Computational Statistics & Probability

Problem Set 2 - Linear Models

Author: Neelesh Bhalla
Collaborators: Nils Marthiensen, Chia-Jung Chang

2022-11-23

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:

```
library(rethinking)
```

```
## Loading required package: rstan
```

```
## Loading required package: StanHeaders
```

```
##
## rstan version 2.26.13 (Stan version 2.26.1)
```

```
## For execution on a local, multicore CPU with excess RAM we recommend calling
## options(mc.cores = parallel::detectCores()).
## To avoid recompilation of unchanged Stan programs, we recommend calling
## rstan_options(auto_write = TRUE)
## For within-chain threading using `reduce_sum()` or `map_rect()` Stan functions,
## change `threads_per_chain` option:
## rstan_options(threads_per_chain = 1)
```

```
## Loading required package: cmdstanr
```

```
## This is cmdstanr version 0.5.3
```

```
## - CmdStanR documentation and vignettes: mc-stan.org/cmdstanr
```

```
## - CmdStan path: /Users/neelesh/.cmdstan/cmdstan-2.30.1
```

```
## - CmdStan version: 2.30.1
```

```
##
## A newer version of CmdStan is available. See ?install_cmdstan() to install it.
## To disable this check set option or environment variable CMDSTANR_NO_VER_CHECK=TRUE.
```

```
## Loading required package: parallel
```

```
## rethinking (Version 2.21)
```

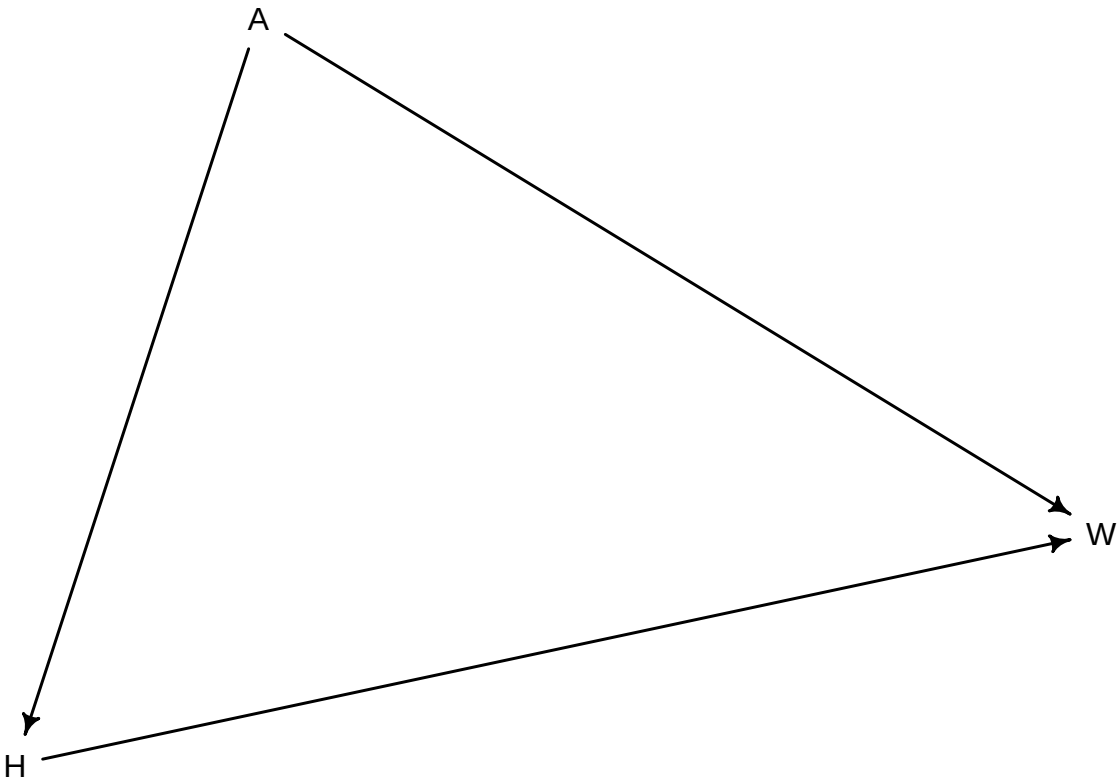
```
##
## Attaching package: 'rethinking'
```

```
## The following object is masked from 'package:rstan':
##
```

```
##     stan

## The following object is masked from 'package:stats':
##
##     rstudent
```

```
library(dagitty)
dag_q1 <- dagitty('dag{ W <- A -> H -> W }')
drawdag( dag_q1 )
```

