Computational Statistics & Probability

Problem Set 2 - Linear Models
Due: 23:59:59 23.nov.2022

Fall 2022

Instructions

Assignments must be submitted through Canvas. See the course Canvas page for policies covering collaboration, acceptable file formats (`.Rmd` & `.pdf`), and late submissions. Completed assignments must include executable code (`.Rmd`) and a corresponding knitted markdown file (`pdf`). An R Markdown cheat sheet is available.

1. Multiple Regression & Causal Models

Return to the `Howell1` dataset and consider the causal relationship between *age* and *weight* in children. Let's define children as anyone younger than 13 and assume that age influences *weight* directly and through age-related physical changes that occur during development – physical attributes that a child's *height* will serve as proxy. We may summarize this causal background knowledge by the DAG:

$$
\begin{tikzcd}
& H \arrow[dr] & \\
A \arrow[ur] \arrow[rr] & & W
\end{tikzcd}
$$

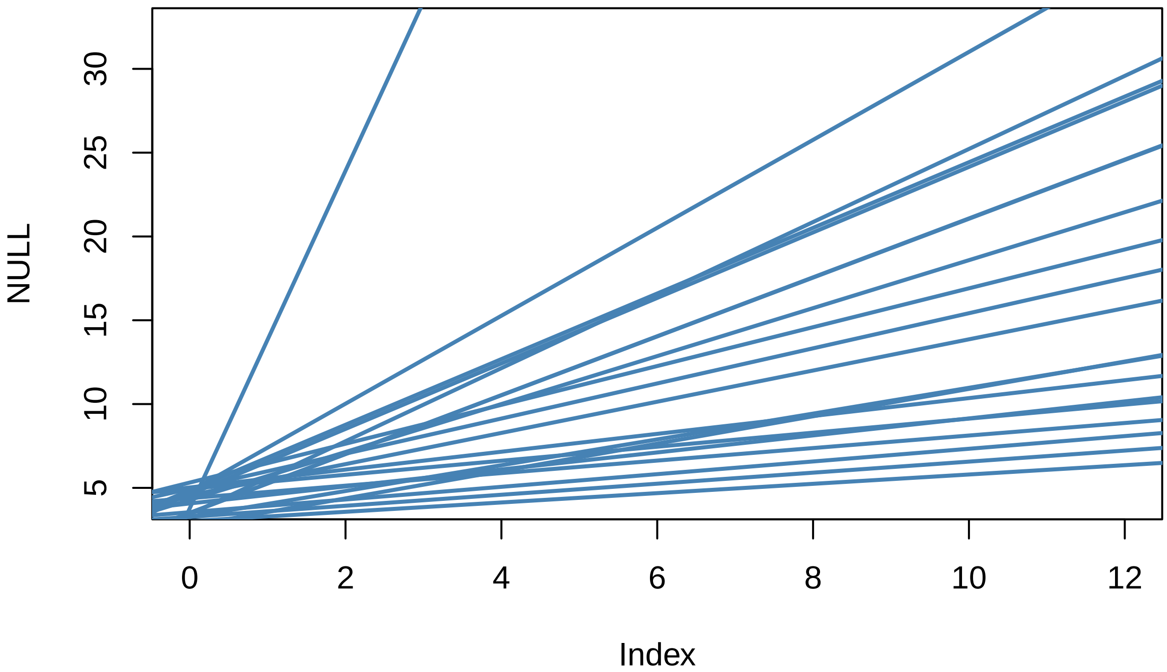
where A_i is *age* of child i, H_i is *height* of child i and W is *weight* of child i.

a) What is the total causal effect of year-by-year growth of !Kung children on their weight? Construct a linear regression (`m1a`) to estimate the total causal effect of each year of growth on a !Kung child's weight. Assume average birth weight is 4kg. Use prior predictive simulation to assess the implications of your priors.

```
# First, select only those people from the Howell dataset whose age is less than
# 13 years old
library(rethinking)
data(Howell1)
d <- Howell1
d <- d[ d$age < 13 , ]
```

```
# Next, we may simulate  priors.  We assume that average birth weight is 4 kilograms,
# which rounds up from the global average of 3.5 kgs. Thus, the prior on the
# intercept, a,is  normally distributed with a mean of 4 and SD = 1. Further, we
# assume that children get heavier as they grow, so the slope term bA is assumed
# to be non-negative.  We encode this assumption with a log-normal prior
# distribution with mean 0 and SD 1. Each prior is an n=20-dimensional vector.
#
set.seed(303)
n <- 20
a <- rnorm(n,4,1)
```

```r
bA <- rlnorm(n,0,1)
# blank(bty="n")
plot( NULL , xlim=range(d$age) , ylim=range(d$weight) )
for ( i in 1:n ) abline( a[i] , bA[i] , lwd=2 , col="steelblue" )
```



