
G_impute[24]   -0.07 0.99 -1.65  1.46  4249     1
G_impute[25]   -0.06 1.04 -1.72  1.62  5993     1
G_impute[26]    0.19 1.01 -1.40  1.78  4973     1
G_impute[27]    0.08 1.04 -1.54  1.70  5765     1
G_impute[28]   -0.07 0.99 -1.66  1.52  3609     1
G_impute[29]    0.04 0.99 -1.55  1.63  4542     1
G_impute[30]   -0.01 0.99 -1.61  1.56  4821     1
G_impute[31]    0.01 1.02 -1.65  1.66  6602     1
G_impute[32]    0.08 1.01 -1.51  1.66  5420     1
G_impute[33]   -0.02 0.99 -1.58  1.60  4241     1
```

5)

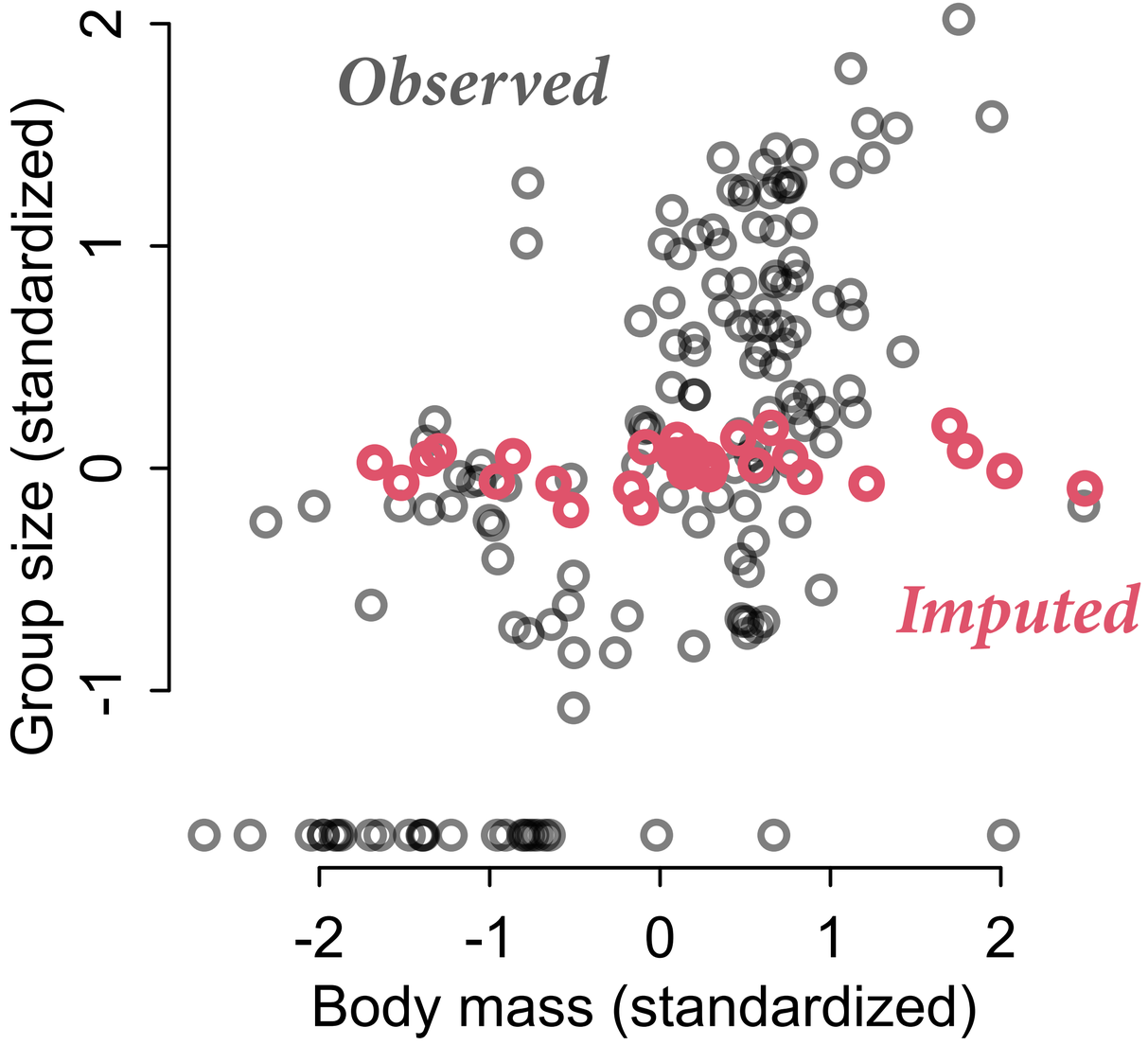
5)

(2) Impute *G* and *M* ignoring models for each



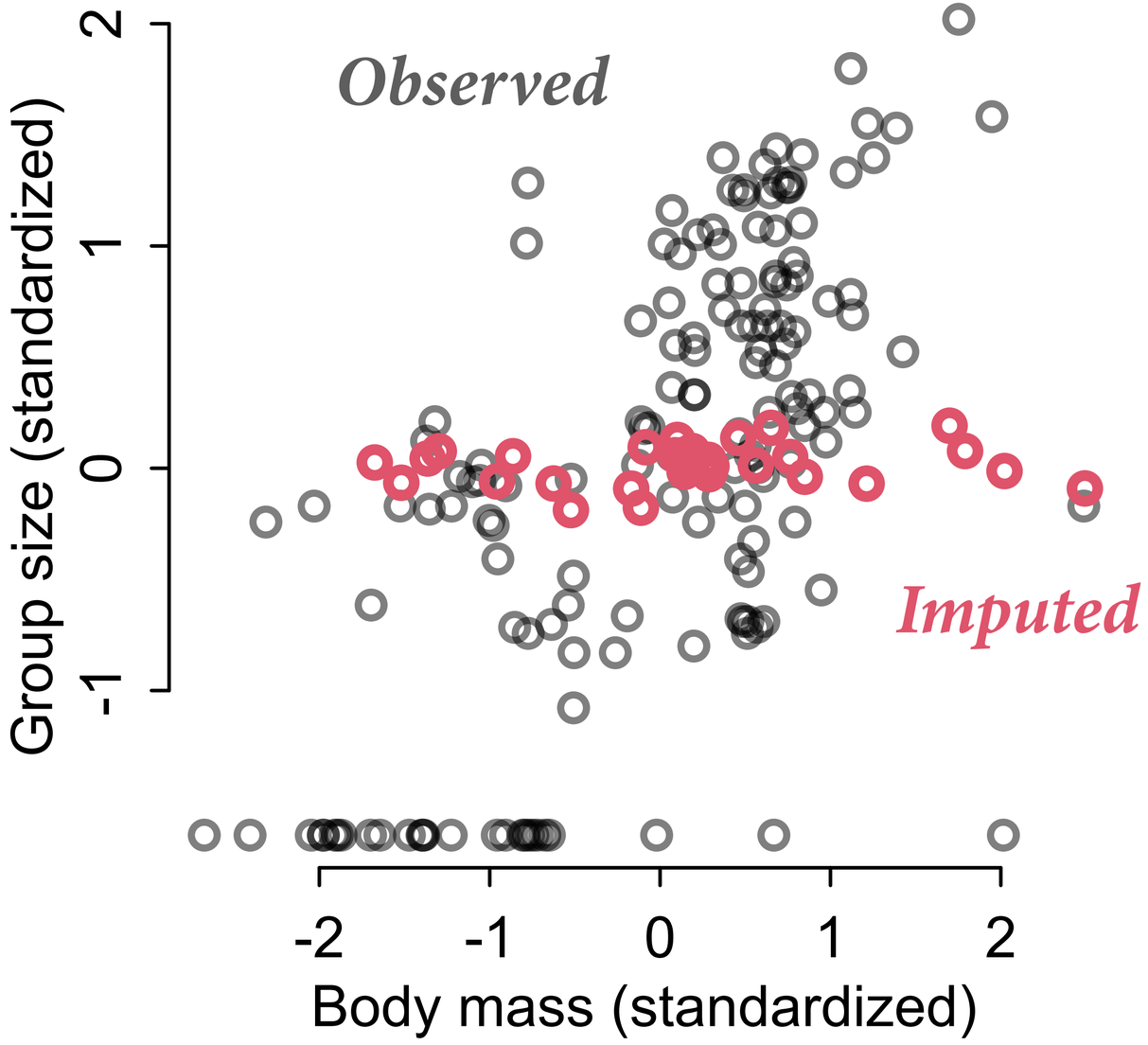
Because *M* strongly associated with *B*, **imputed *M* values** follow the regression relationship

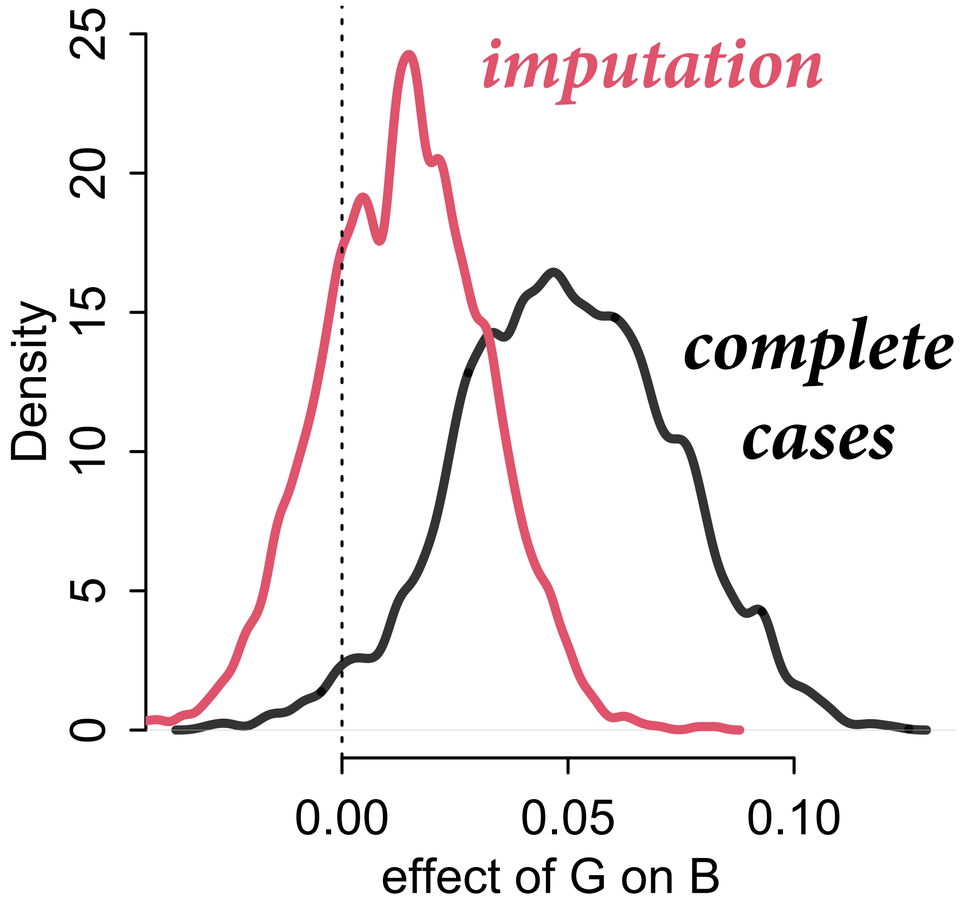
(2) Impute G and M ignoring models for each



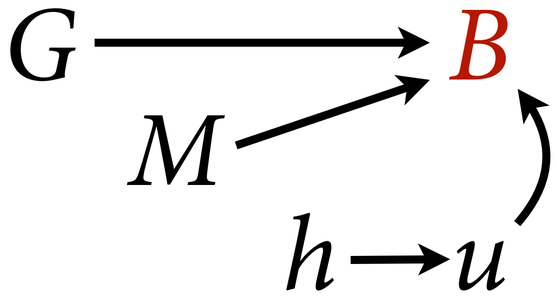
Because association between M and G not modeled, **imputed G values** do not follow the regression relationship

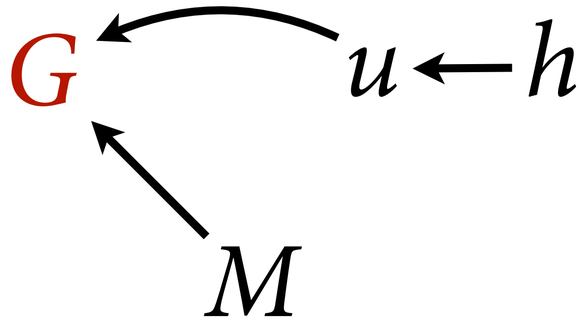
(2) Impute *G* and *M* ignoring models for each

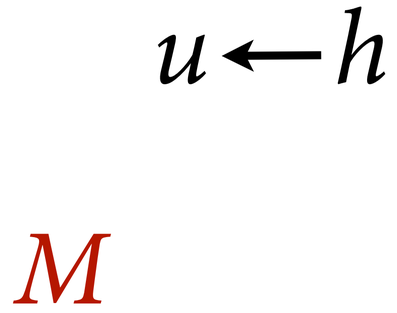