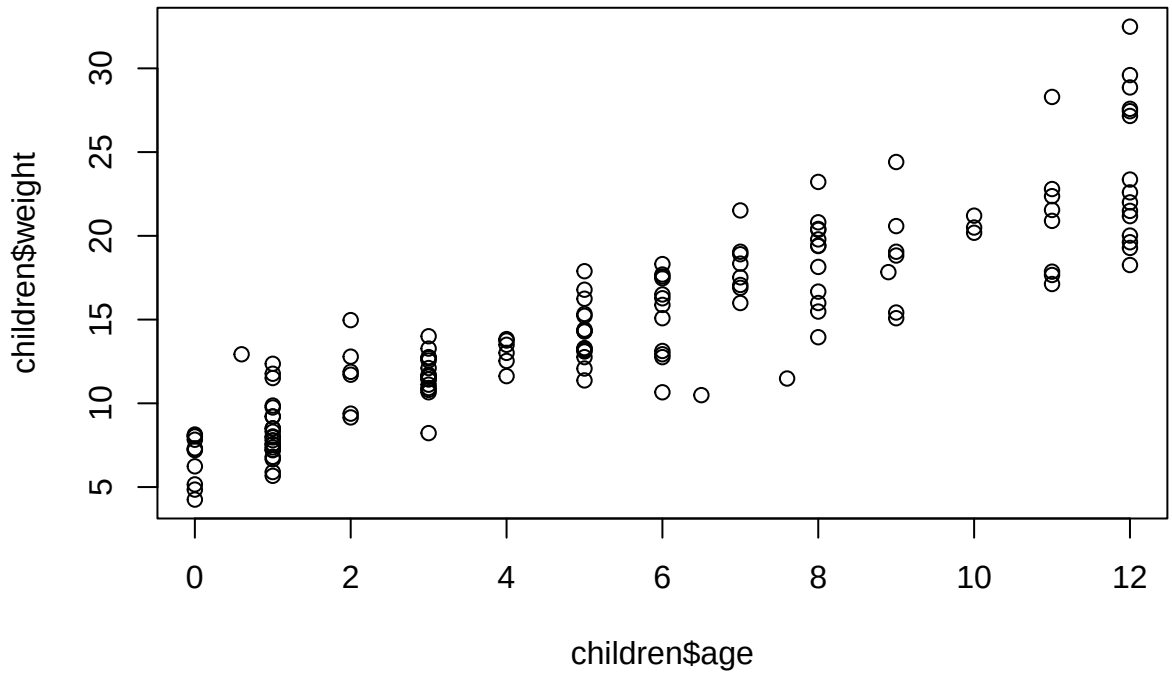


where Ai is age of child i, Hi is height of child i and W is weight of child i.

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.

```
impliedConditionalIndependencies(dag_q1)
#There are no conditional independence, so there is no output to display
```

