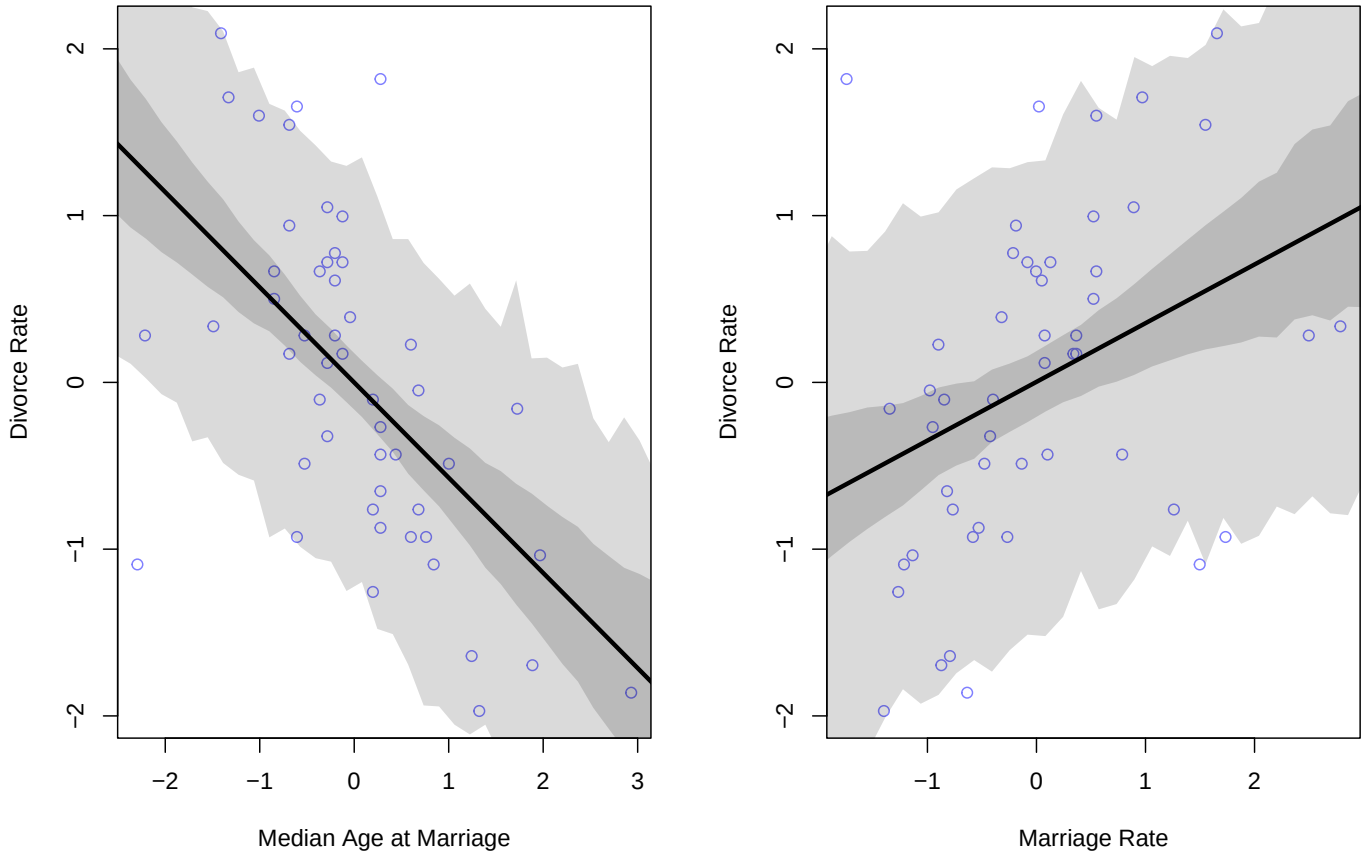
Causal Models and Multivariate Regression

Divorce rate, Milk and Apes

Nikodem Lewandowski

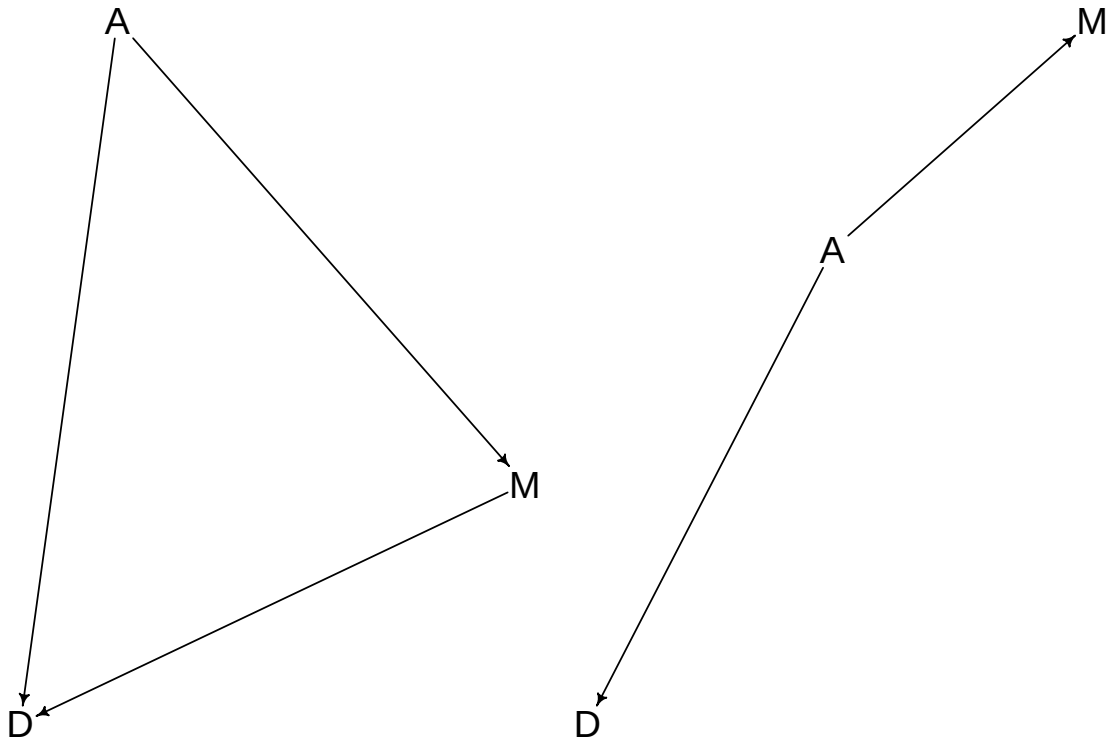
Post. Pred. Checks of Our Models



- Points are true values
- Lines represent posterior's mean mu value
- Dark Grey areas are 0.89 HPDI of posterior's mu
- Light Grey areas are 0.89 HPDI of simulated predictions

University of Gdańsk

DAGs



- The second model implies that:

$$I(D, M)|A$$

Multiple Regression

$$
\begin{aligned}
\text{D} &\sim \text{Normal}(\mu, \sigma) \\
\mu &\sim a + bA * A + bM * M \\
a &\sim \text{Normal}(0, 0.2) \\
bA &\sim \text{Normal}(0, 0.5) \\
bM &\sim \text{Normal}(0, 0.5) \\
\sigma &\sim \text{Exp}(1)
\end{aligned}
$$

- The linear equation can be interpreted as:

A State's divorce rate can be a function of its marriage rate **or** its median age at marriage

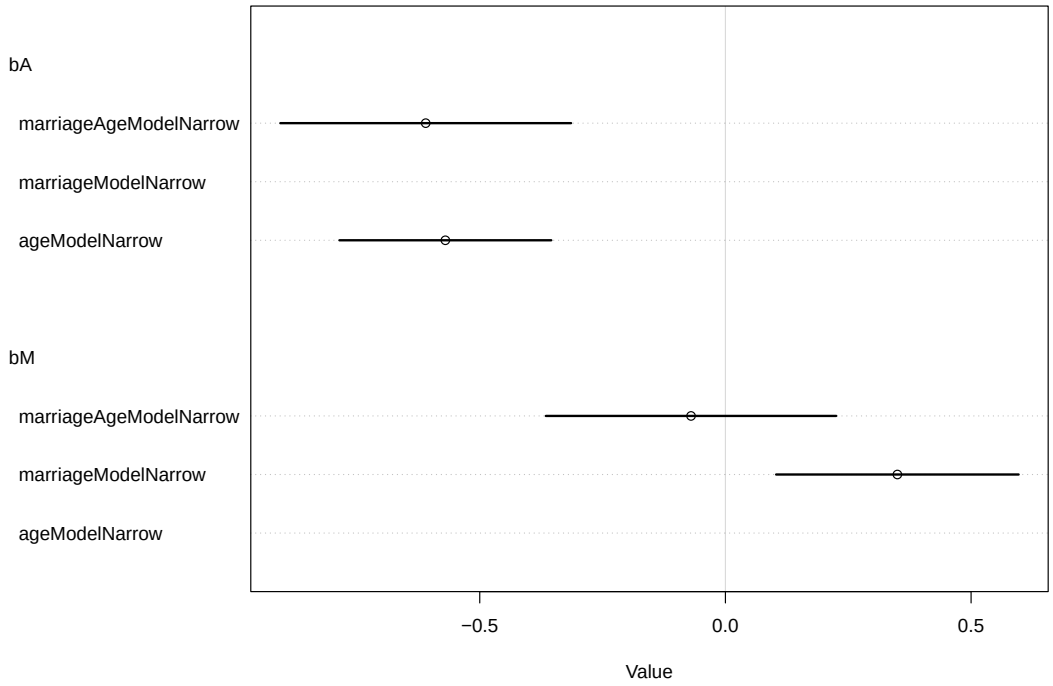
```
marriageAgeModelNarrow <- quap(
  alist(
    D ~ dnorm(mu, sigma ) ,
    mu <- a + bA * A + bM * M,
    a ~ dnorm(0, .2),
    bA ~ dnorm( 0, .5),
    bM ~ dnorm( 0, .5),
    sigma ~ dexp( 1 )
  ), data = d
)
```