(3) Impute G using model







$$B \sim \text{MVNormal}(\mu, \mathbf{K})$$
$$\mu_i = \alpha + \beta_G G_i + \beta_M M_i$$
$$\mathbf{K} = \eta^2 \exp(-\rho d_{i,j})$$
$$\alpha \sim \text{Normal}(0,1)$$
$$\beta_G, \beta_M \sim \text{Normal}(0,0.5)$$
$$\eta^2 \sim \text{HalfNormal}(1,0.25)$$
$$\rho \sim \text{HalfNormal}(3,0.25)$$

$$G \sim \text{MVNormal}(\nu, \mathbf{K}_G)$$
$$\nu_i = \alpha_G + \beta_{MG} M_i$$
$$\mathbf{K}_G = \eta_G^2 \exp(-\rho_G d_{i,j})$$
$$\alpha_G \sim \text{Normal}(0,1)$$
$$\beta_{MG} \sim \text{Normal}(0,0.5)$$
$$\eta_G^2 \sim \text{HalfNormal}(1,0.25)$$
$$\rho_G \sim \text{HalfNormal}(3,0.25)$$

$$M_i \sim \text{Normal}(0,1)$$

```
# no phylogeny on G but have submodel M -> G
mBMG_OU_G <- ulam(
    alist(
        B ~ multi_normal( mu , K ),
        mu <- a + bM*M + bG*G,
        G ~ normal(nu,sigma),
        nu <- aG + bMG*M,
        M ~ normal(0,1),
        matrix[N_spp,N_spp]:K <- cov_GPL1(Dmat,etasq,rho,0.01),
        c(a,aG) ~ normal( 0 , 1 ),
        c(bM,bG,bMG) ~ normal( 0 , 0.5 ),
        c(etasq) ~ half_normal(1,0.25),
        c(rho) ~ half_normal(3,0.25),