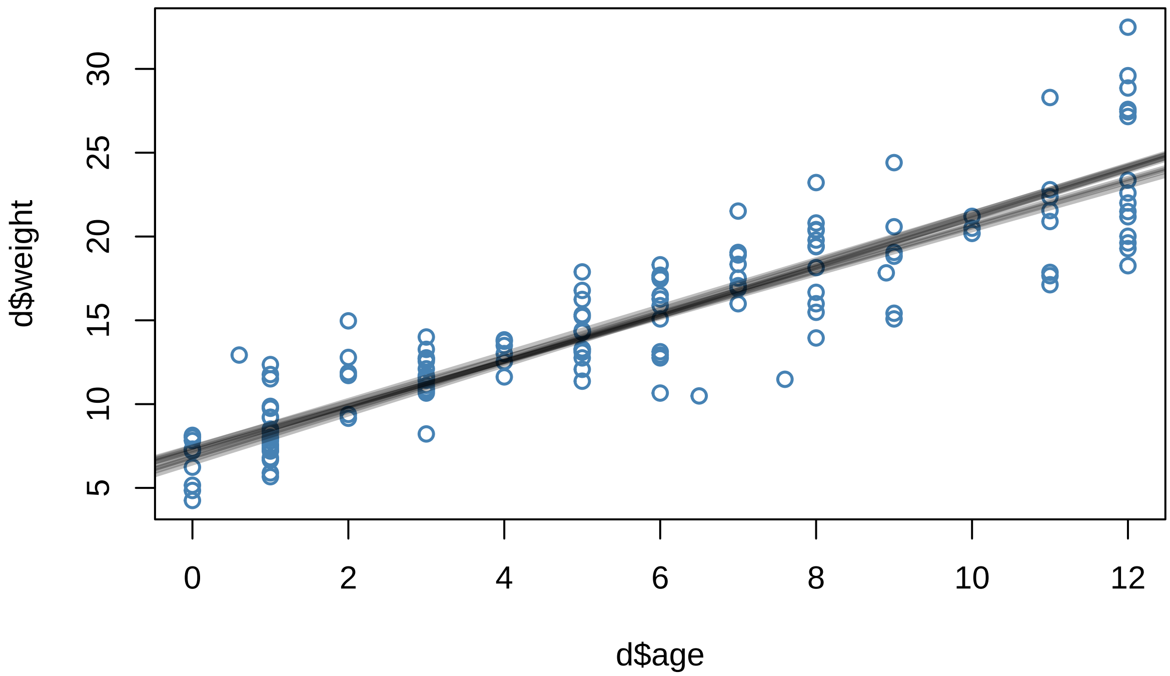

```r
# With these assumptions, we construct a linear model.
m1a <- quap(
    alist(
        W ~ dnorm( mu , sigma ),
        mu <- a + bA*A,
        a ~ dnorm(4,1),
        bA ~ dlnorm(0,1),
        sigma ~ dexp(1)
    ), data=list(W=d$weight,A=d$age) )
```

```r
precis(m1a)
```

```
##           mean         sd     5.5%    94.5%
## a     7.062987 0.34157033 6.517092 7.608883
## bA    1.388135 0.05264097 1.304005 1.472266
## sigma 2.512291 0.14613996 2.278731 2.745850
```

```r
# The causal effect of each year of grouth is given by bA, which is an average of
# 1.39 kg per year with an 89% CI of [1.30, 1.47].
```