Small Summary and Comparison

```
round(precis(marriageAgeModelNarrow),3)
```

```
        mean    sd   5.5%  94.5%
a       0.000 0.109 -0.174  0.174
bA     -0.613 0.152 -0.855 -0.371
bM     -0.065 0.151 -0.307  0.177
sigma  0.788 0.079  0.663  0.914
```

```
plot( coeftab(ageModelNarrow,
              marriageModelNarrow,marriageAgeModelNarrow), par=c("bA","bM") )
```



University of Gdańsk

Evaluating Multivariate Models

- It is not obvious how to evaluate a multivariate model that has e.g. two variables assigned to two different slope parameters (our case). Different approaches are possible:
  - Predictor residual plots, the outcome against residual predictor values
  - Posterior prediction plots, model-based predictions against raw data
  - Counterfactual plots, implied predictions for imaginary experiments

University of Gdańsk

1. Residuals

- A predictor variable residual is the average prediction error when we use all of the other predictor variables to model a predictor of interest.
- We have two predictors: (1) marriage rate (M) and (2) median age at marriage (A). To compute predictor residuals for either, we just use the other predictor to model it:

```
marriageAndAge <- quap(
alist(
M ~ dnorm( mu , sigma ) ,
mu <- a + bAM * A ,
a ~ dnorm( 0 , 0.2 ) ,
bAM ~ dnorm( 0 , 0.5 ) ,
sigma ~ dexp( 1 )
) , data = d )

AgeAndMarriage <- quap(
alist(
A ~ dnorm( mu , sigma ) ,
mu <- a + bM * M ,
a ~ dnorm( 0 , 0.2 ) ,
bM ~ dnorm( 0 , 0.5 ) ,
sigma ~ dexp( 1 )
) , data = d )
```

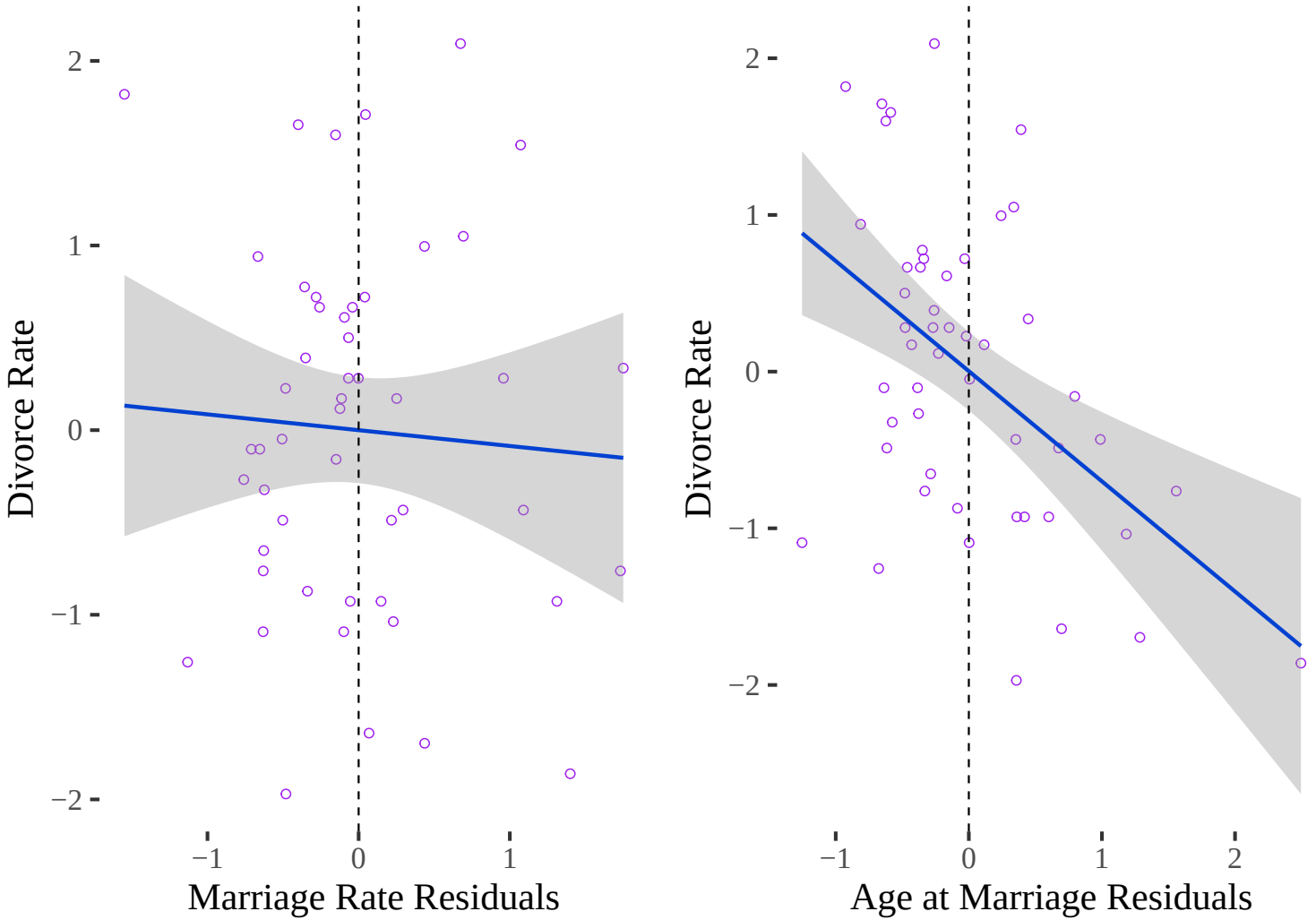
1. Residuals

- And then we compute the residuals by subtracting the observed marriage rate in each State from the predicted rate, based upon the model above
- We will plot those residuals with the divorce rate

```
muM <- link(marriageAndAge)
mu_meanM <- apply( muM , 2 , mean )
mu_residM <- d$M - mu_meanM

muA <- link(AgeAndMarriage)
mu_meanA <- apply( muA , 2 , mean )
mu_residA <- d$A - mu_meanA
```

University of Gdańsk

1. Residuals



University of Gdańsk

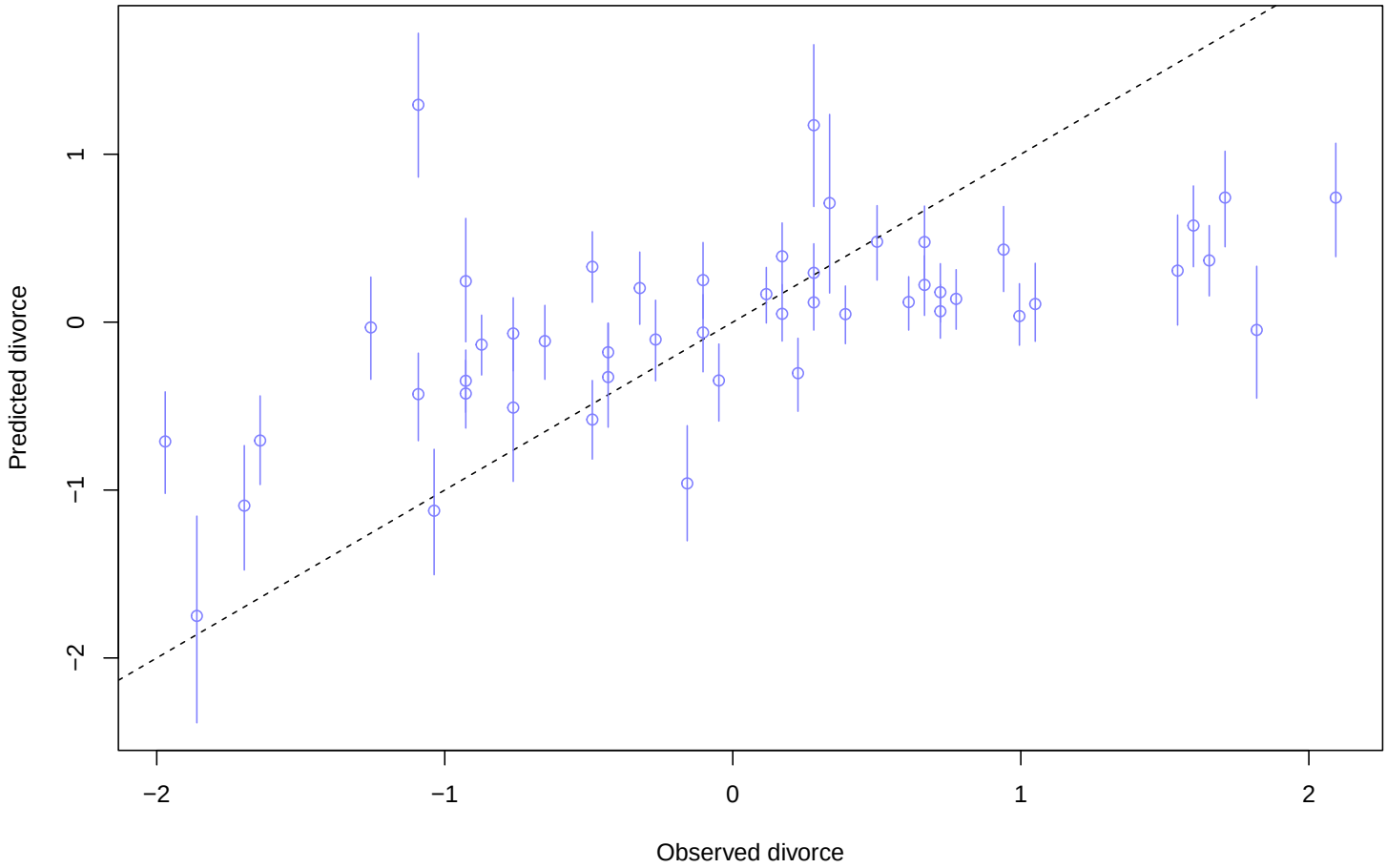
2. Posterior prediction

- Simmilarly to the posterior checks that we did before:
  - we simulate states from our model to compare them with the observed values

```
mu <- link( marriageAgeModelNarrow )
# summarize samples across cases
mu_mean <- apply( mu , 2 , mean )
mu_PI <- apply( mu , 2 , PI )
# simulate observations
# again no new data, so uses original data
D_sim <- sim( marriageAgeModelNarrow , n=1e4 )
D_PI <- apply( D_sim , 2 , PI )
```