        sigma ~ exponential(1)
    ), data=dat_all , chains=4 , cores=4 , sample=TRUE )

# phylogeny information for G imputation (but no M -> G model)
mBMG_OU2 <- ulam(
    alist(
        B ~ multi_normal( mu , K ),
        mu <- a + bM*M + bG*G,
        M ~ normal(0,1),
        G ~ multi_normal( 'rep_vector(0,N_spp)' ,KG),
        matrix[N_spp,N_spp]:K <- cov_GPL1(Dmat,etasq,rho,0.01),
        matrix[N_spp,N_spp]:KG <- cov_GPL1(Dmat,etasqG,rhoG,0.01),
        a ~ normal( 0 , 1 ),
        c(bM,bG) ~ normal( 0 , 0.5 ),
        c(etasq,etasqG) ~ half_normal(1,0.25),
        c(rho,rhoG) ~ half_normal(3,0.25)
    ), data=dat_all , chains=4 , cores=4 , sample=TRUE )
```

(3) Impute G using model

Just M –> G model

$$G \sim \text{MVNormal}(\nu, \mathbf{I}\sigma^2)$$

$$\nu_i = \alpha_G + \beta_{MG} M_i$$

Just phylogeny

$$G \sim \text{MVNormal}(0, \mathbf{K}_G)$$

$$\mathbf{K}_G = \eta_G^2 \exp(-\rho_G d_{i,j})$$

```
# no phylogeny on G but have submodel M -> G
mBMG_OU_G <- ulam(
    alist(
        B ~ multi_normal( mu , K ),
        mu <- a + bM*M + bG*G,
        G ~ normal(nu,sigma),
        nu <- aG + bMG*M,