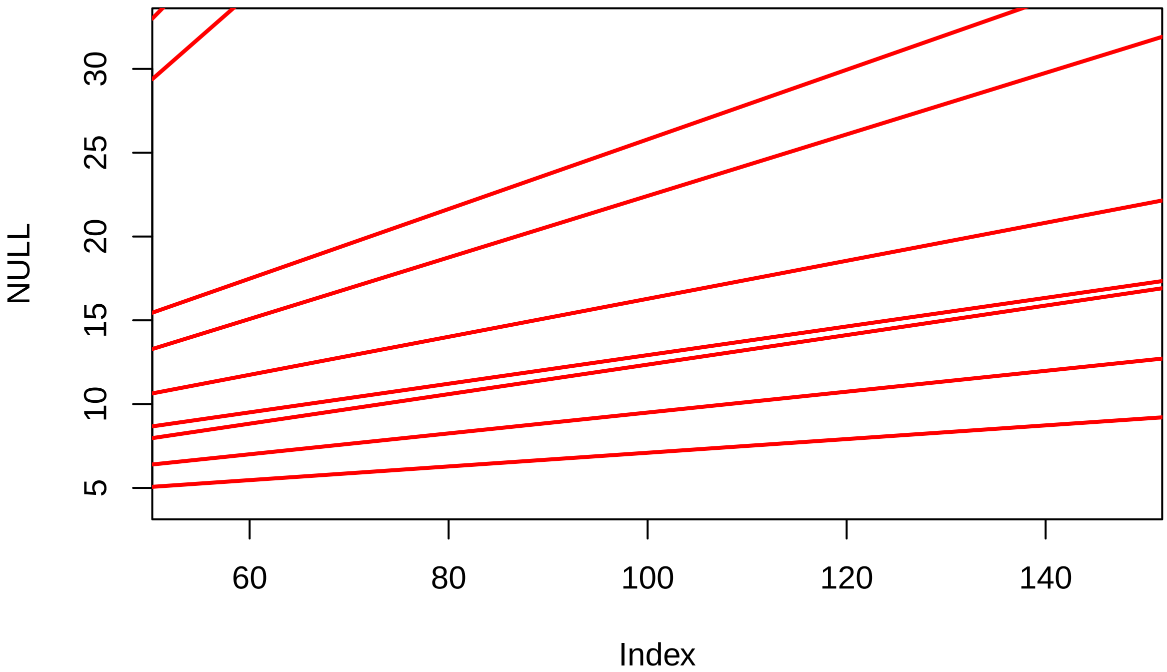
```r
# The following overlays 10 regression lines from the posterior `post` over the
# census data of !Kung children:

plot( d$age , d$weight , lwd=1.5, col="steelblue" )
post <- extract.samples(m1a)
for ( i in 1:10 ) abline( post$a[i] , post$b[i] , lwd=2.5 , col=alpha("black", 0.25) )
```




```
# Observe that the relationship between age and weight is approximately linear, but
# heteroskedastic: variance increases with age.
```

b) What is the total causal effect of height on weight? Construct a linear regression (`m1b`) to estimate the total causal effect height on a !Kung child's weight. Use prior predictive simulation to assess the implication of your priors.

```
# To simulate priors, we assume as before that average birth weight is 4 kilograms,
# which rounds up from the global average of 3.5 kgs. Thus, the prior on the
# intercept, a,is  normally distributed with a mean of 4 and SD = 1.

# Next we assume that children get heavier as they grow in height, so the slope
# term bH is assumed to be non-negative.  As before, this assumption is encoded
# a log-normal prior distribution with mean 0 but with SD 2.5.
# Each prior is an n=20-dimensional vector.
#
set.seed(303)
n <- 20
a <- rnorm(n,4,1)
bH <- rlnorm(n,0,2.5)
# blank(bty="n")
plot( NULL , xlim=range(d$height) , ylim=range(d$weight) )
for ( i in 1:n ) abline( a[i] , bH[i] , lwd=2 , col="red" )
```