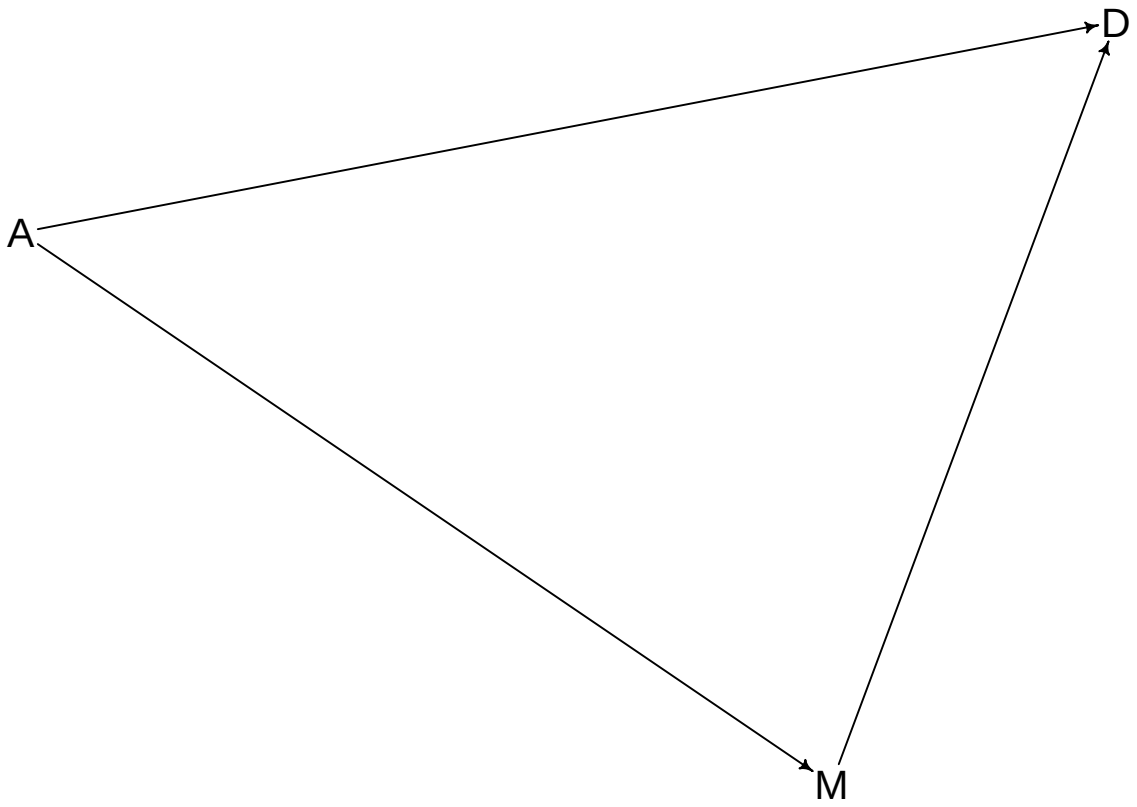
University of Gdańsk

2. Posterior prediction



University of Gdańsk

3. Counterfactual plots



- We will test different possible scenarios of values that A and M can take
- No only A or M on D, but also A on M

University of Gdańsk

3. Counterfactual plots

```
DAG_model <- quap(
alist(
## A -> D <- M
D ~ dnorm( mu , sigma ) ,
mu <- a + bM*M + bA*A ,
a ~ dnorm( 0 , 0.2 ) ,
bM ~ dnorm( 0 , 0.5 ) ,
bA ~ dnorm( 0 , 0.5 ) ,
sigma ~ dexp( 1 ),
## A -> M
M ~ dnorm( mu_M , sigma_M ),
mu_M <- aM + bAM*A,
aM ~ dnorm( 0 , 0.2 ),
bAM ~ dnorm( 0 , 0.5 ),
sigma_M ~ dexp( 1 )
) , data = d )
```

University of Gdańsk

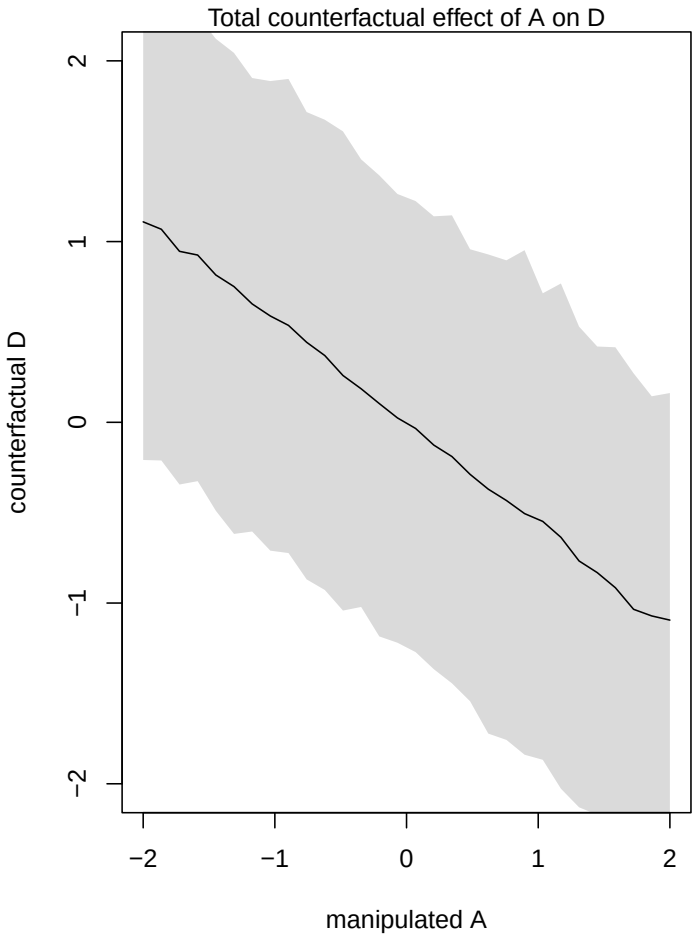
3. Counterfactual plots

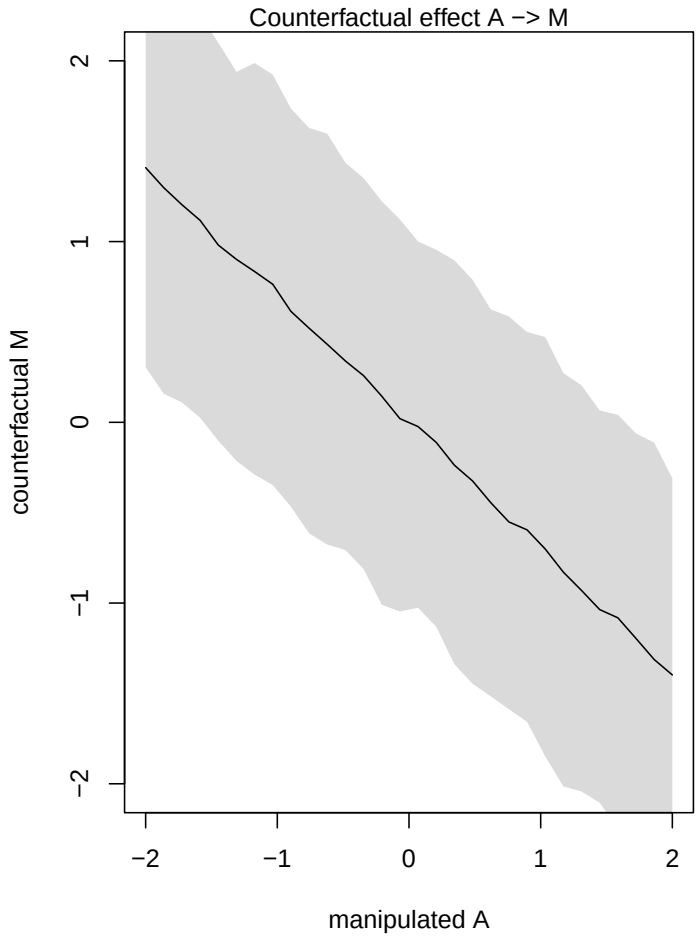
- We create artificial sequence of values to test what will happen if we maniuplate A

```
A_seq <- seq( from=-2 , to=2 , length.out=30 )
sim_dat <- data.frame( A=A_seq )

s <- sim( DAG_model , data=sim_dat , vars=c("M","D") )
```