        M ~ normal(0,1),
        matrix[N_spp,N_spp]:K <- cov_GPL1(Dmat,etasq,rho,0.01),
        c(a,aG) ~ normal( 0 , 1 ),
        c(bM,bG,bMG) ~ normal( 0 , 0.5 ),
        c(etasq) ~ half_normal(1,0.25),
        c(rho) ~ half_normal(3,0.25),
        sigma ~ exponential(1)
    ), data=dat_all , chains=4 , cores=4 , sample=TRUE )

# phylogeny information for G imputation (but no M -> G model)
mBMG_OU2 <- ulam(
    alist(
        B ~ multi_normal( mu , K ),
        mu <- a + bM*M + bG*G,
        M ~ normal(0,1),
        G ~ multi_normal( 'rep_vector(0,N_spp)' ,KG),
        matrix[N_spp,N_spp]:K <- cov_GPL1(Dmat,etasq,rho,0.01),
        matrix[N_spp,N_spp]:KG <- cov_GPL1(Dmat,etasqG,rhoG,0.01),
        a ~ normal( 0 , 1 ),
        c(bM,bG) ~ normal( 0 , 0.5 ),
        c(etasq,etasqG) ~ half_normal(1,0.25),
        c(rho,rhoG) ~ half_normal(3,0.25)
    ), data=dat_all , chains=4 , cores=4 , sample=TRUE )
```

(3) Impute G using model

Just M –> G model

$$G \sim \text{MVNormal}(\nu, \mathbf{I}\sigma^2)$$

$$\nu_i = \alpha_G + \beta_{MG} M_i$$

Just phylogeny

$$G \sim \text{MVNormal}(0, \mathbf{K}_G)$$