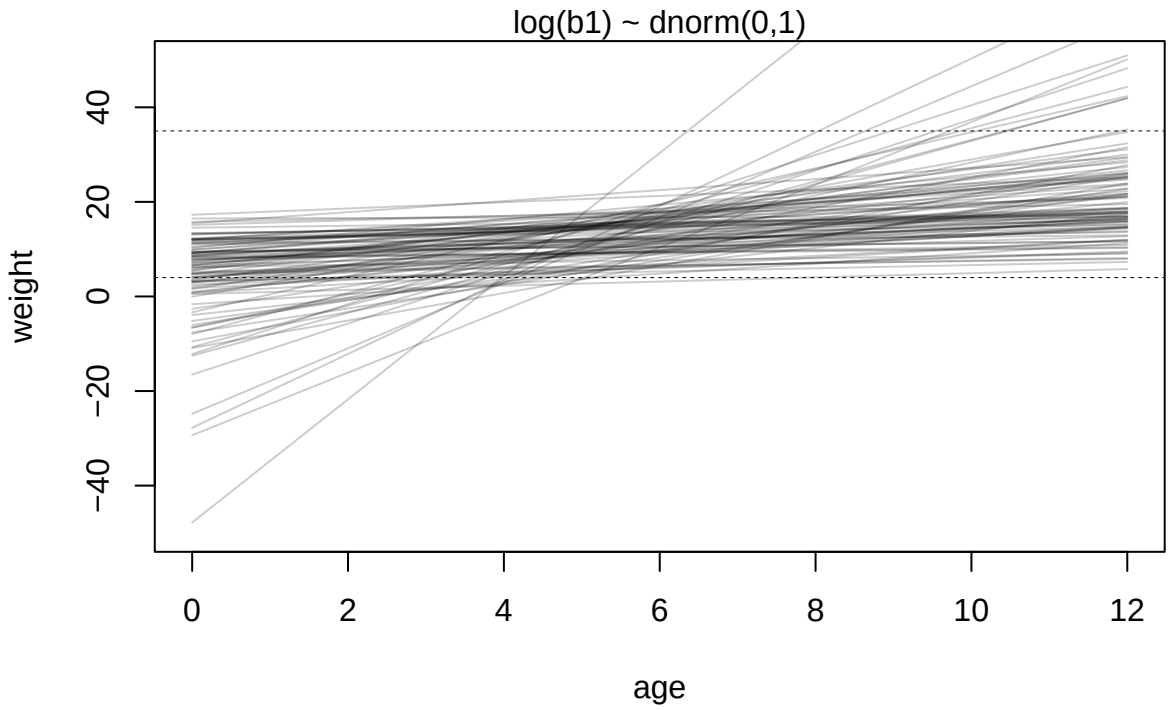
```
data(Howell1)
d <- Howell1
children <- d[ d$age < 13 , ]
plot( children$weight ~ children$age )
```




```
# taking an idea of the relation between age and weight
```

```
N <- 100
a <- rnorm( N , 12 , 4 )
b1 <- rlnorm( N , 0 , 1 )
```

```
# Prior predictive simulation for the weight and age model
plot( NULL , xlim=range(children$age) , ylim=c(-50,50) ,
xlab="age" , ylab="weight" )
abline( h=4 , lty=2, lwd=0.5 )
abline( h=35 , lty=2 , lwd=0.5 )
mtext( "log(b1) ~ dnorm(0,1)" )
xbar <- mean(children$age)
for ( i in 1:N ) curve( a[i] + b1[i]*(x - xbar) ,
from=min(children$age) , to=max(children$age) , add=TRUE ,
col=col.alpha("black",0.2) )
```




```
# Posterior distribution for weight and age model
xbar <- mean(children$age)
# fit model
m1a <- quap(
alist(
weight ~ dnorm( mu , sigma ) ,
mu <- a + exp(log_b1)*( age - xbar ) ,
a ~ dnorm( 12 , 4 ) ,
log_b1 ~ dnorm( 0 , 1 ) ,
sigma ~ dunif( 0 , 10 )
) ,
data=children )
```

```
precis( m1a )
```

```
##              mean         sd      5.5%      94.5%
## a      14.6860027 0.20861765 14.352591 15.0194140
## log_b1  0.2934821 0.04084369  0.228206  0.3587582
## sigma   2.5241507 0.14771611  2.288072  2.7602296
```

```
# the marginal posterior distributions
```