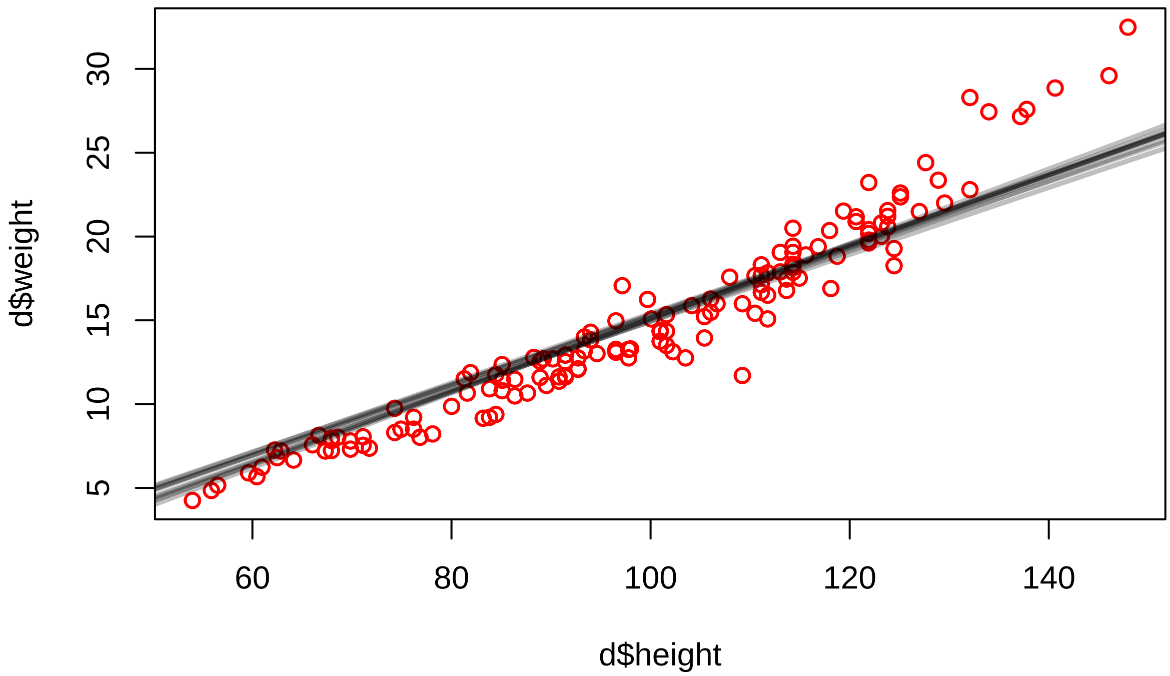

```r
# With these assumptions, we construct a linear model.
m1b <- quap(
    alist(
        W ~ dnorm( mu , sigma ),
        mu <- a + bH*H,
        a ~ dnorm(4,1),
        bH ~ dlnorm(0,2.5),
        sigma ~ dexp(1)
    ), data=list(W=d$weight,H=d$height) )
```

```r
precis(m1b)
```

```
##             mean          sd       5.5%      94.5%
## a     -5.8493950 0.763262371 -7.0692357 -4.6295543
## bH     0.2101009 0.007523872  0.1980763  0.2221255
## sigma  1.7955573 0.138347745  1.5744509  2.0166638
```

```r
#
# The causal effect of each year of grouth is given by bH, which is an average of
# 0.21 kg per centimeter of growth with an 89% CI of [0.20, 0.22].
```

```r
# The following overlays the 10 regression lines over the data of children
# in the !Kung census:

plot( d$height , d$weight , lwd=1.5, col="red" )
post <- extract.samples(m1b)
for ( i in 1:10 ) abline( post$a[i] , post$bH[i] , lwd=2.5 , col=alpha("black", 0.25) )
```




```
# Observe that the relationship between height and weight is not linear, but
# more homoskedastic: variance appears to vary much less with respect to height
# than it varies wrt age.

# Thus, model m1a has higher variance than m1b and m1b has higher bias than m1a.
# Specifically, predicting weight by age by m1a is more uncertain for older children
# than for younger children, and predicting weight by height by m1b is less accurate
# for older children than for younger children.
```

c) After knowing the age of a !Kung child, what additional value is there in also knowing the child's height? Conversely, after knowing the height of a !Kung child, what additional value is there in also knowing the child's age?

```
# To answer this causal question, we first must standardize our variables and
# add them to the dataframe d:

d$A <- standardize( d$age )
d$H <- standardize( d$height )
d$W <- standardize( d$weight )

# and rerun the bivariate regressions w/ standardized variables:

# for regressor `age`
m1aa <- quap(
    alist(
        W ~ dnorm( mu , sigma ),
        mu <- a + bA*A,
        a ~ dnorm(1,1),
        bA ~ dlnorm(0,1),
        sigma ~ dexp(1)
    ), data=d )
precis(m1aa)
```

```
##              mean         sd       5.5%      94.5%
## a     0.001326491 0.03642481 -0.0568874 0.05954038
## bA    0.895395080 0.03657110  0.8369474 0.95384276
## sigma 0.440410973 0.02571479  0.3993138 0.48150818
```

```
# for regressor `height`
m1bb <- quap(
    alist(
        W ~ dnorm( mu , sigma ),
        mu <- a + bH*H,
        a ~ dnorm(1,1),
        bH ~ dlnorm(0,2.5),
        sigma ~ dexp(1)
    ), data=d )
precis(m1bb)
```