$$\mathbf{K}_G = \eta_G^2 \exp(-\rho_G d_{i,j})$$

```
# no phylogeny on G but have submodel M -> G
mBMG_OU_G <- ulam(
    alist(
        B ~ multi_normal( mu , K ),
        mu <- a + bM*M + bG*G,
        G ~ normal(nu,sigma),
        nu <- aG + bMG*M,
        M ~ normal(0,1),
        matrix[N_spp,N_spp]:K <- cov_GPL1(Dmat,etasq,rho,0.01),
        c(a,aG) ~ normal( 0 , 1 ),
        c(bM,bG,bMG) ~ normal( 0 , 0.5 ),
        c(etasq) ~ half_normal(1,0.25),
        c(rho) ~ half_normal(3,0.25),
        sigma ~ exponential(1)
    ), data=dat_all , chains=4 , cores=4 , sample=TRUE )

# phylogeny information for G imputation (but no M -> G model)
mBMG_OU2 <- ulam(
    alist(
        B ~ multi_normal( mu , K ),
        mu <- a + bM*M + bG*G,
        M ~ normal(0,1),
        G ~ multi_normal( 'rep_vector(0,N_spp)' ,KG),
        matrix[N_spp,N_spp]:K <- cov_GPL1(Dmat,etasq,rho,0.01),
        matrix[N_spp,N_spp]:KG <- cov_GPL1(Dmat,etasqG,rhoG,0.01),
        a ~ normal( 0 , 1 ),
        c(bM,bG) ~ normal( 0 , 0.5 ),
        c(etasq,etasqG) ~ half_normal(1,0.25),
        c(rho,rhoG) ~ half_normal(3,0.25)