University of Gdańsk

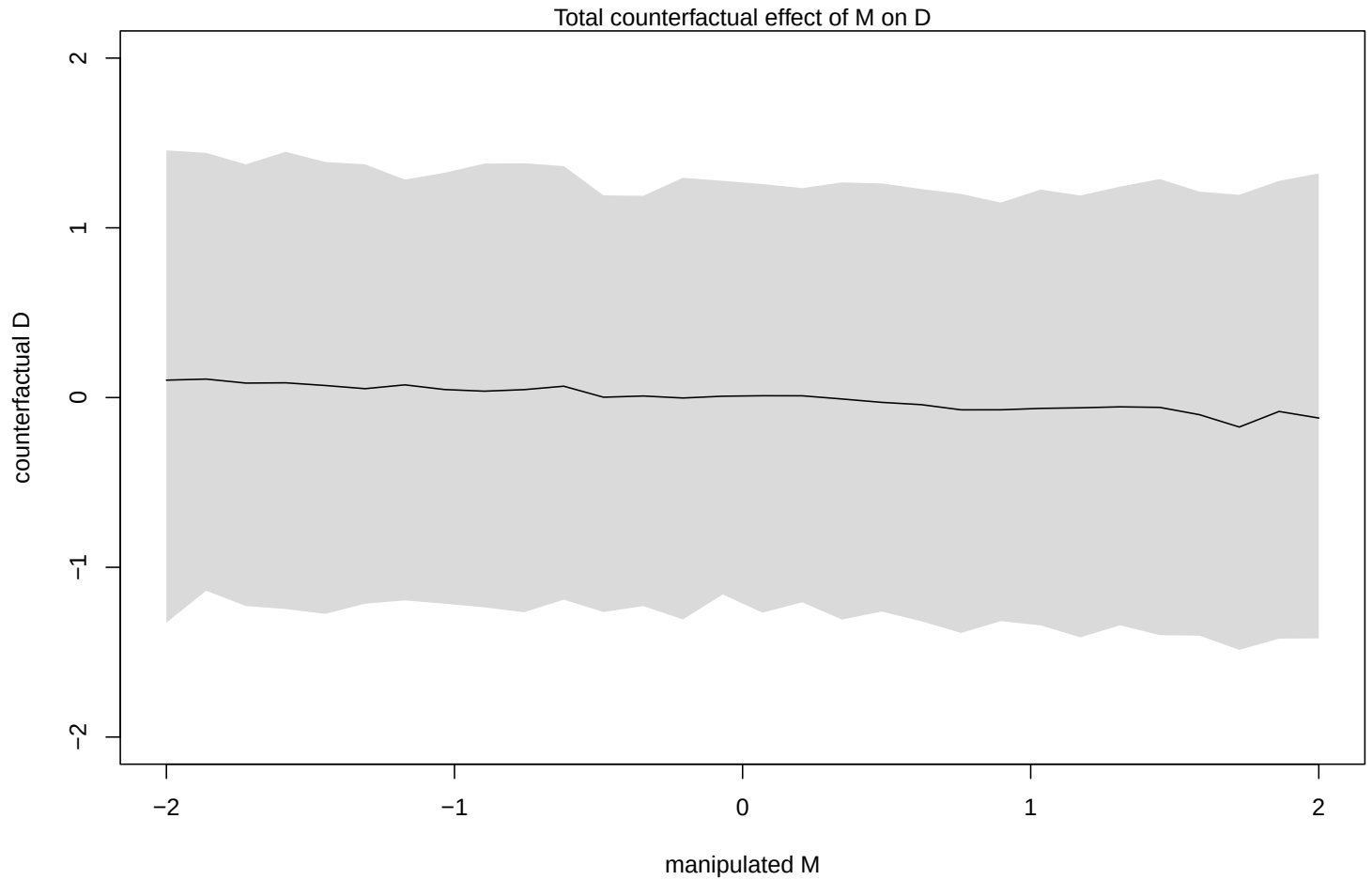
3. Counterfactual plots





University of Gdańsk

3. Counterfactual plots: the third wheel



University of Gdańsk

Conclusions

- Manipulating marriage rate has no significant effect on divorce rate
- We were able to knock out spurious association with multiple predictor model
- Now we will see an another use of this general strategy

University of Gdańsk

Masked relationship

- A new dataset!

```
data(milk)
d <- milk
head(d, n=3)
```

```
          clade        species kcal.per.g perc.fat perc.protein perc.lactose
1 Strepsirrhine Eulemur fulvus       0.49    16.60        15.42        67.98
2 Strepsirrhine       E macaco       0.51    19.27        16.91        63.82
3 Strepsirrhine       E mongoz       0.46    14.11        16.85        69.04
  mass neocortex.perc
1 1.95          55.16
2 2.09             NA
3 2.51             NA
```

kcal.per.g : Kilocalories of energy per gram of milk.
mass : Average female body mass, in kilograms.

neocortex.perc : The percent of total brain mass that is neocortex mass.

University of Gdańsk

Masked relationship

```
d$K <- scale( d$kcal.per.g )
d$N <- scale( d$neocortex.perc )
d$M <- scale( log(d$mass) )


# excluding NAs
dcc <- d[ complete.cases(d$K,d$N,d$M) , ]
```

University of Gdańsk

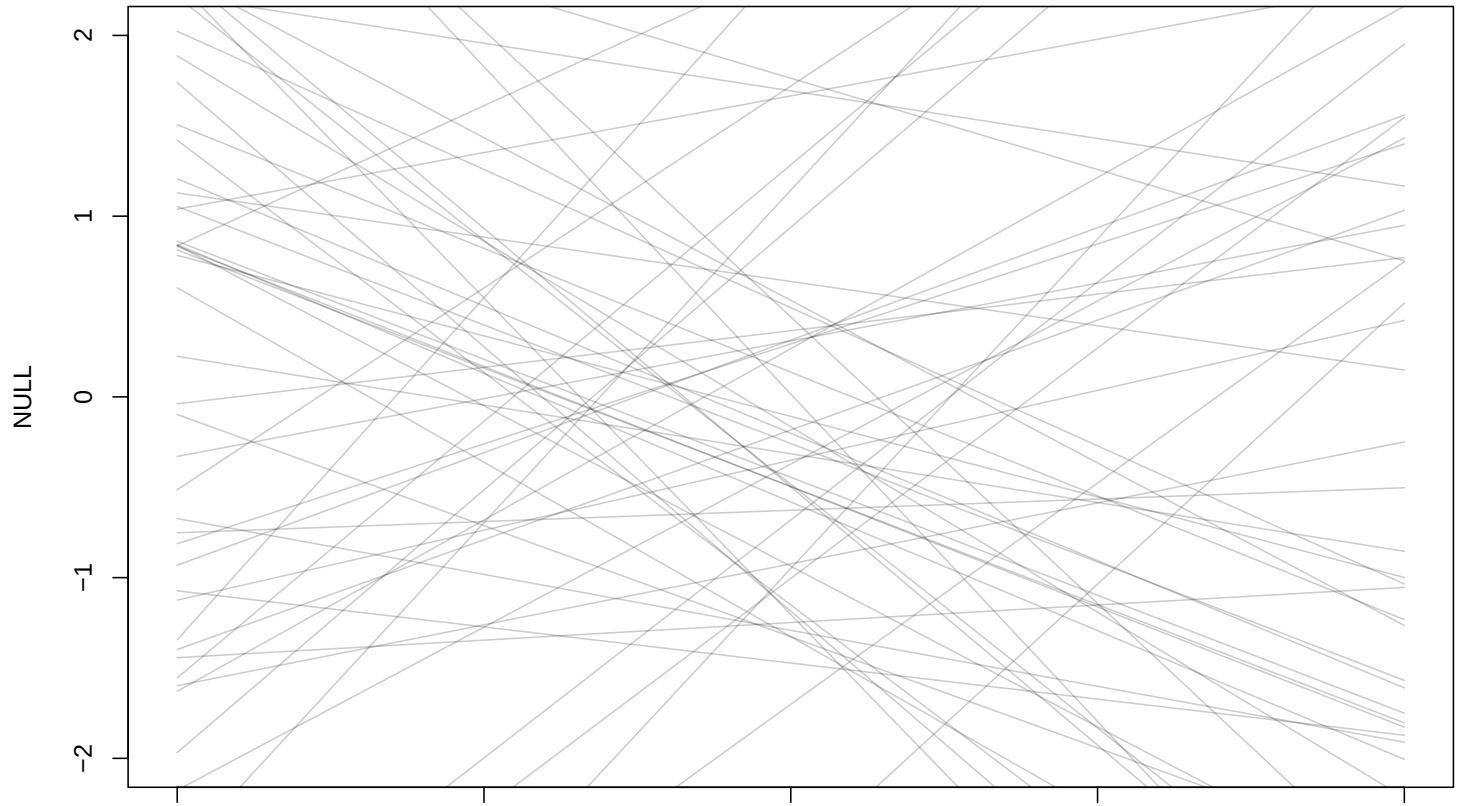
First Model: Milk Richness and Neocrtex perc.

```
milk_try2 <- quap(
  alist(
    K ~ dnorm( mu, sigma),
    mu <- a + bN * N,
    a ~ dnorm(0, 1),
    bN ~ dnorm( 0, 1),
    sigma ~ dexp(1)
  ), data = dcc
)
```

University of Gdańsk

Prior Check

```
prior <- extract.prior(milk_try2)
xseq <- seq(-2,2,length.out = 30)
mu <- link(milk_try2, post = prior, data = list(N = xseq))

plot( NULL, xlim = c(-2,2), ylim = c(-2,2))
for (i in 1:50 ) lines (xseq, mu[i,], col = col.alpha("black", .2))
```



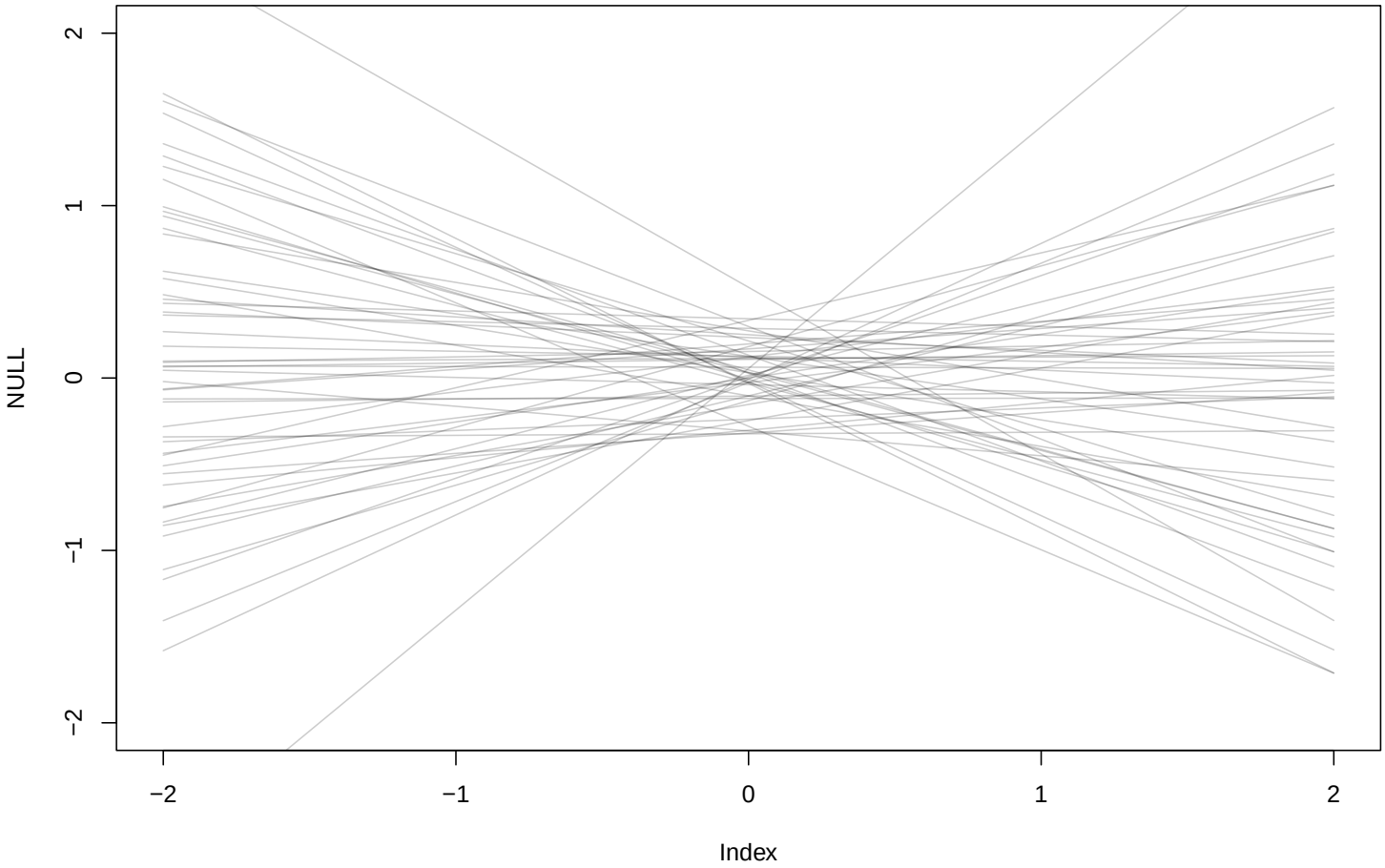
University of Gdańsk

Revising Priors

```
milk_n <- quap(
  alist(
    K ~ dnorm( mu, sigma),
    mu <- a + bN * N,
    a ~ dnorm(0, .2),
    bN ~ dnorm( 0, .5),
    sigma ~ dexp(1)
  ), data = dcc
)
```