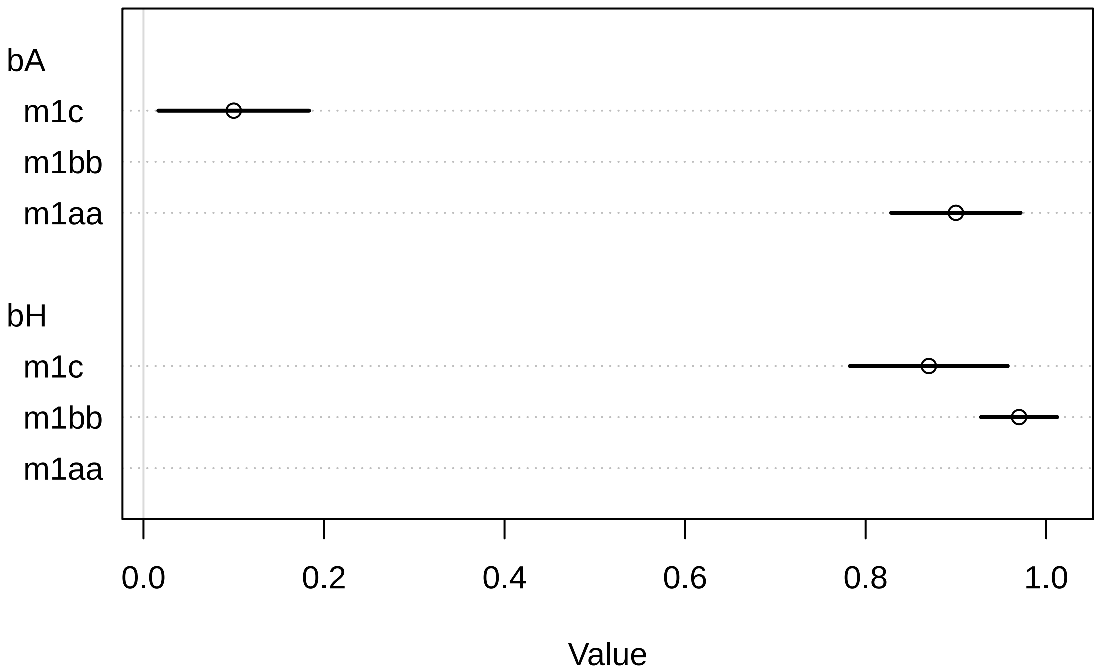
```
##               mean         sd        5.5%      94.5%
## a     0.0004650283 0.02143553 -0.03379309 0.03472315
## bH    0.9650794312 0.02151873  0.93068835 0.99947052
## sigma 0.2590652021 0.01513941  0.23486950 0.28326090
```

```
# Next, we construct a multiple regression model with
# regressors age and height

m1c <- quap(
    alist(
        W ~ dnorm( mu , sigma ),
        mu <- a + bA*A + bH*H,
        a ~ dnorm(0,1),
        bA ~ dlnorm(0,1),
        bH ~ dlnorm(0,1),
        sigma ~ dexp(1)
    ), data=d )
precis(m1c)
```

```
##                mean         sd        5.5%      94.5%
## a     -1.573231e-06 0.02136877 -0.03415300 0.03414985
## bA     1.033884e-01 0.04259037  0.03532078 0.17145605
## bH     8.701726e-01 0.04460096  0.79889164 0.94145354
## sigma  2.582588e-01 0.01512515  0.23408592 0.28243173
```

```
# Inspecting the coefficients table reveals that both age and height are
# positively associated with weight. But the question asks whether, once
# age is known, is there any additional predictive value in also knowing
# that child's height. To answer that question we need to compare the
# joint model with the two bivariate regressions from above.
```

```
plot( coeftab(m1aa, m1bb, m1c), par=c("bA", "bH"))
```




```
# Given the causal dag, both age and height are positively associated with weight.
# However, one is much more informative in predicting weight than the other.
# Once we know the height H of a child, there is only moderate if any improved
# predictive power in also knowing the child's age, as the mean for bH wrt `m1bb`is
# just above the upper bound of the 89% CI for `m1c`. On the other hand, age is
# only very strongly associated with predicting weight when height is missing
# from the model (m1aa).
```