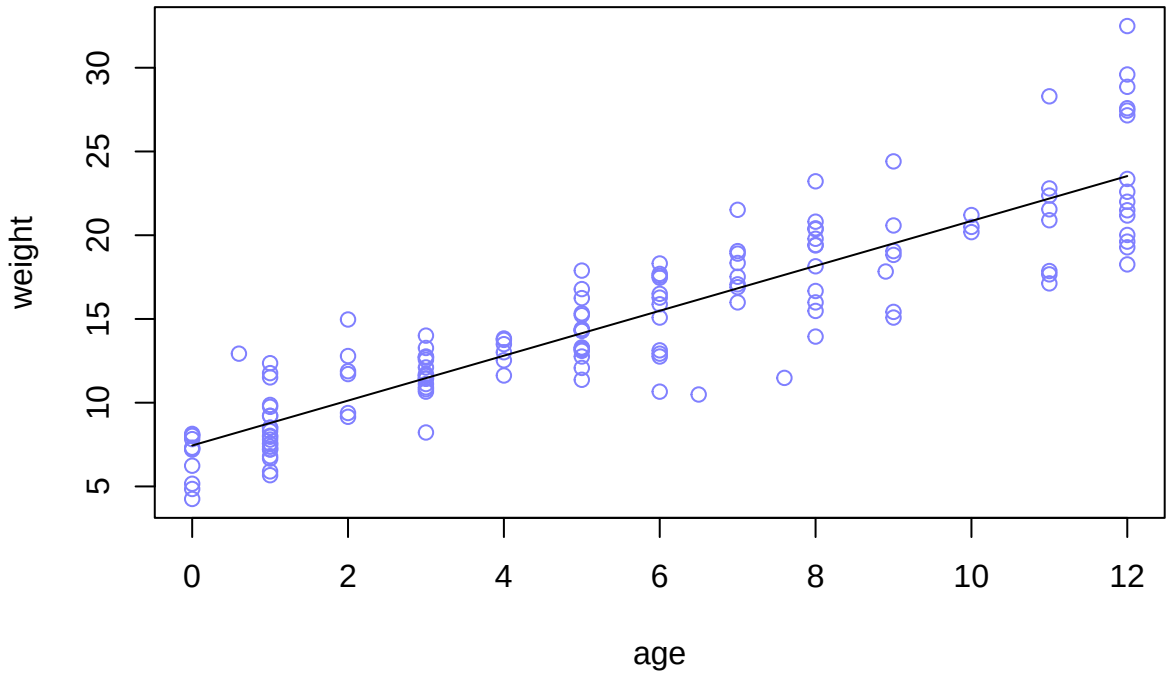
```
round( vcov( m1a ) , 3 )
```

```
##            a log_b1 sigma
## a      0.044  0.000 0.000
## log_b1 0.000  0.002 0.000
## sigma  0.000  0.000 0.022
```

```
# the variance-covariance matrix
```