University of Gdańsk

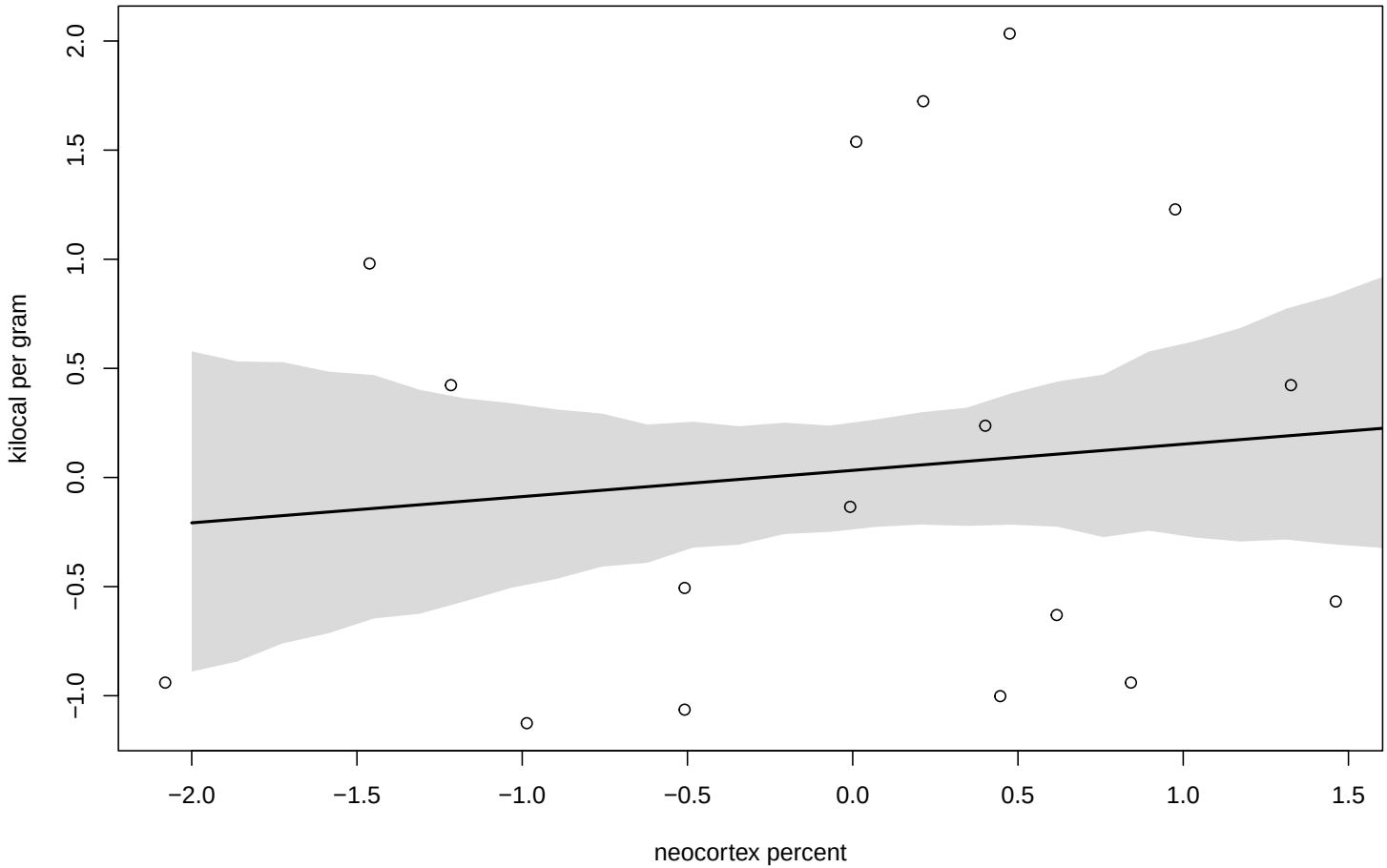
Next Prior Check



University of Gdańsk

Posterior for neocortex percentage

	mean	sd	5.5%	94.5%
a	0.03993647	0.1544903	-0.2069689	0.2868418
bN	0.13322436	0.2237456	-0.2243643	0.4908130
sigma	0.99981242	0.1647048	0.7365823	1.2630426



University of Gdańsk

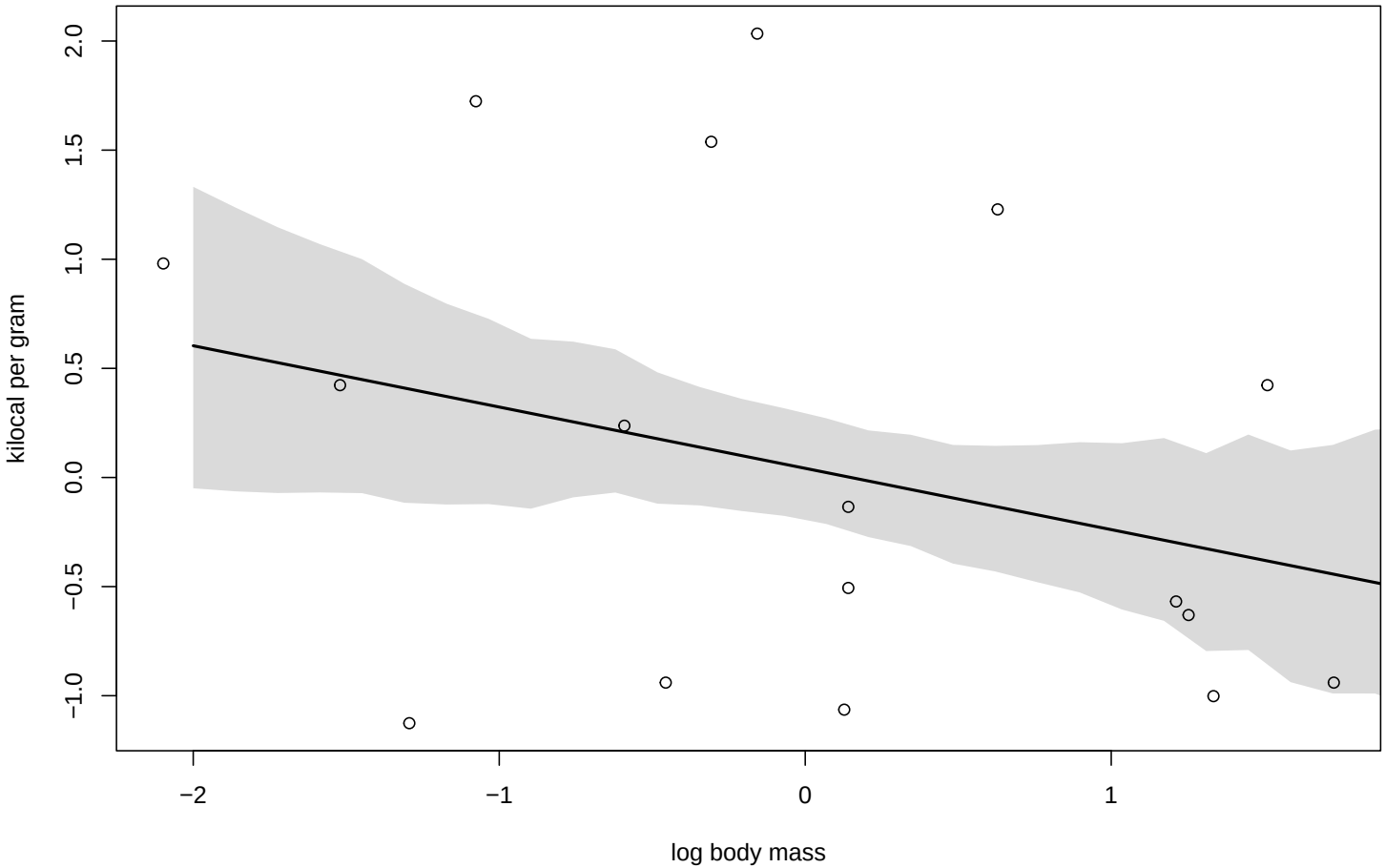
How about mass?

```
milk_m <- quap(
  alist(
    K ~ dnorm( mu, sigma),
    mu <- a + bM * M,
    a ~ dnorm(0, .2),
    bM ~ dnorm( 0, .5),
    sigma ~ dexp(1)
  ), data = dcc
)
```

University of Gdańsk

Posterior for mass

	mean	sd	5.5%	94.5%
a	0.04653301	0.1512793	-0.1952405	0.28830655
bM	-0.28253155	0.1928798	-0.5907908	0.02572768
sigma	0.94926762	0.1570567	0.6982607	1.20027453