    ), data=dat_all , chains=4 , cores=4 , sample=TRUE )
```

(3) Impute G using model

Just M –> G model

$$G \sim \text{MVNormal}(\nu, \mathbf{I}\sigma^2)$$

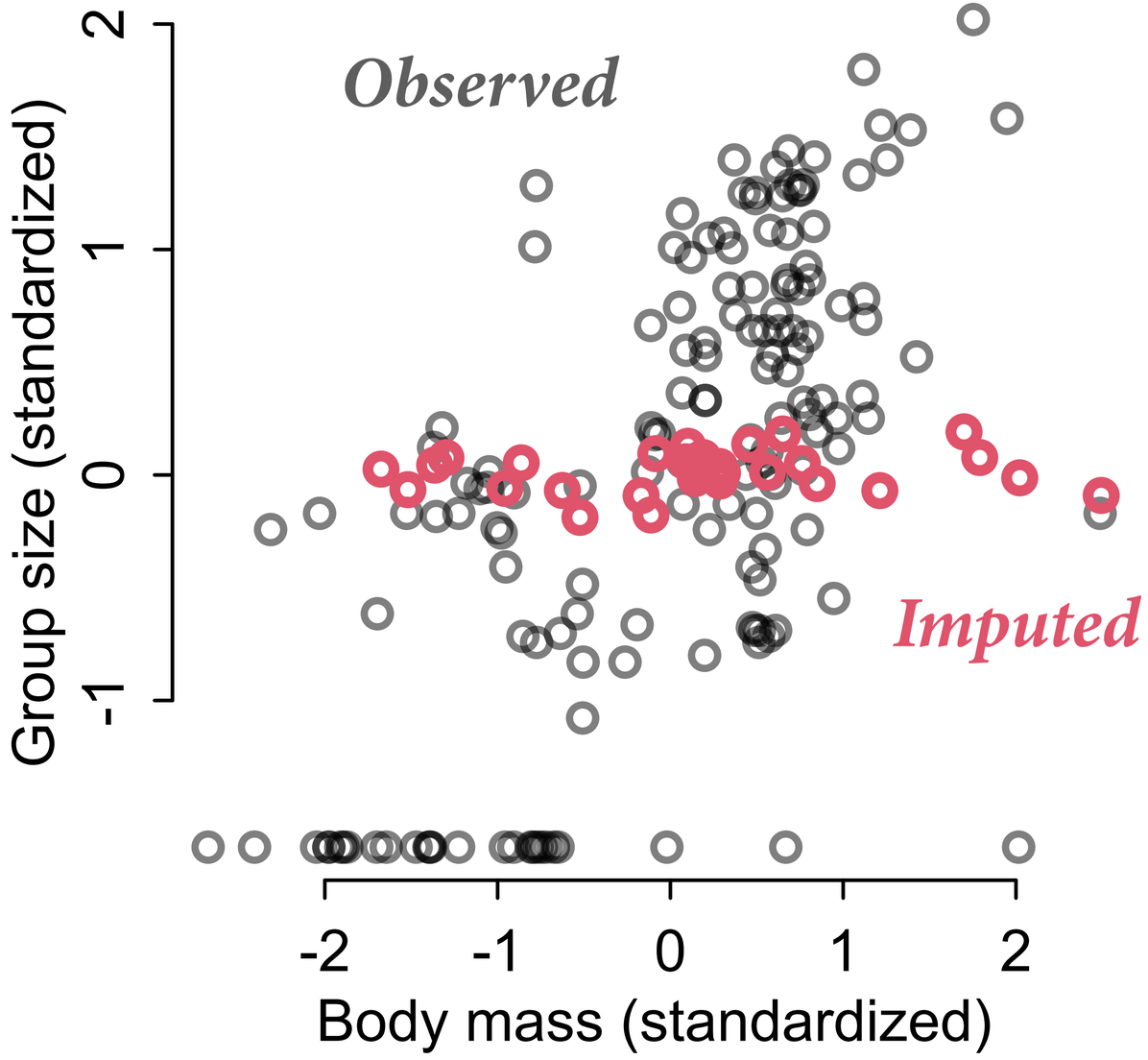
$$\nu_i = \alpha_G + \beta_{MG} M_i$$

Just phylogeny

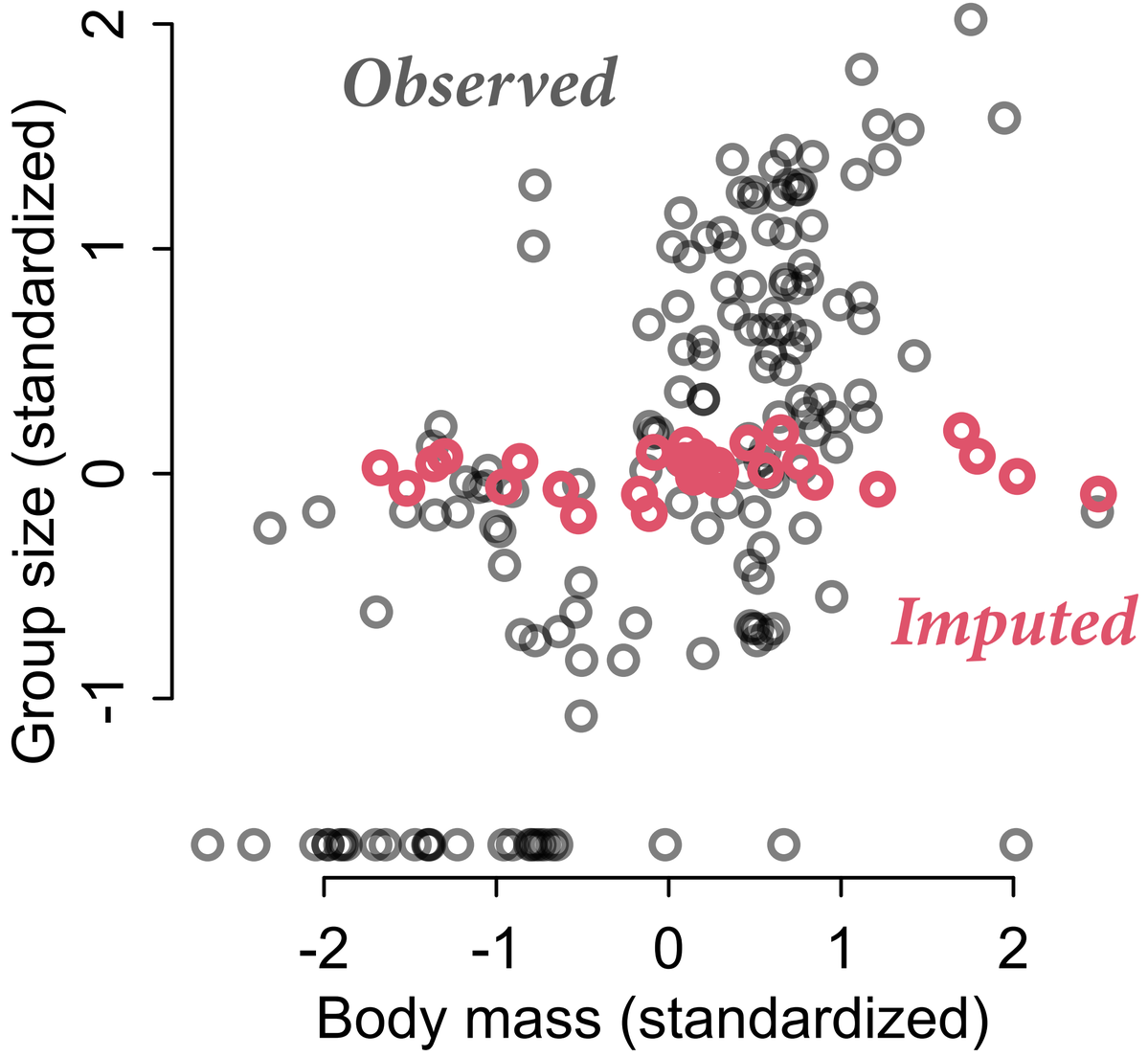
$$G \sim \text{MVNormal}(0, \mathbf{K}_G)$$

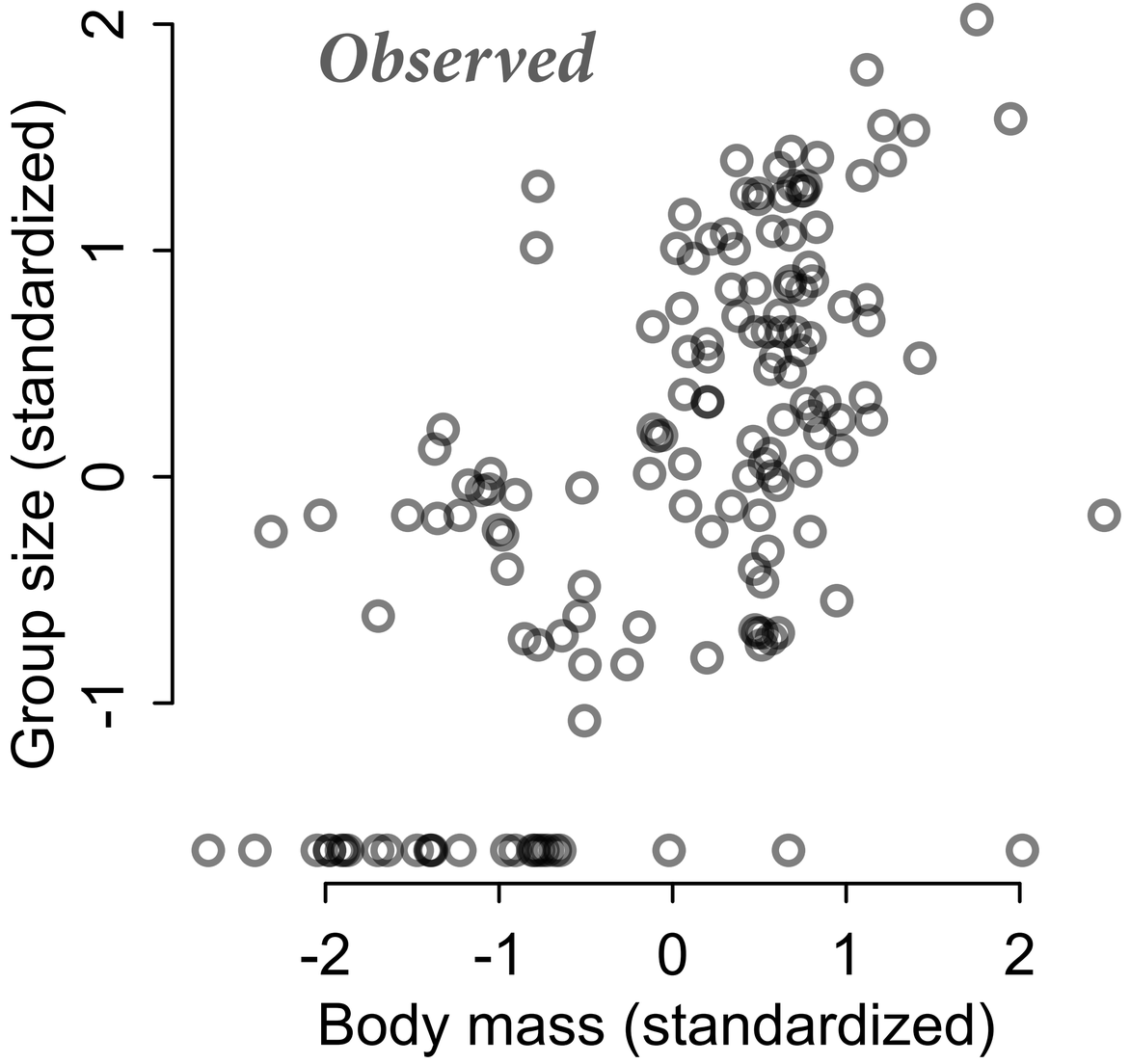
$$\mathbf{K}_G = \eta_G^2 \exp(-\rho_G d_{i,j})$$

(3) Impute *G* using model

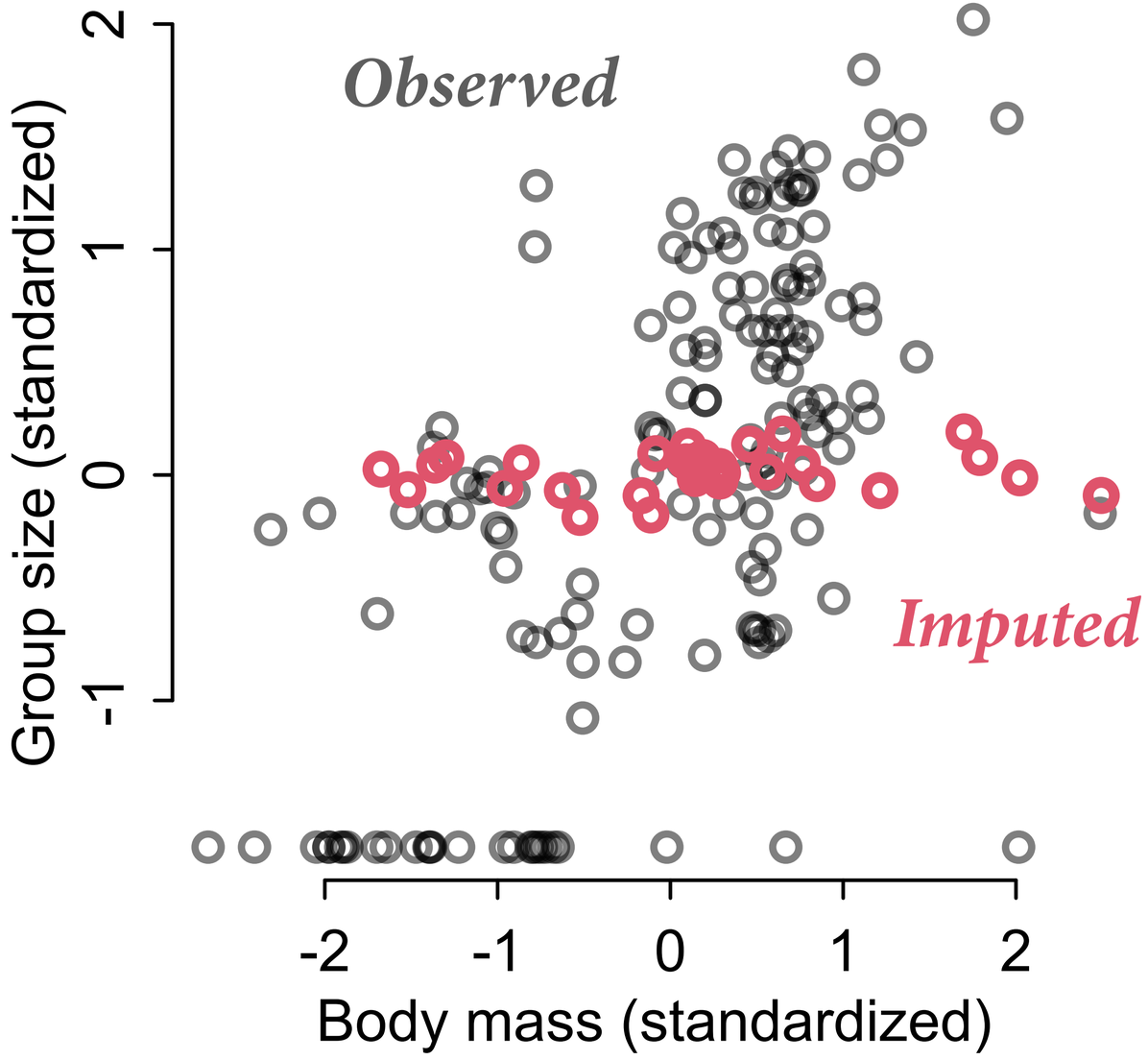


(3) Impute *G* using model

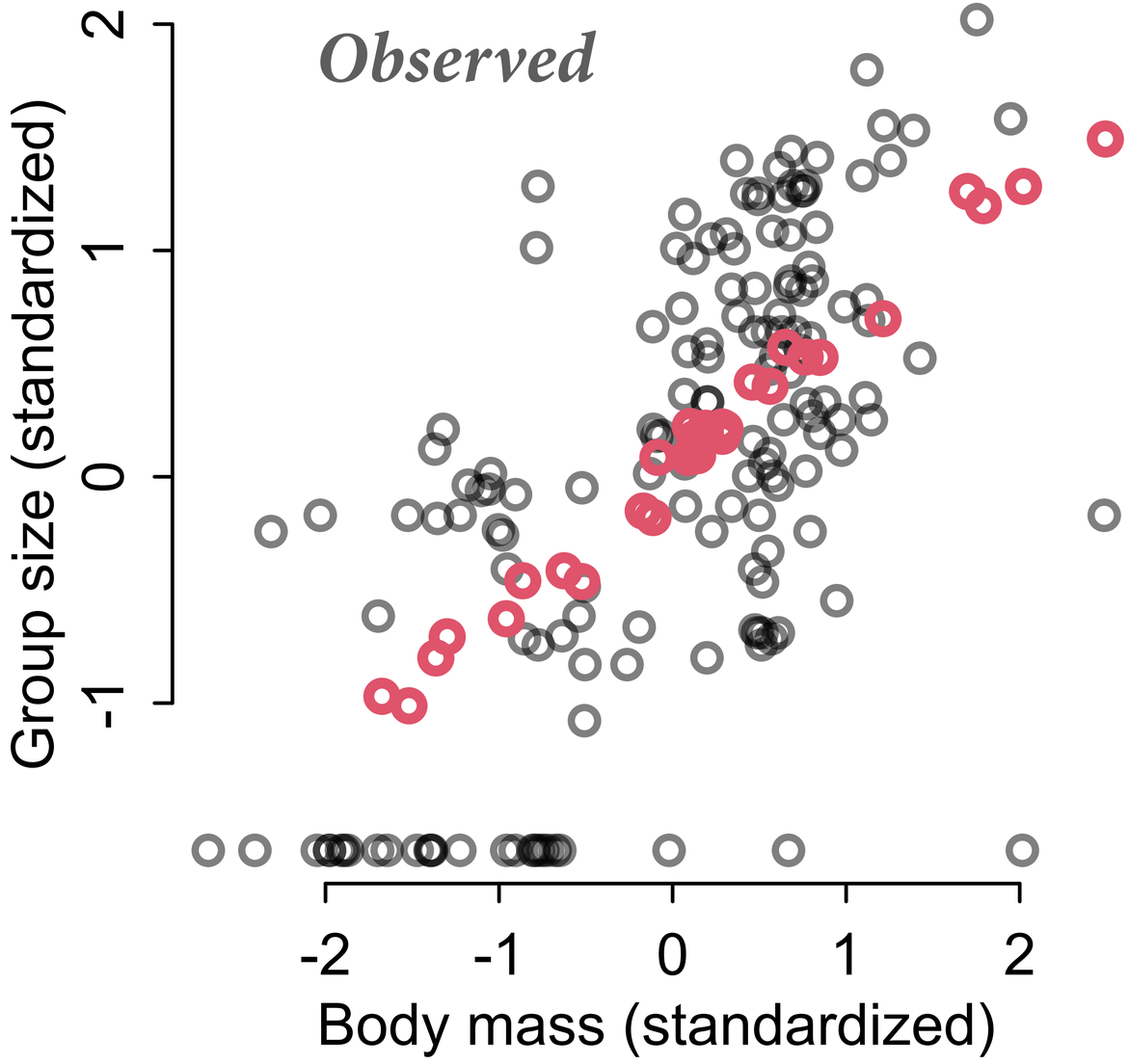




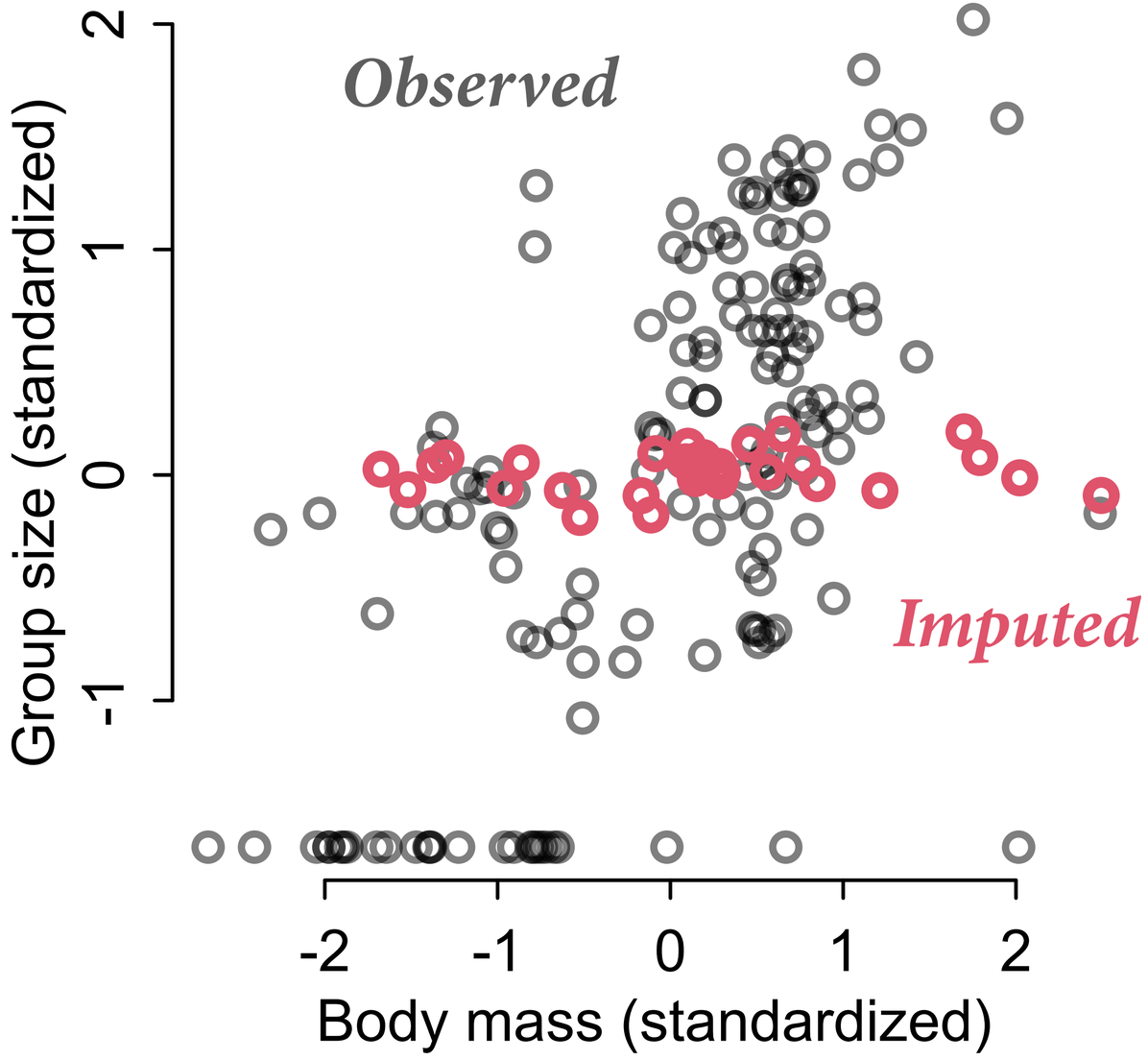
(3) Impute *G* using model

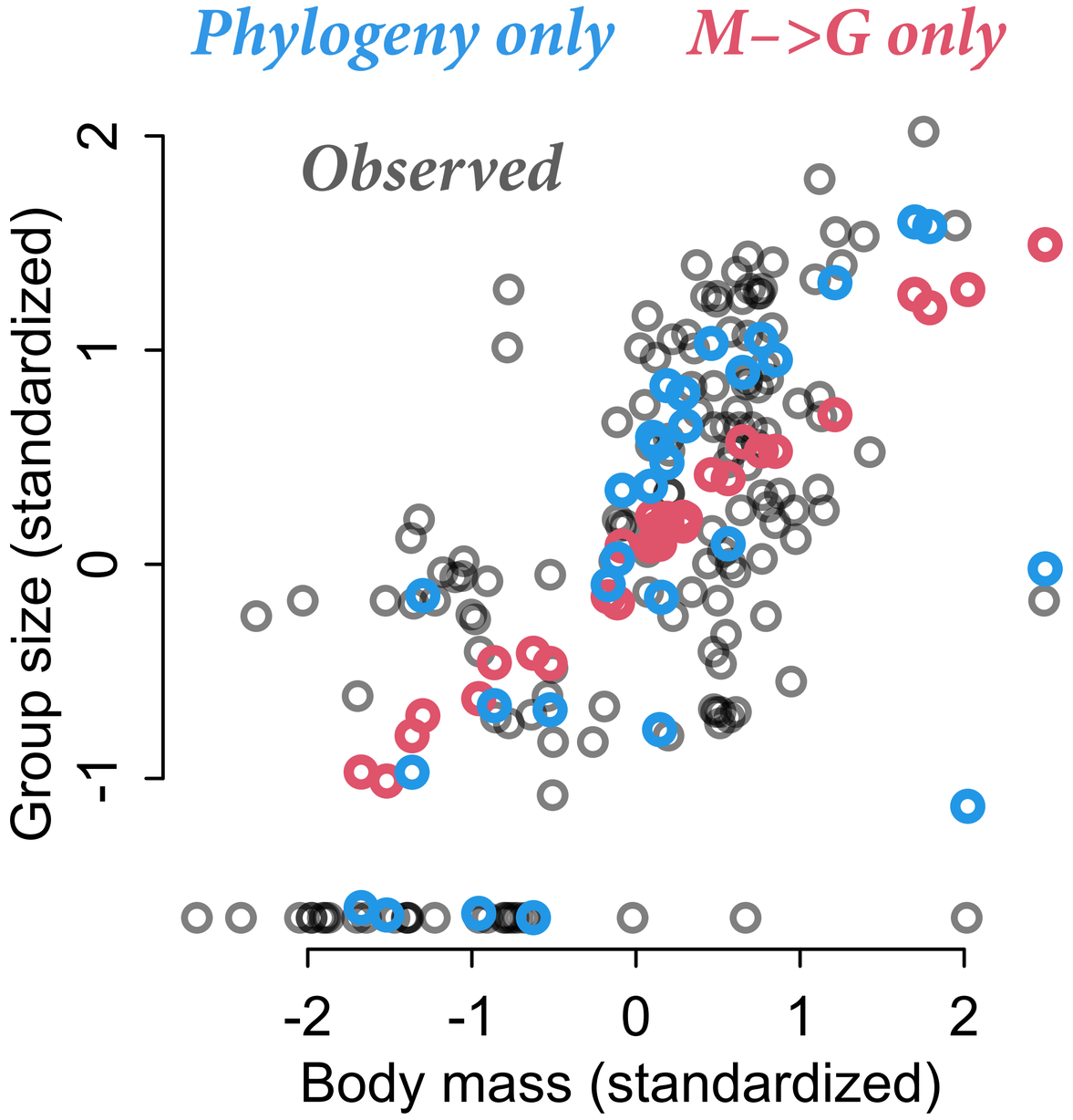


M–>G only

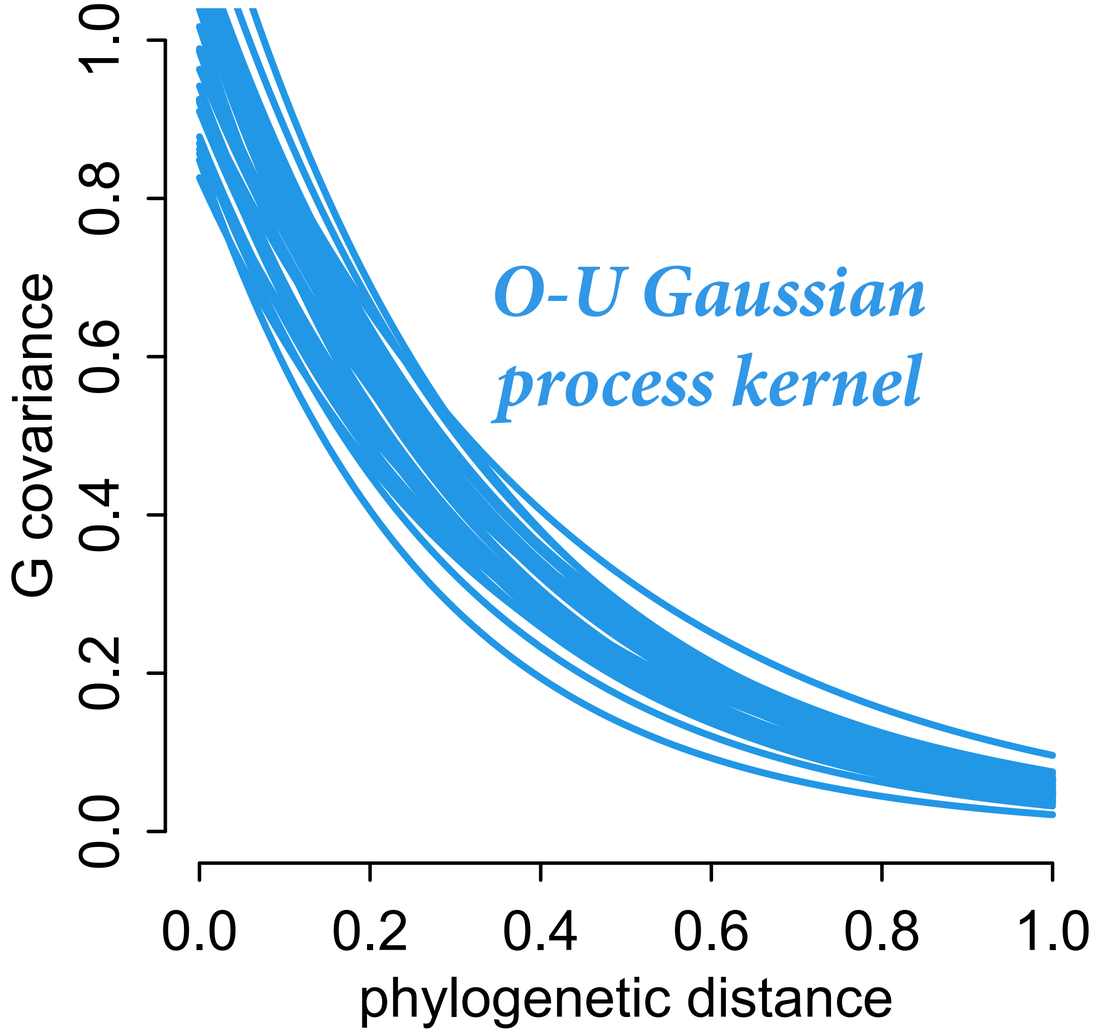


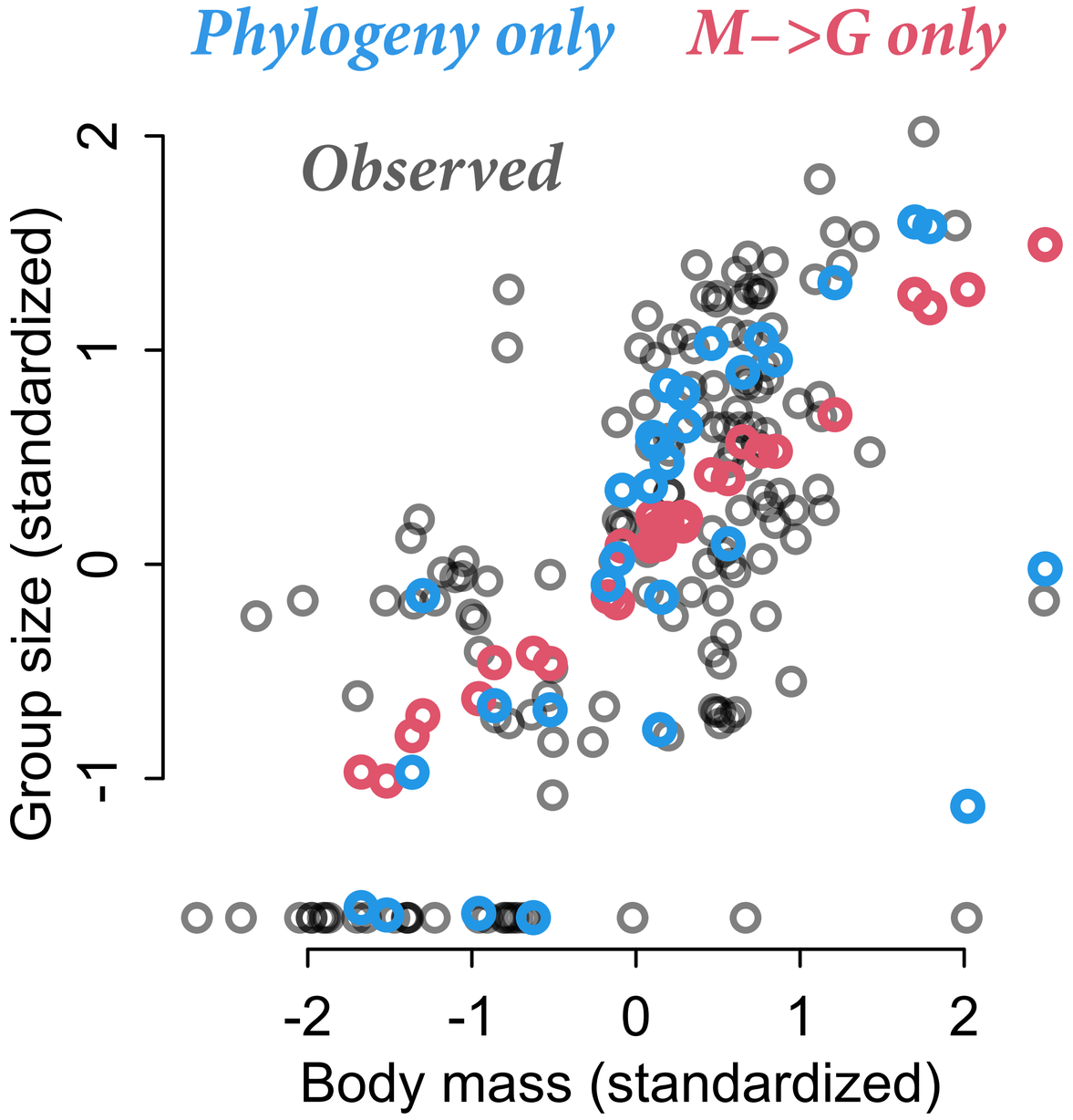
(3) Impute *G* using model



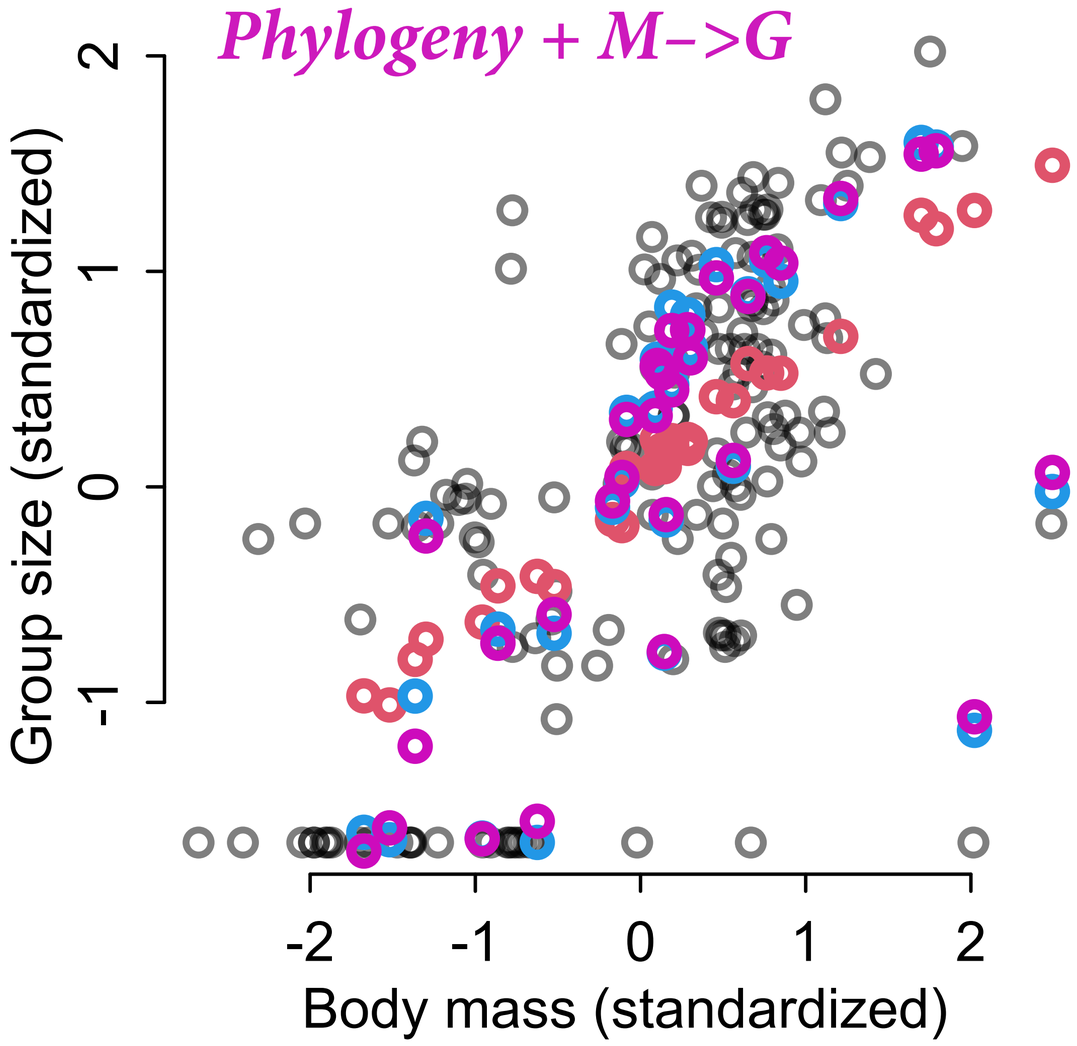


(3) Impute *G* using model

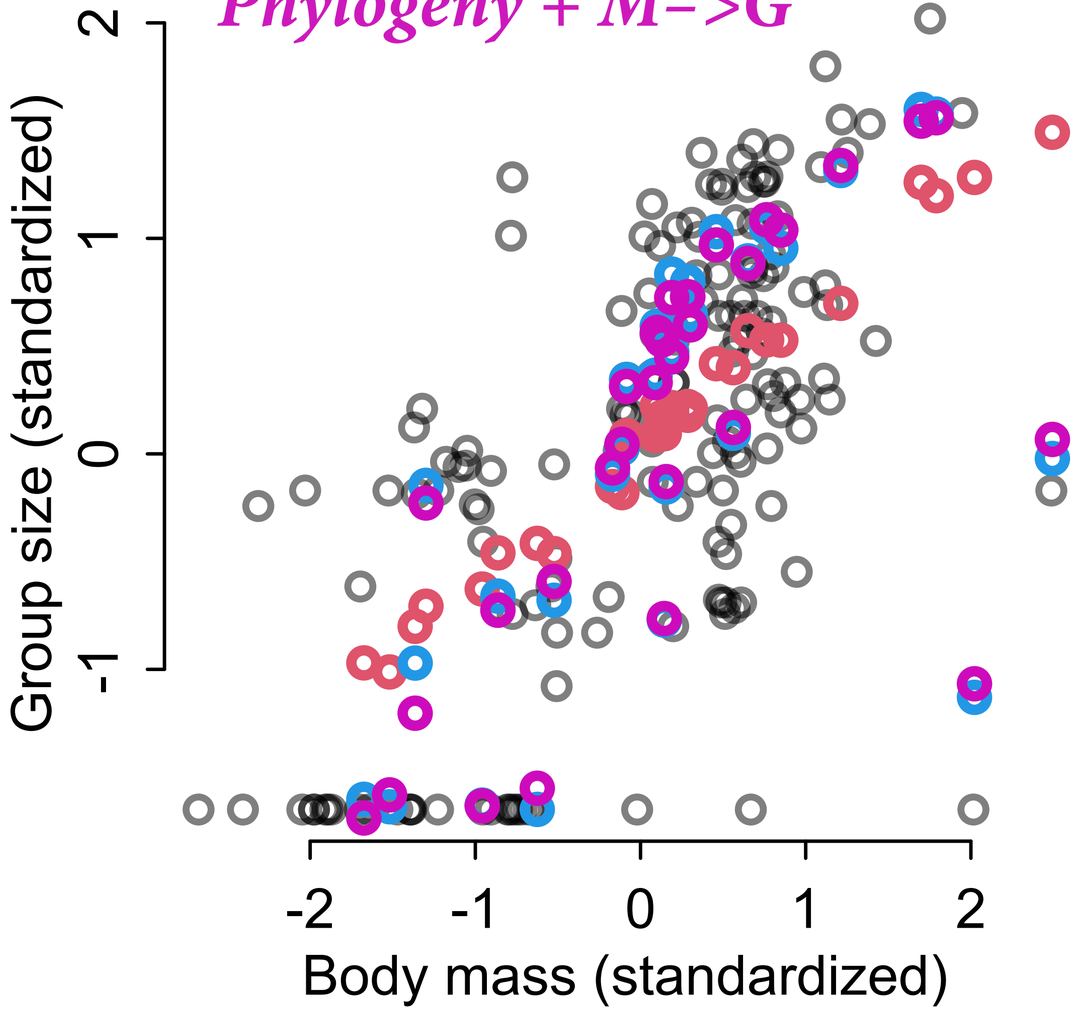


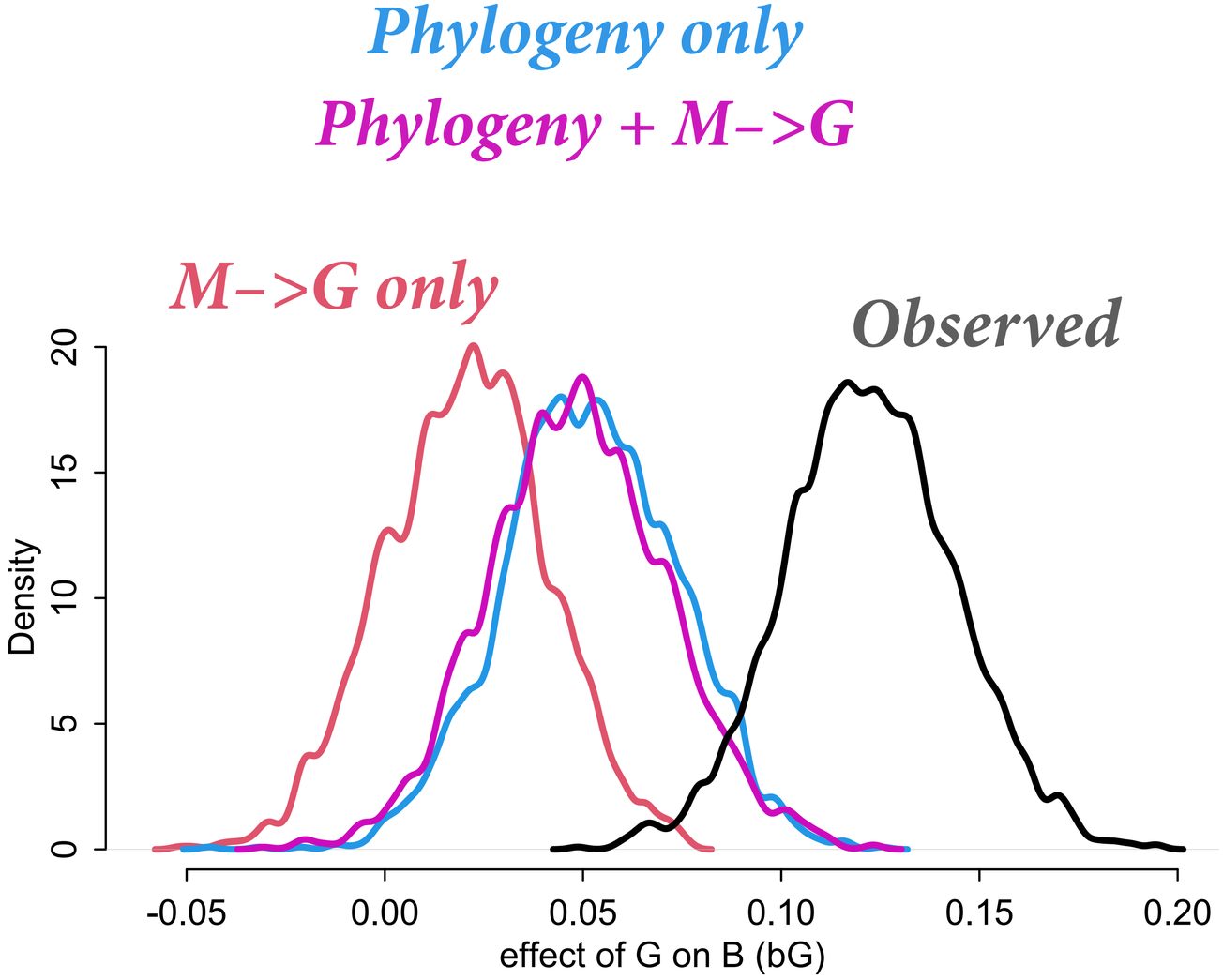


(3) Impute *G* using model



(3) Impute *G* using model





Draw the Missing Owl

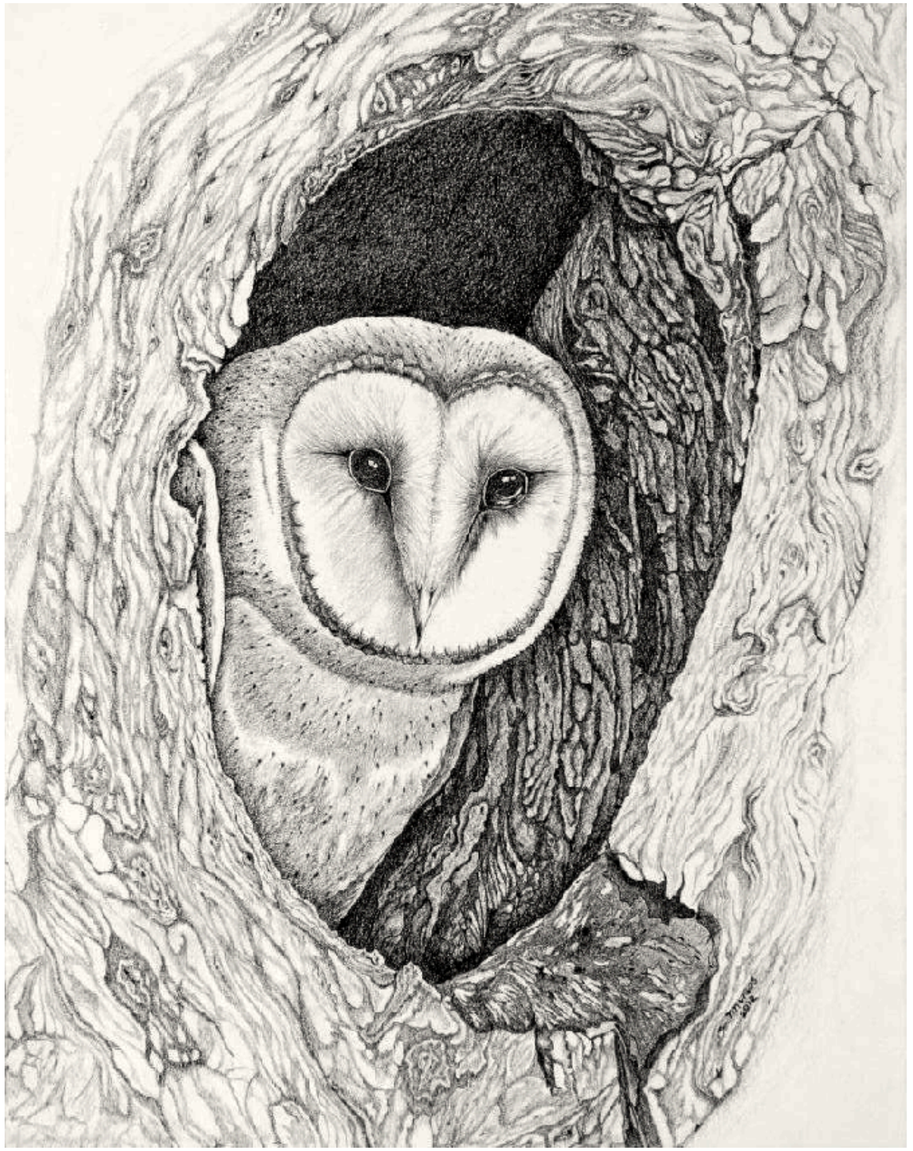
Let's take it slow…