2. Causal Influence with Categorical Variables

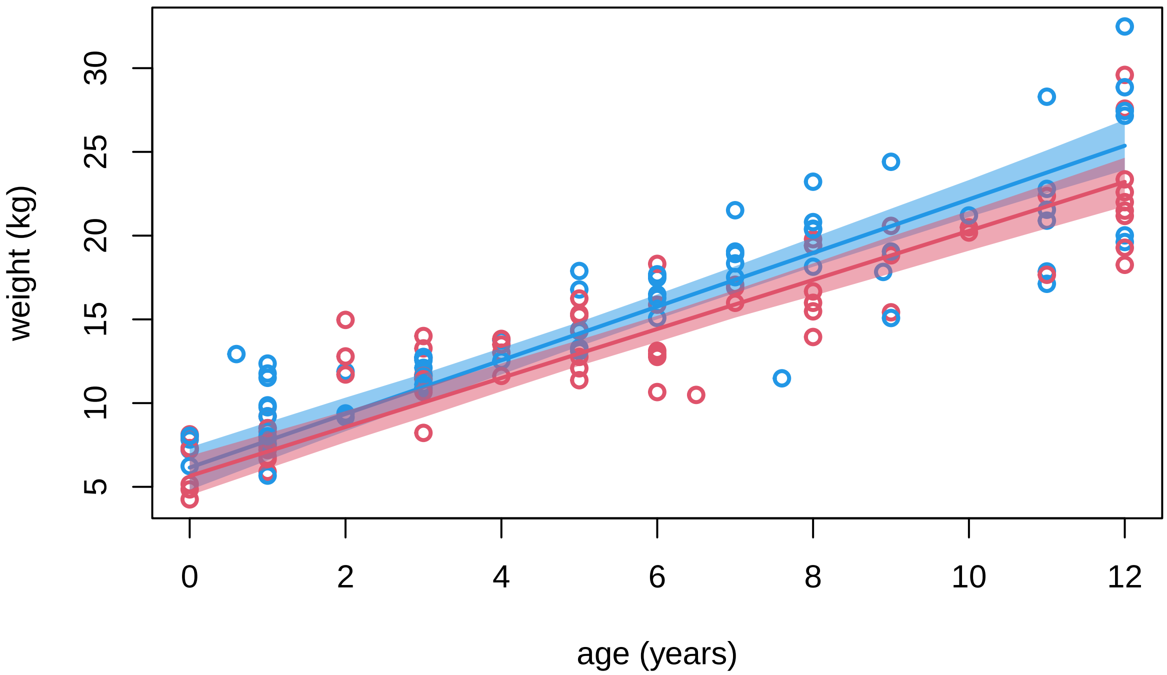
The causal relationship between age and weight might be different for girls and boys.

a) To investigate whether this is so, construct a single linear regression with a categorical variable for sex to estimate the total causal effect of age on weight separately for !Kung boys and girls. Plot your data and overlay the two regression lines, one for girls and one for boys.

```
# The model `m1a` can be modified to statify by sex. We construct an index
# variable, S, and change the coding from 0,1 to 1, 2.
data(Howell1)
d <- Howell1
d <- d[ d$age < 13 , ]

m2 <- quap(
    alist(
        W ~ dnorm( mu , sigma ),
        mu <- a[S] + b[S]*A,
        a[S] ~ dnorm(0,1),
        b[S] ~ dlnorm(0,1),
        sigma ~ dexp(1)
), data=list(W=d$weight,A=d$age,S=d$male+1) )
```

```
# Next, plot the data with regression lines:
plot( d$age , d$weight , lwd=2, col=ifelse(d$male==1,4,2) ,
    xlab="age (years)" , ylab="weight (kg)" )
```