University of Gdańsk

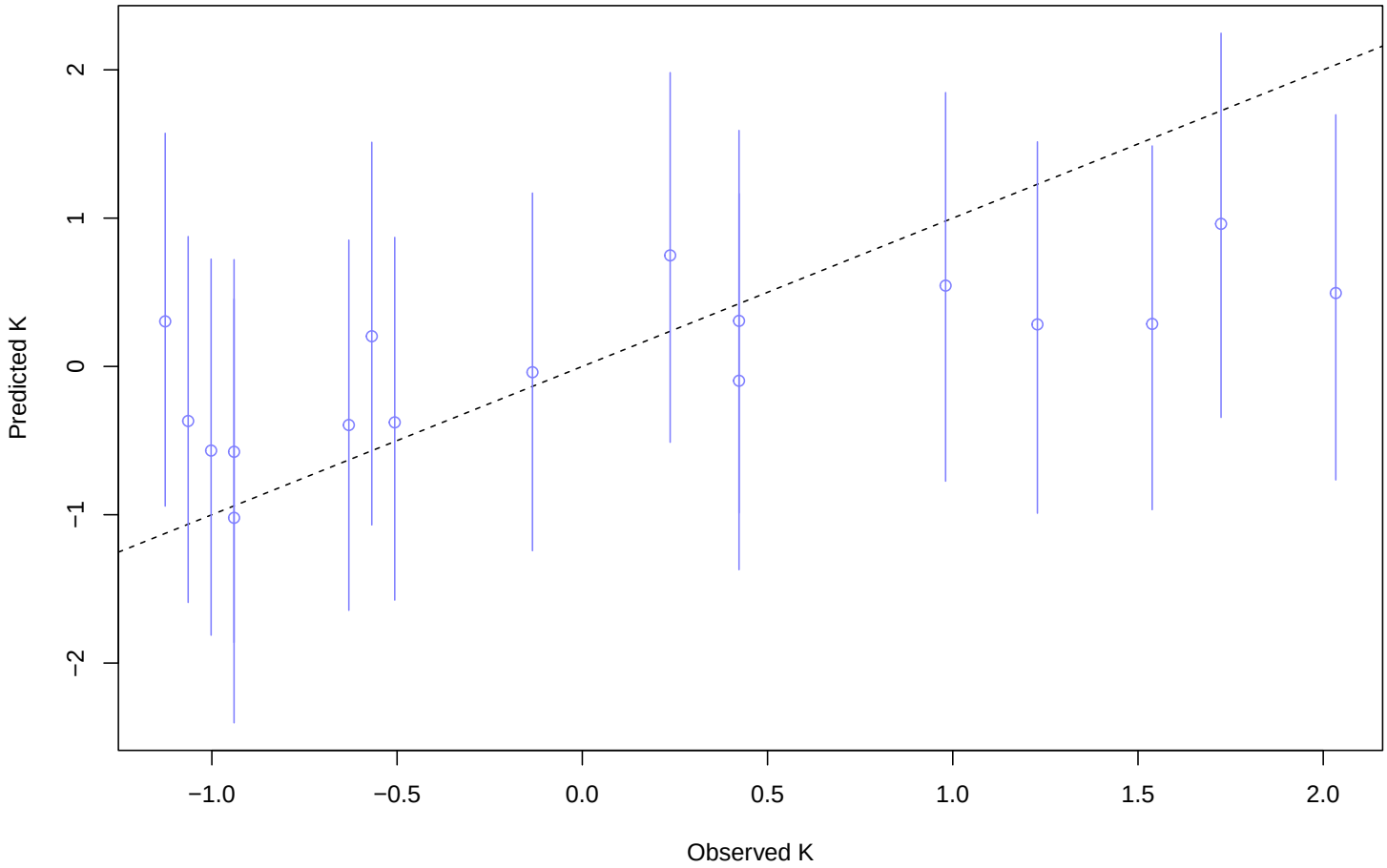
Now with both predictors

```
milk_mn <- quap(
  alist(
    K ~ dnorm( mu, sigma),
    mu <- a + bN * N + bM * M,
    a ~ dnorm(0, .2),
    bM ~ dnorm( 0, .5),
    bN ~ dnorm( 0, .5),
sigma ~ dexp(1)
  ), data = dcc
)
```

```
              mean         sd       5.5%      94.5%
a       0.06799322 0.1339995 -0.1461639  0.2821503
bM     -0.70297915 0.2207899 -1.0558441 -0.3501142
bN      0.67510990 0.2483012  0.2782766  1.0719432
sigma   0.73802270 0.1324655  0.5263172  0.9497282
```

University of Gdańsk

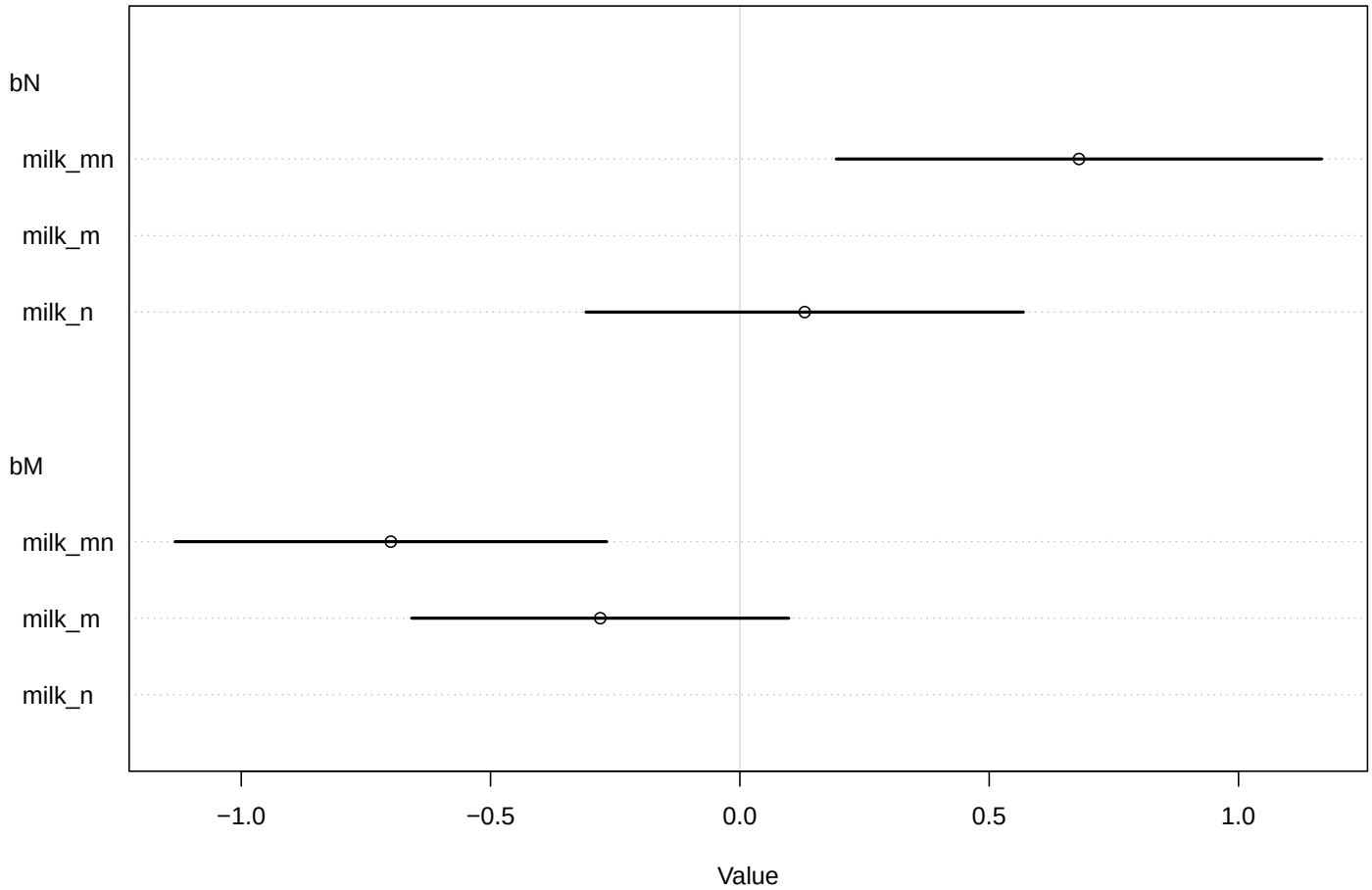
Quick Predictive Check



University of Gdańsk

Comparison of the models

```
plot(coeftab(milk_n, milk_m, milk_mn), pars = c("bN", "bM"))
```



University of Gdańsk

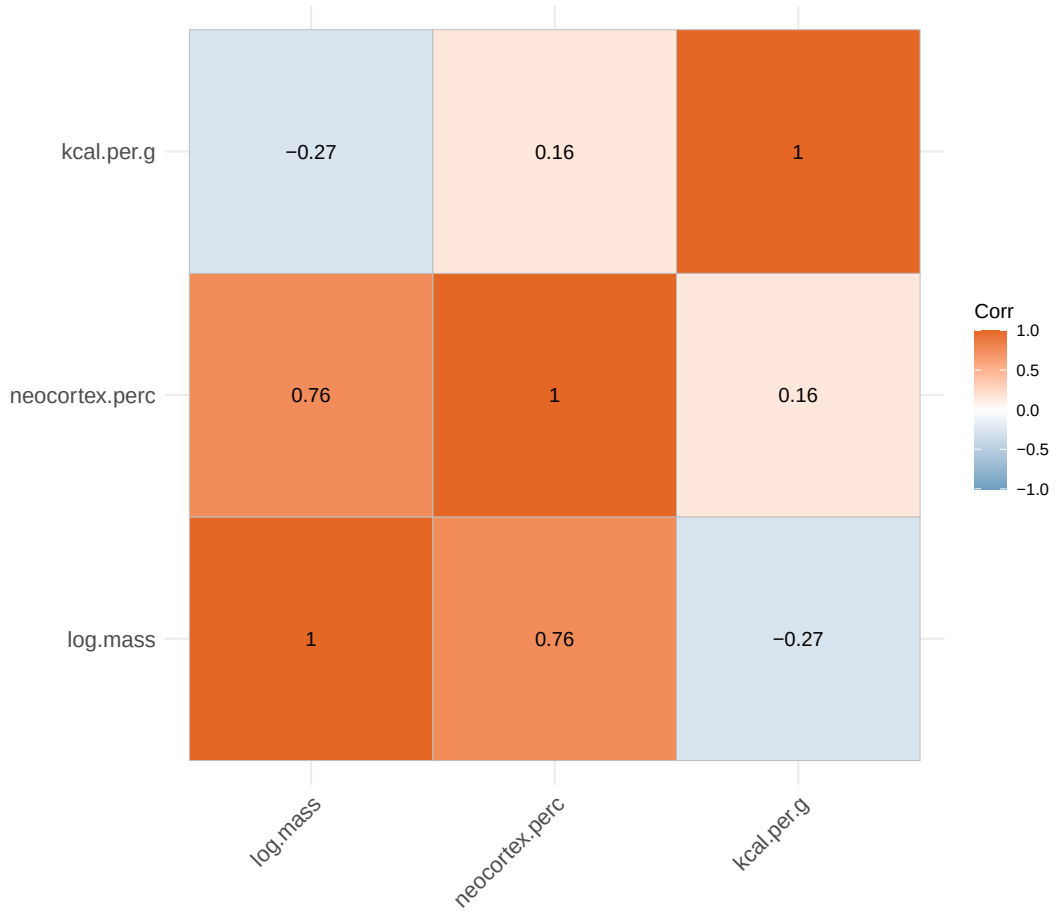
Comparison of the models

- Adding neocortex and body mass to the same model lead to a larger estimated effects of both!
- There are two variables correlated with the outcome, but one is positively correlated with it and the other is negatively correlated with it
- The correlation between those two predictors is very high