(1) Ignore cases with missing *B* values (for now)

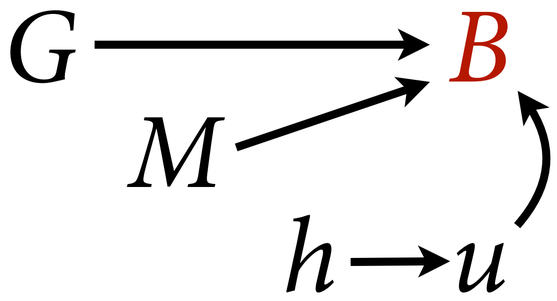
(2) Impute *G* and *M* ignoring models for each

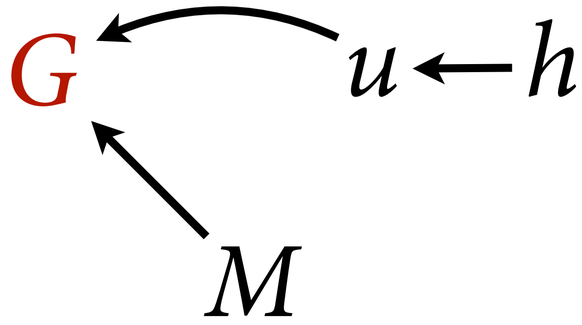
(3) Impute *G* using model

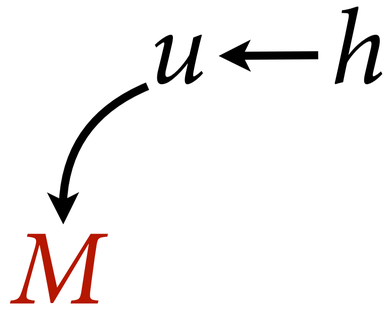
(4) Impute *B*, *G*, *M* using model



(4) Impute B, G, M using model



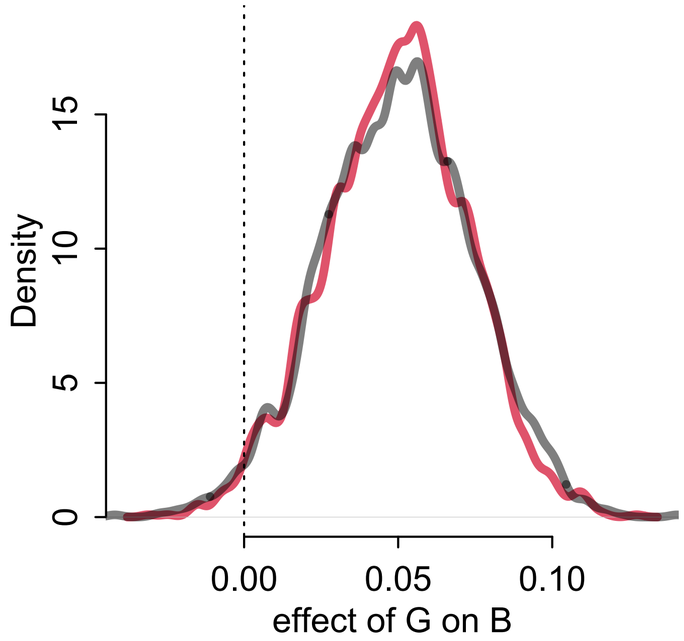


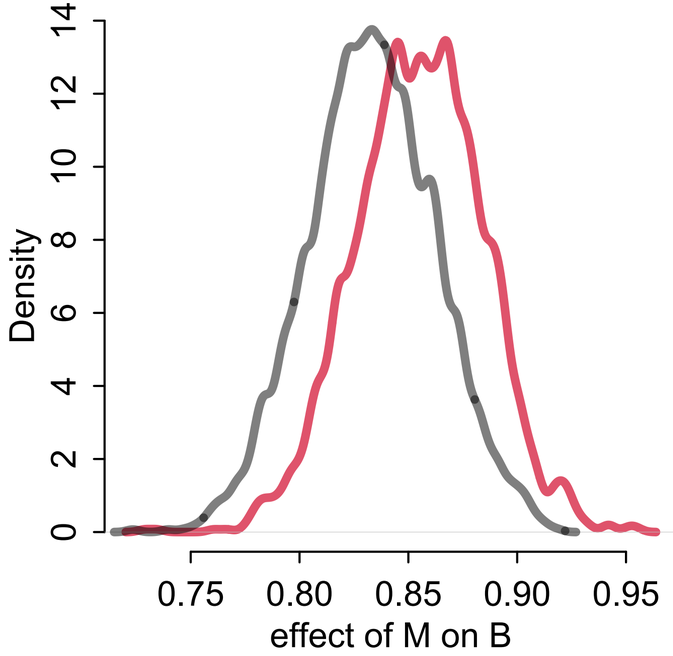


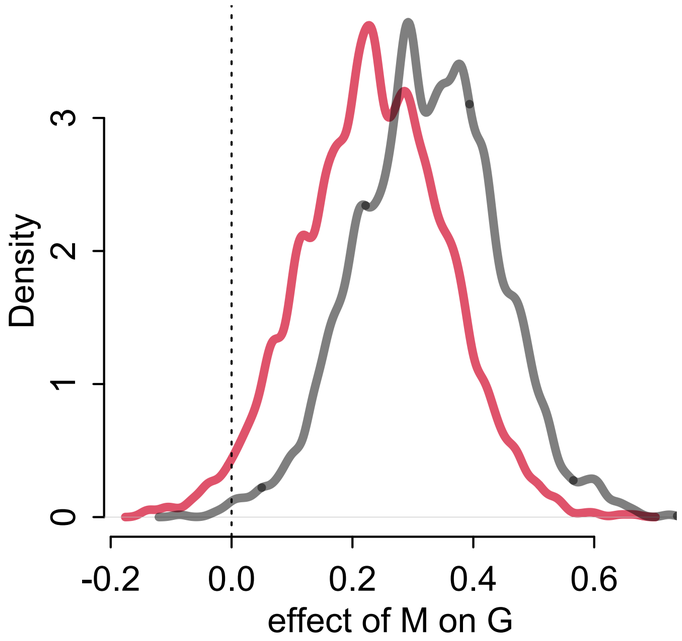
$$B \sim \text{MVNormal}(\mu, \mathbf{K})$$
$$\mu_i = \alpha + \beta_G G_i + \beta_M M_i$$
$$\mathbf{K} = \eta^2 \exp(-\rho d_{i,j})$$
$$\alpha \sim \text{Normal}(0,1)$$
$$\beta_G, \beta_M \sim \text{Normal}(0,0.5)$$
$$\eta^2 \sim \text{HalfNormal}(1,0.25)$$
$$\rho \sim \text{HalfNormal}(3,0.25)$$

$$G \sim \text{MVNormal}(\nu, \mathbf{K}_G)$$
$$\nu_i = \alpha_G + \beta_{MG} M_i$$
$$\mathbf{K}_G = \eta_G^2 \exp(-\rho_G d_{i,j})$$
$$\alpha_G \sim \text{Normal}(0,1)$$
$$\beta_{MG} \sim \text{Normal}(0,0.5)$$
$$\eta_G^2 \sim \text{HalfNormal}(1,0.25)$$
$$\rho_G \sim \text{HalfNormal}(3,0.25)$$

$$M \sim \text{MVNormal}(0, \mathbf{K}_M)$$
$$\mathbf{K}_M = \eta_M^2 \exp(-\rho_M d_{i,j})$$