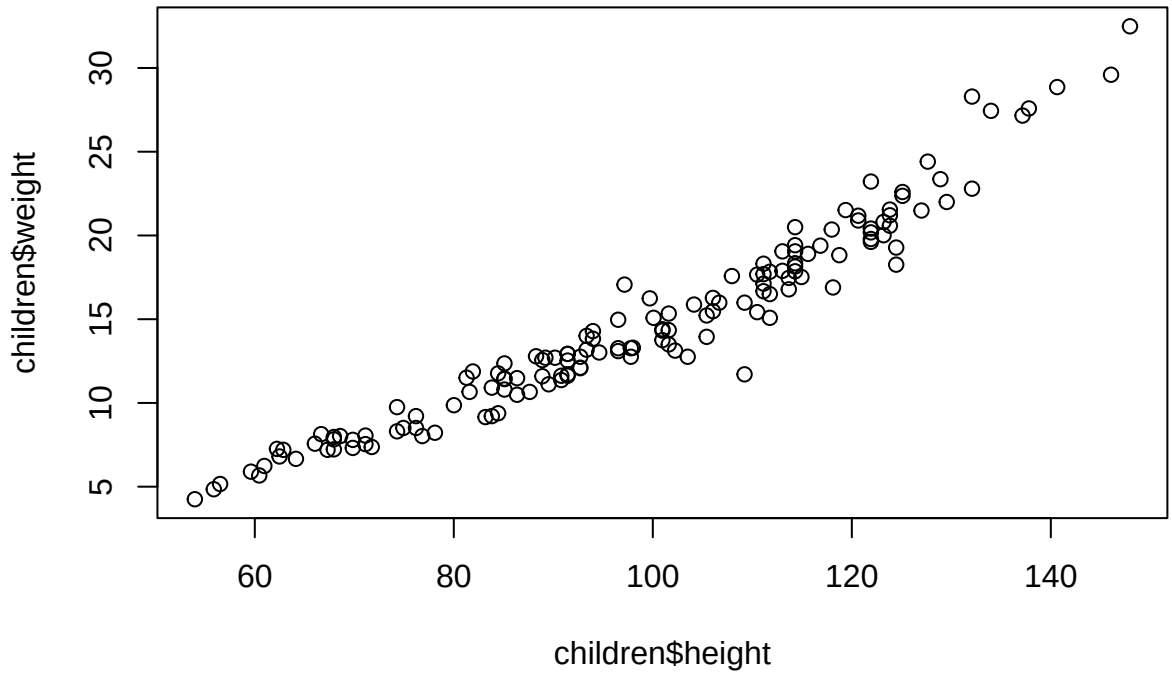
```
plot( weight ~ age , data=children , col=rangi2 )
post <- extract.samples( m1a )
a_map <- mean(post$a)
```

```
b1_map <- exp(mean(post$log_b1))
curve( a_map + b1_map*(x - xbar) , add=TRUE )
```



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.

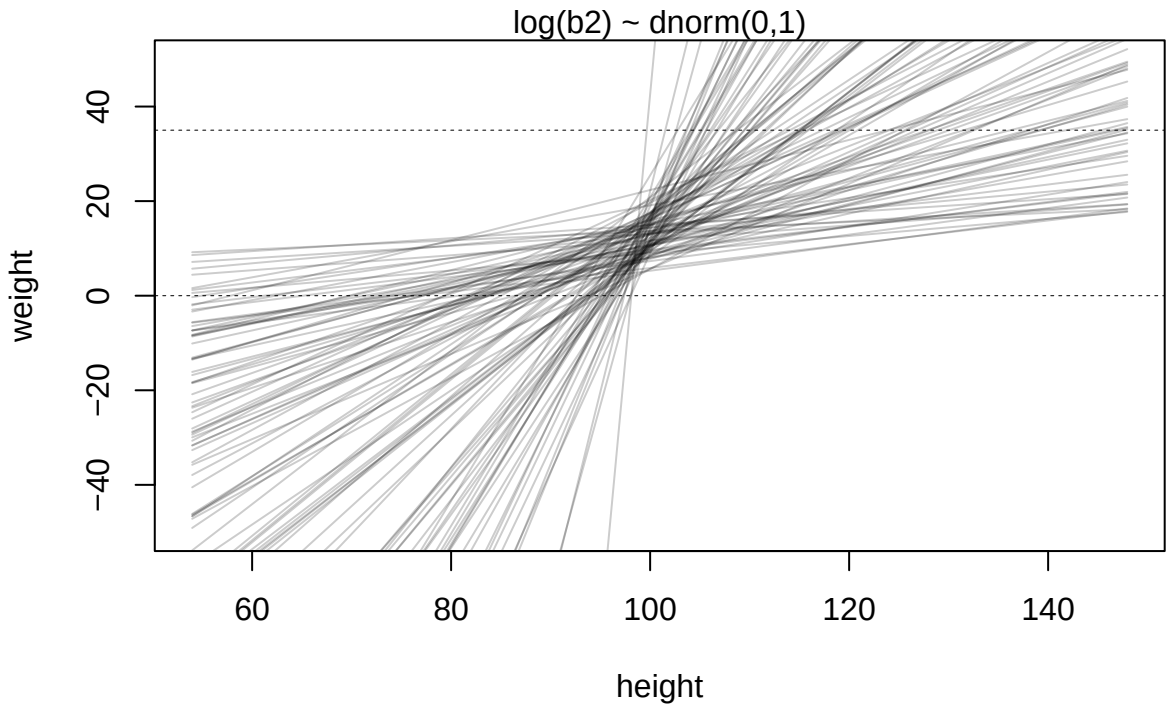
```
plot( children$weight ~ children$height )
```




```
# taking an idea of the relation between height and weight
```

```
N <- 100
a <- rnorm( N , 12 , 4)
b2 <- rlnorm( N , 0 , 1 )
```

```
# Prior predictive simulation for the weight and height model
plot( NULL , xlim=range(children$height) , ylim=c(-50,50) ,
xlab="height" , ylab="weight" )
abline( h=0 , lty=2, lwd=0.5 )
abline( h=35 , lty=2 , lwd=0.5 )
mtext( "log(b2) ~ dnorm(0,1)" )
xbar <- mean(children$height)
for ( i in 1:N ) curve( a[i] + b2[i]*(x - xbar) ,
from=min(children$height) , to=max(children$height) , add=TRUE ,
col=col.alpha("black",0.2) )
```




```
# Posterior distribution for weight and height model
xbar <- mean(children$height)
# fit model
m1b <- quap(
alist(
weight ~ dnorm( mu , sigma ) ,
mu <- a + exp(log_b2)*( height - xbar ) ,
a ~ dnorm( 12 , 4 ) ,
log_b2 ~ dnorm( 0 , 1 ) ,
sigma ~ dunif( 0 , 10 )
) ,
data=children )
```

```
precis( m1b )
```

```
##             mean         sd      5.5%     94.5%
## a      14.691379 0.12275967 14.495186 14.887573
```

```
## log_b2 -1.357512 0.02227601 -1.393114 -1.321911
## sigma   1.484007 0.08685568  1.345195  1.622819
```