University of Gdańsk

Corrplot

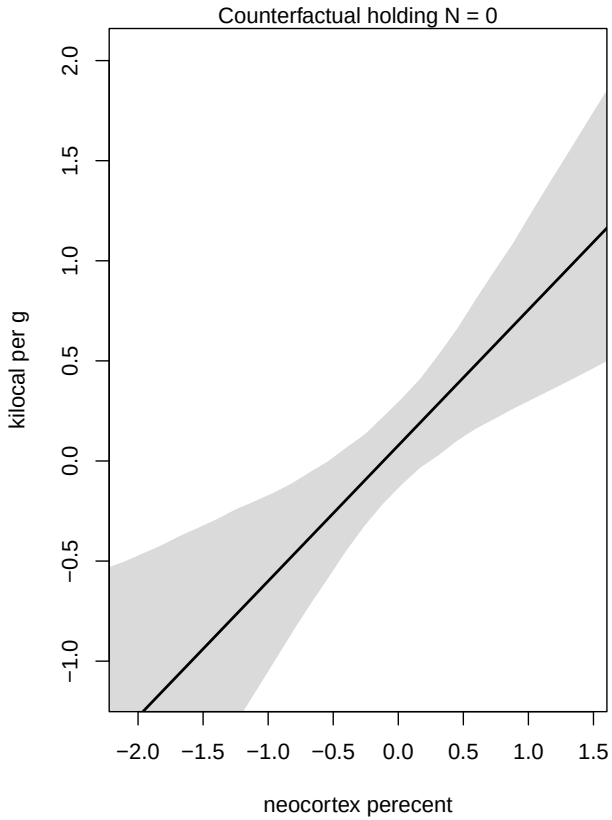
- Here we can observe cancelling each other out:
  - any influence log.mass may have on kcal.per.g is somewhat offset or "cancels out" by the influence of neocortex.perc

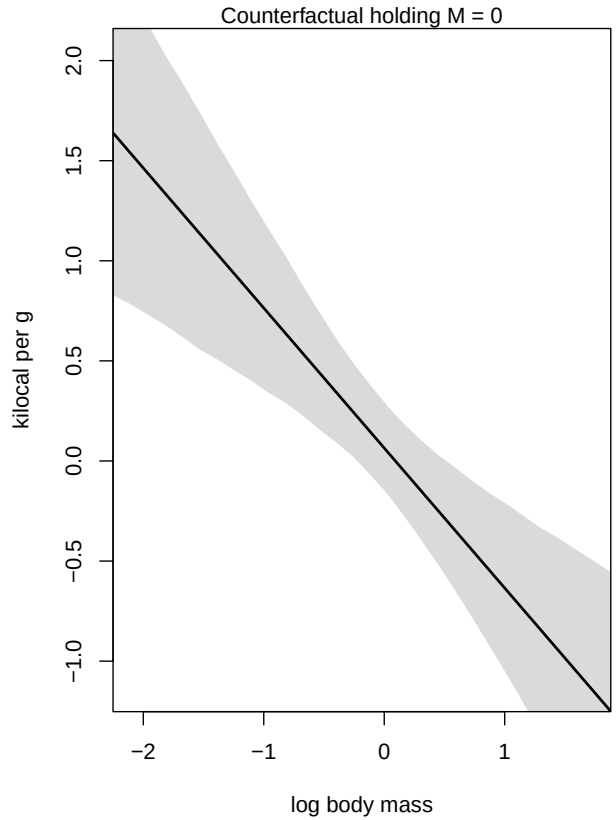


University of Gdańsk

Counterfactuals

- Bigger species, like apes, have milk with less energy. But species with more neocortex tend to have richer milk.
  - The fact that these two variables, body size and neocortex, are correlated across species makes it hard to see these relationships, unless we account for both

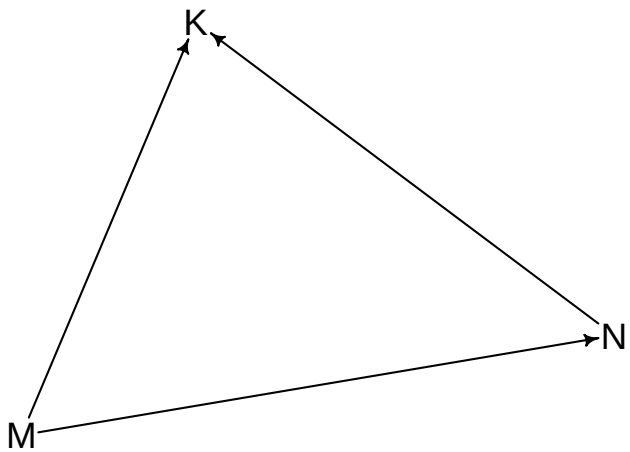


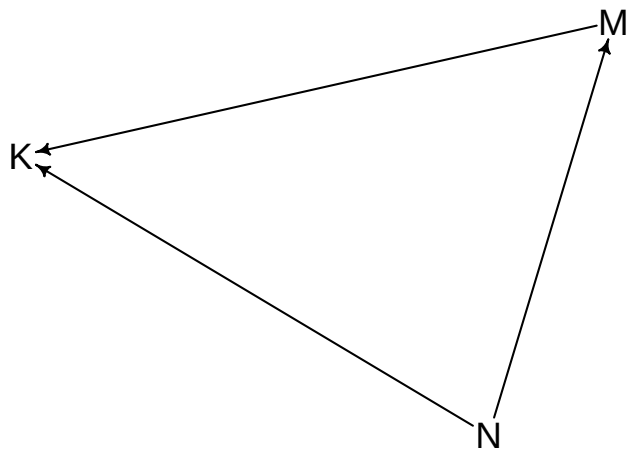


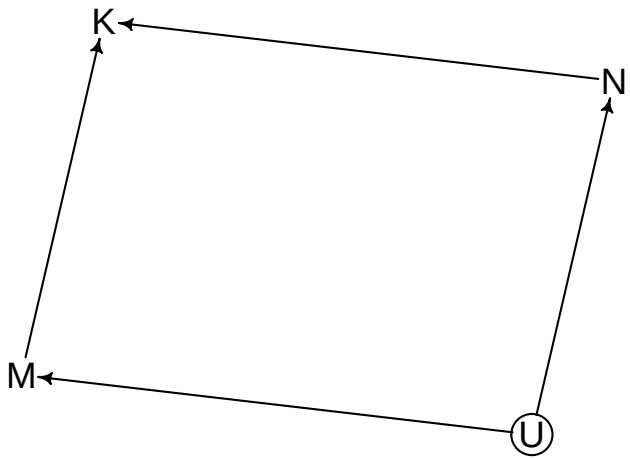
University of Gdańsk

Counterfactuals

```
par(mfrow = c(2, 2))
drawdag(milkDAG1, cex = 2, radius = 5)
drawdag(milkDAG2, cex = 2, radius = 5)
drawdag(milkDAG3, cex = 2, radius = 5)
```







University of Gdańsk

Categorical Variables

- For now we were dealing with continuous numerical variables
- But there are many variables of different type, one of them are categorical variables
  - Think about the difference between e.g. Height and Sex
- Consider a new dataset:

```
data(Howell1)
d <- Howell1
str(d)
```

```
'data.frame':   544 obs. of  4 variables:
 $ height: num  152 140 137 157 145 ...
 $ weight: num  47.8 36.5 31.9 53 41.3 ...
 $ age   : num  63 63 65 41 51 35 32 27 19 54 ...
 $ male  : int  1 0 0 1 0 1 0 1 0 1 ...
```

University of Gdańsk

Binary category: Sex vs. Height

```
d$sex <- ifelse( d$male==1 , 2 , 1 )
str( d$sex )
```

```
 num [1:544] 2 1 1 2 1 2 1 2 1 2 ...
```

$$
\begin{aligned}
\text{height}_i &\sim \text{Normal}(\mu_i, \sigma) \\
\mu_i &= \alpha_{\text{sex}[i]} \\
\alpha_j &\sim \text{Normal}(178, 20), \quad \text{for } j = 1, 2 \\
\sigma &\sim \text{Uniform}(0, 50)
\end{aligned}
$$