$$\eta_M^2 \sim \text{HalfNormal}(1,0.25)$$
$$\rho_M \sim \text{HalfNormal}(3,0.25)$$

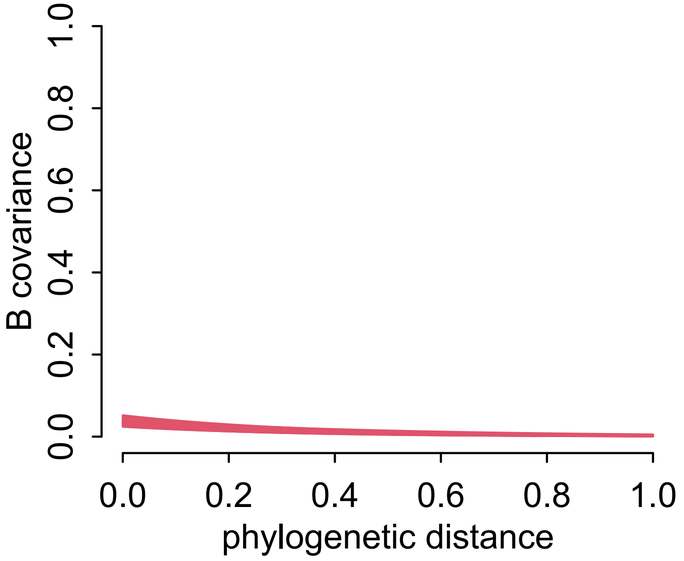


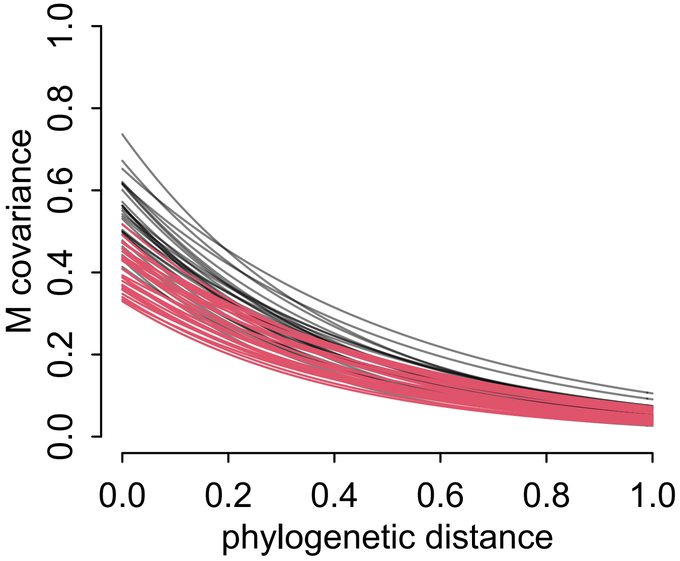


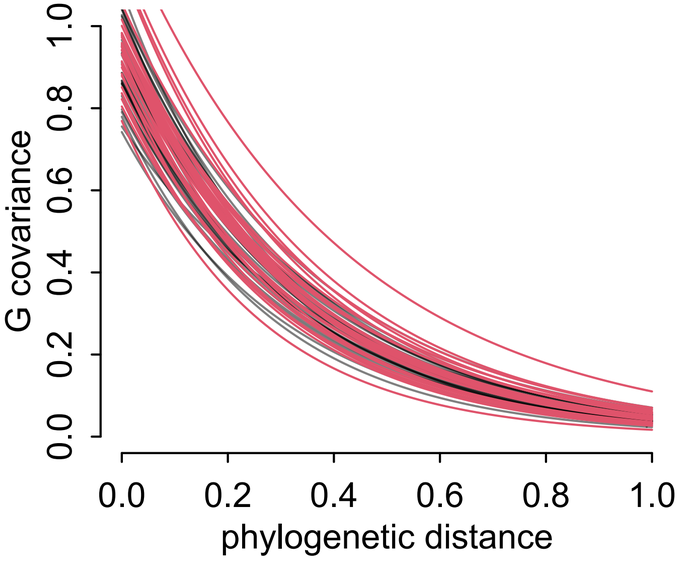


complete cases

full luxury imputation







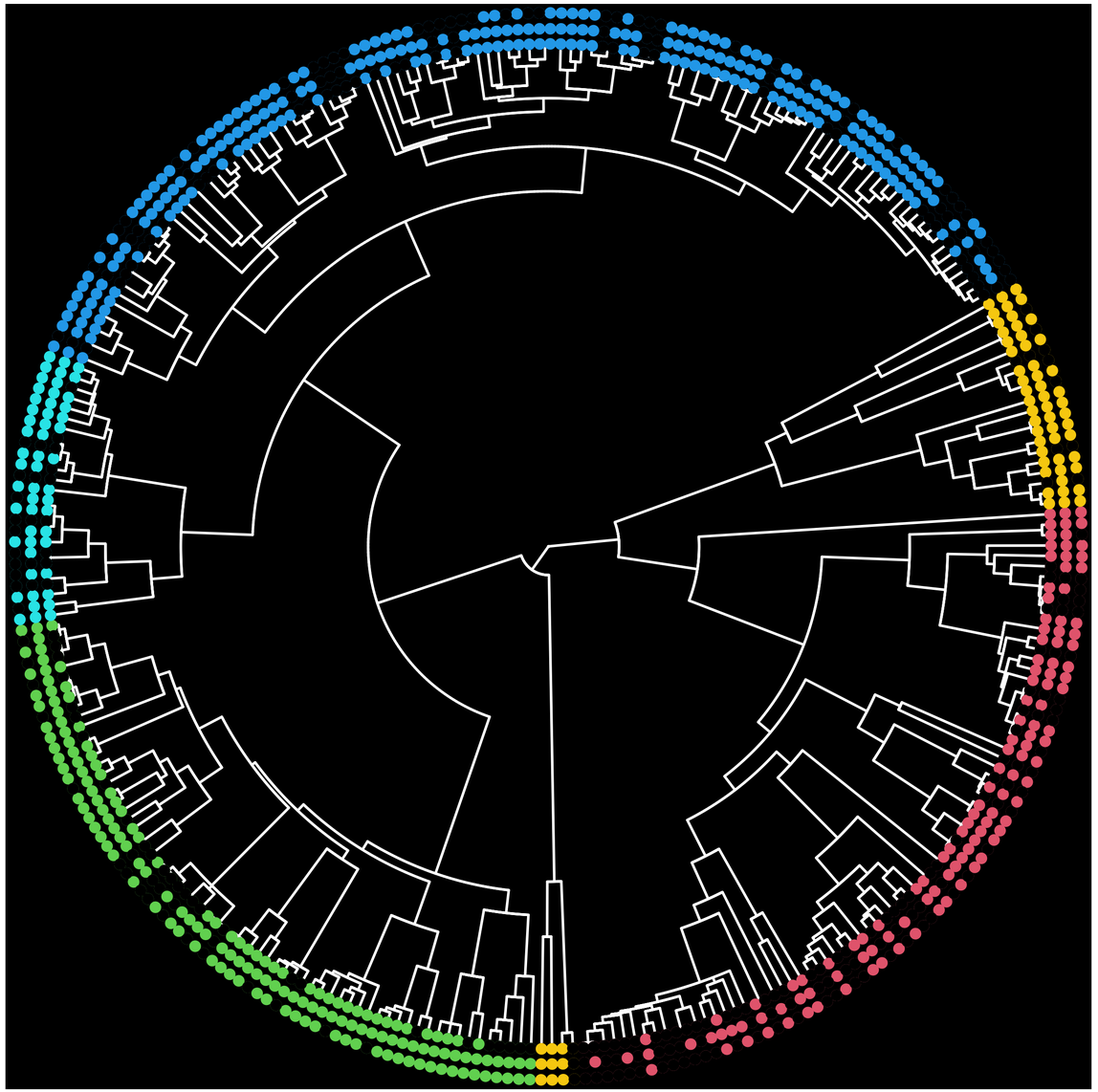
Imputing Primates

Key idea: Missing values already have probability distributions

Think like a graph, not like a regression

Imputation without relationships among predictors risky

Even if doesn't change result, it's our duty



Course Schedule

Week 1	Bayesian inference	Chapters 1, 2, 3
Week 2	Linear models & Causal Inference	Chapter 4
Week 3	Causes, Confounds & Colliders	Chapters 5 & 6
Week 4	Overfitting / MCMC	Chapters 7, 8, 9
Week 5	Generalized Linear Models	Chapters 10, 11
Week 6	Ordered categories & Multilevel models	Chapters 12 & 13
Week 7	More Multilevel models	Chapters 13 & 14
Week 8	Social Networks & Gaussian Processes	Chapter 14
Week 9	Measurement & Missingness	Chapter 15
Week 10	Generalized Linear Madness	Chapter 16

https://github.com/rmcelreath/stat_rethinking_2023