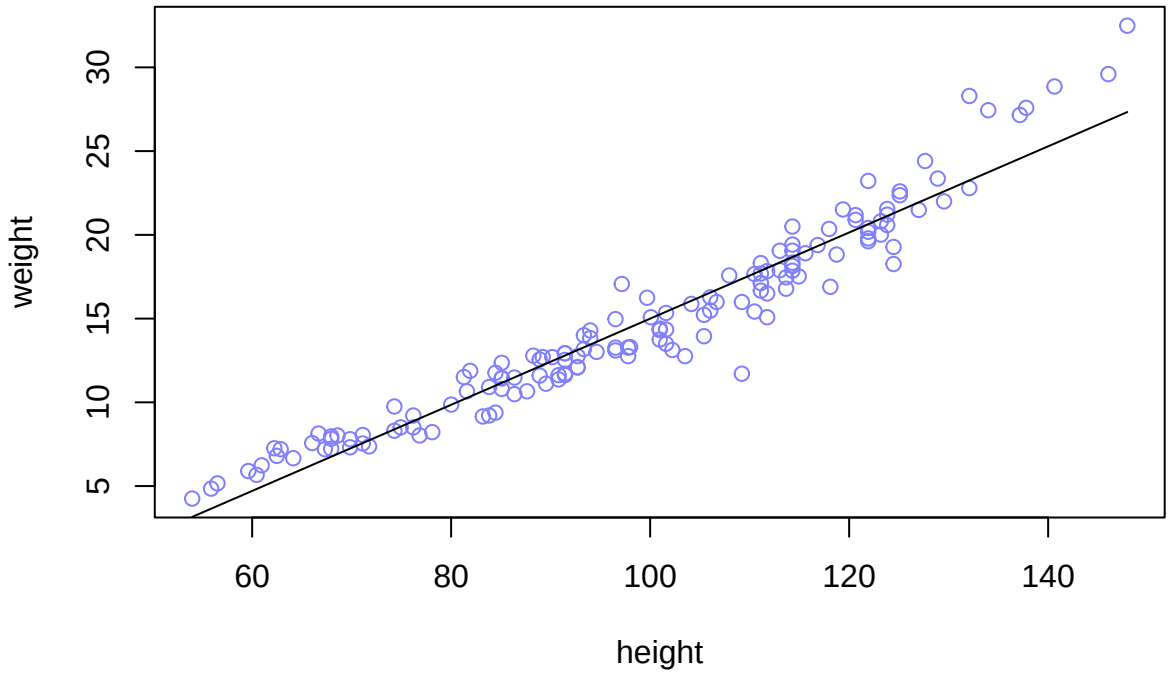
```
# the marginal posterior distributions
```

```
round( vcov( m1b ) , 3 )
```

```
##            a log_b2 sigma
## a      0.015      0 0.000
## log_b2 0.000      0 0.000
## sigma  0.000      0 0.008
```

```
# the variance-covariance matrix
```

```
plot( weight ~ height , data=children , col=rangi2 )
post <- extract.samples( m1b )
a_map <- mean(post$a)
b2_map <- exp(mean(post$log_b2))
curve( a_map + b2_map*(x - xbar) , add=TRUE )
```