University of Gdańsk

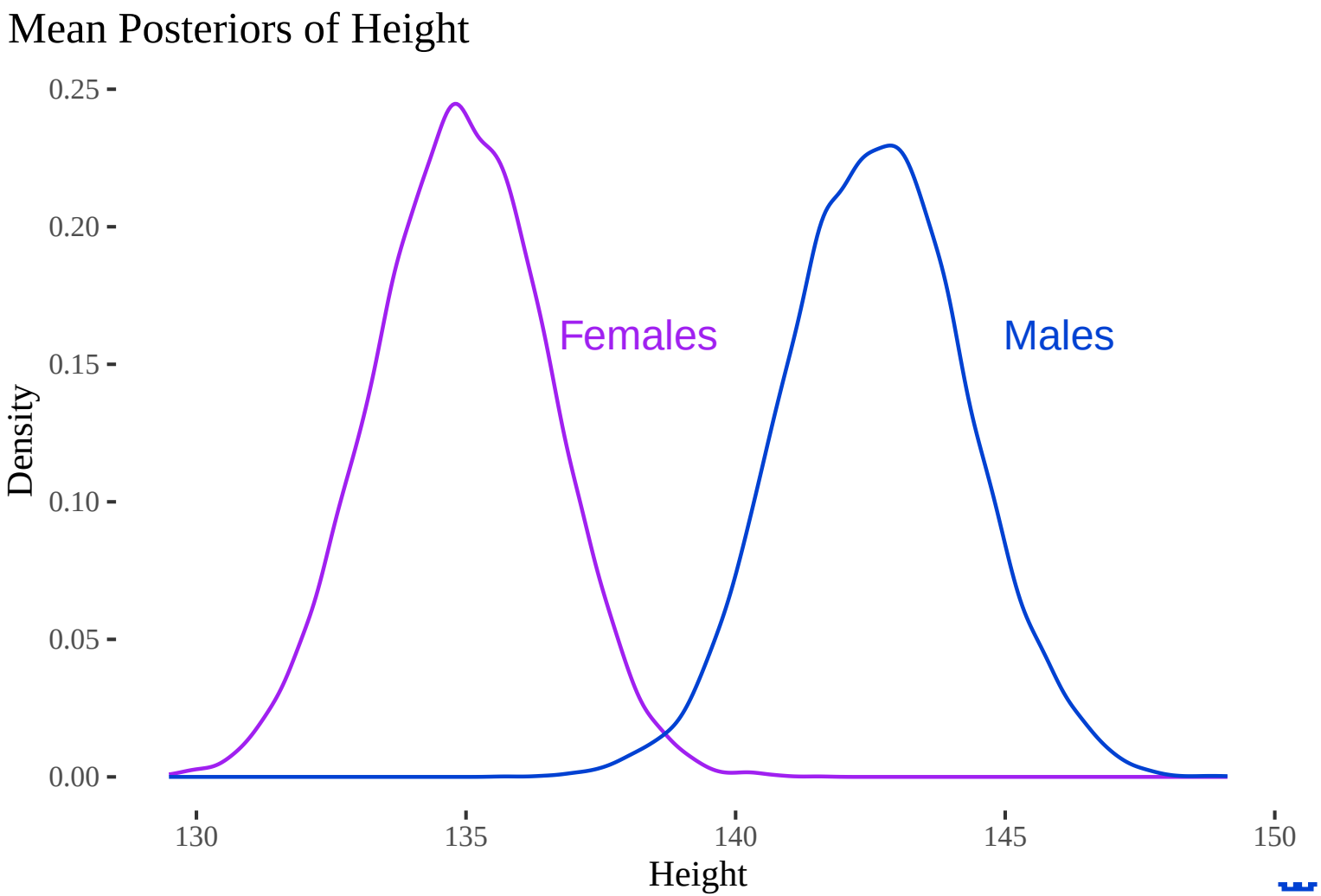
Binary category: Sex vs. Height

```
modelHeightSex <- quap(
alist(
height ~ dnorm( mu , sigma ) ,
mu <- a[sex] , # Notice square brackets '[]'
a[sex] ~ dnorm( 178 , 20 ) ,
sigma ~ dunif( 0 , 50 )
) , data=d )

# Notice depth = 2 !
precis(modelHeightSex, depth= 2)
```

```
          mean        sd      5.5%     94.5%
a[1]  134.9102 1.6069298 132.34200 137.47837
a[2]  142.5781 1.6974692 139.86526 145.29103
sigma  27.3099 0.8280375  25.98654  28.63326
```

University of Gdańsk

Binary category: Sex vs. Height



University of Gdańsk

Multiple Categorical Predictors

- Milk again!
- Here categories are not numbers, we will then change them into integers
  - So our model can recognize them

```
data(milk)
d <- milk
unique(d$clade)
```

```
[1] Strepsirrhine    New World Monkey Old World Monkey Ape
Levels: Ape New World Monkey Old World Monkey Strepsirrhine
```

```
d$K <- scale( d$kcal.per.g )
d$clade_id <- as.integer( d$clade )
```

University of Gdańsk

Multiple Categories: Clade vs. Milk Richness

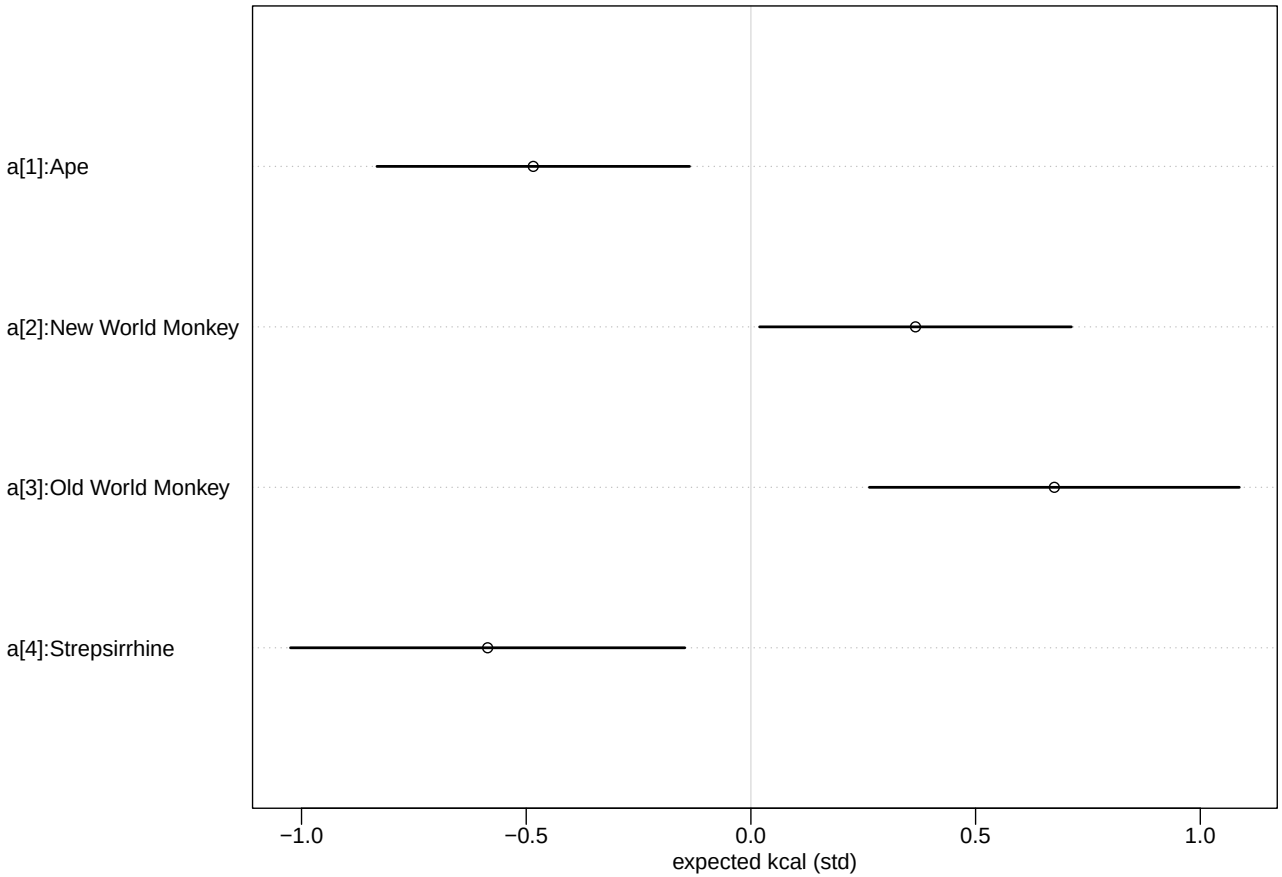
```
modelMilkClade <- quap(
alist(
K ~ dnorm( mu , sigma ),
mu <- a[clade_id],
a[clade_id] ~ dnorm( 0 , 0.5 ),
sigma ~ dexp( 1 )
) , data=d )

precis(modelMilkClade, depth = 2)
```

```
             mean         sd        5.5%      94.5%
a[1]  -0.4843413 0.21763503 -0.83216414 -0.1365185
a[2]   0.3662467 0.21705270  0.01935453  0.7131388
a[3]   0.6752459 0.25752626  0.26366921  1.0868226
a[4]  -0.5858275 0.27450142 -1.02453378 -0.1471212
sigma  0.7196196 0.09652481  0.56535434  0.8738849
```

University of Gdańsk

Multiple Categories: Clade vs. Milk Richness

```
labels <- paste( "a[" , 1:4 , "]:" , levels(d$clade) , sep="" )
plot( precis( modelMilkClade , depth=2 , pars="a" ) , labels=labels ,
xlab="expected kcal (std)" )
```



University of Gdańsk

More Variables!

- Of course we can add more then one categorical variable
- Here we made up new category, let's say that those monkeys are studying at Hogwart

```
set.seed(63)
#[1] Gryffindor, [2] Hufflepuff, [3] Ravenclaw, and [4]Slytherin
d$house <- sample( rep(1:4,each=8) , size=nrow(d) )
```

University of Gdańsk

More Variables: Clade, Milk, Hogwart Houses

```
modelMilkCladeHogwart <- quap(
alist(
K ~ dnorm( mu , sigma ),
mu <- a[clade_id] + h[house],
a[clade_id] ~ dnorm( 0 , 0.5 ),
h[house] ~ dnorm( 0 , 0.5 ),
sigma ~ dexp( 1 )
) , data=d )

precis(modelMilkCladeHogwart, depth = 2)
```

```
             mean         sd        5.5%        94.5%
a[1]  -0.4205602 0.26035044 -0.83665051 -0.004469931
a[2]   0.3836799 0.25968056 -0.03133976  0.798699616
a[3]   0.5664583 0.28903284  0.10452803  1.028388637
a[4]  -0.5055426 0.29664574 -0.97963974 -0.031445367
h[1]  -0.1025749 0.26170883 -0.52083620  0.315686313
h[2]  -0.1997060 0.27544061 -0.63991325  0.240501338
h[3]  -0.1603286 0.26905491 -0.59033029  0.269673121
h[4]   0.4866417 0.28751296  0.02714046  0.946142955
sigma  0.6631313 0.08812531  0.52229006  0.803972596
```

University of Gdańsk