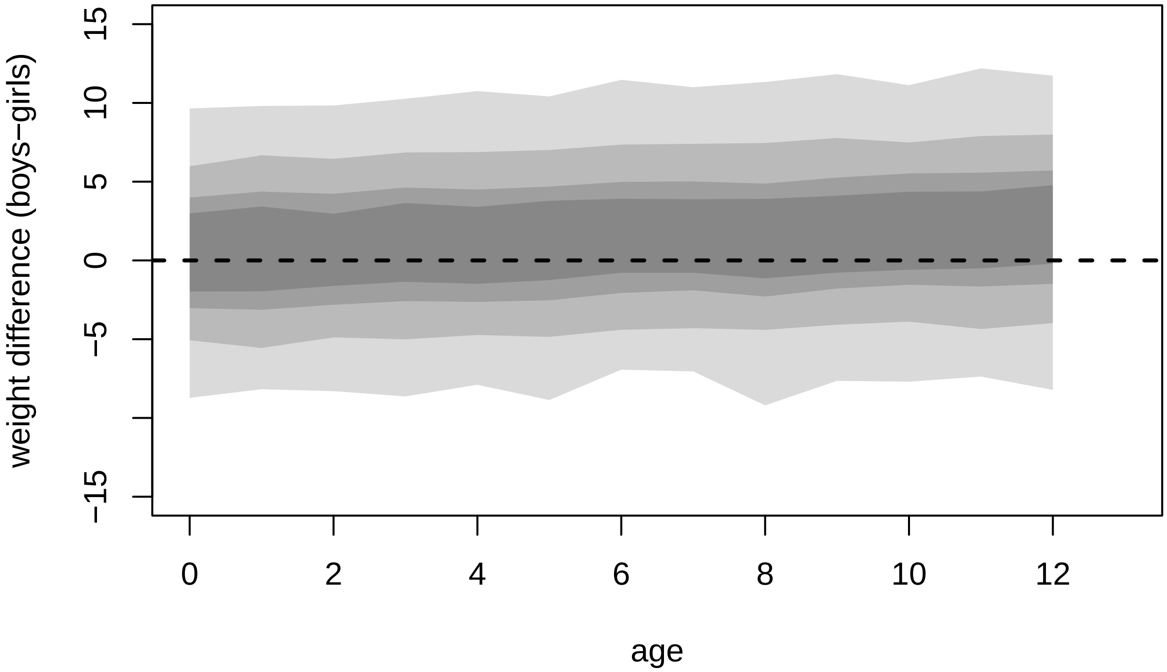
```
seq <- 0:12

# boys
muM <- link(m2,data=list(A=seq,S=rep(2,13)))
shade( apply(muM,2,PI,0.99) , seq , col=col.alpha(4,0.5) )
lines( seq , apply(muM,2,mean) , lwd=2 , col=4 )

# girls
muF <- link(m2,data=list(A=seq,S=rep(1,13)))
shade( apply(muF,2,PI,0.99) , seq , col=col.alpha(2,0.5) )
lines( seq , apply(muF,2,mean) , lwd=2 , col=2 )
```




```
# Boys appear to be slightly heavier than girls and to increase in weight slightly
# faster.
```

b) Do they differ? If so, provide one or more posterior contrasts as a summary.

```
# A posterior contrast across ages checks whether the trend which appears
# to be present in the simple regression plot is in fact true

seq <- 0:12

mu1 <- sim(m2,data=list(A=seq,S=rep(1,13)))
mu2 <- sim(m2,data=list(A=seq,S=rep(2,13)))

mu_contrast <- mu1
for ( i in 1:13 ) mu_contrast[,i] <- mu2[,i] - mu1[,i]

plot( NULL , xlim=c(0,13) , ylim=c(-15,15) , xlab="age" ,
    ylab="weight difference (boys-girls)" )
for ( p in c(0.5,0.67,0.89,0.99) ) # credibility intervals
shade( apply(mu_contrast,2,PI,prob=p) , seq )
abline(h=0,lty=2,lwd=2)
```




```
# The contrast plot uses the entire distribution, not just the expectation values.
# Even though the distributions overlap, boys tend to be heavier than girls at all
# ages and the difference increases with age.

# Note that, although the distributions overlap zero, it would be a mistake to
# infer that there is no difference in weight between girls and boys. Intervals
# that overlap zero do not entail that one's estimate is exactly zero.
```