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?

```
children$A <- scale( children$age )
children$H <- scale( children$height )
children$W <- scale( children$weight )
#fit models
m1_age <- quap(
alist(
W ~ dnorm( mu , sigma ) ,
mu <- a + bA * A ,
a ~ dnorm( 0 , 0.2 ) ,
bA ~ dnorm( 0 , 0.5 ) ,
sigma ~ dexp( 1 )
```

```
) , data = children )
m1_height <- quap(
alist(
W ~ dnorm( mu , sigma ) ,
mu <- a + bH * H ,
a ~ dnorm( 0 , 0.2 ) ,
bH ~ dnorm( 0 , 0.5 ) ,
sigma ~ dexp( 1 )
) , data = children )
m1_all <- quap(
alist(
W ~ dnorm( mu , sigma ) ,
mu <- a + bA*A + bH*H ,
a ~ dnorm( 0 , 0.2 ) ,
bA ~ dnorm( 0 , 0.5 ) ,
bH ~ dnorm( 0 , 0.5 ) ,
sigma ~ dexp( 1 )
) , data = children )
# Visualize results
```

```
precis( m1_all )
```

```
##                mean         sd        5.5%      94.5%
## a     -1.034309e-06 0.02121835 -0.03391205 0.03390998
## bA     7.348459e-02 0.05348912 -0.01200136 0.15897054
## bH     8.964452e-01 0.05349125  0.81095582 0.98193452
## sigma  2.578374e-01 0.01507132  0.23375052 0.28192426
```

```
round( vcov( m1_all ) , 3 )
```

```
##       a     bA     bH sigma
## a     0  0.000  0.000     0
## bA    0  0.003 -0.003     0
## bH    0 -0.003  0.003     0
## sigma 0  0.000  0.000     0
```

```
# the variance-covariance matrix
```

```
library(arulesViz)
```

```
## Loading required package: arules
```

```
## Loading required package: Matrix
```

```
##
## Attaching package: 'arules'
```

```
## The following objects are masked from 'package:base':
##
##     abbreviate, write
```

```
library(raster)
```

```
## Loading required package: sp
```

```
##
## Attaching package: 'sp'
```

```
## The following objects are masked from 'package:dagitty':
```

```
##
##     coordinates, coordinates<-
```

```
library(Rgraphviz)
```

```
## Loading required package: graph
```

```
## Loading required package: BiocGenerics
```

```
##
## Attaching package: 'BiocGenerics'
```

```
## The following objects are masked from 'package:raster':
##
##     as.data.frame, intersect, match, union, unique, which.max,
##     which.min
```

```
## The following objects are masked from 'package:arules':
##
##     duplicated, intersect, match, setdiff, sort, union, unique
```

```
## The following objects are masked from 'package:rethinking':
##
##     dims, normalize
```

```
## The following objects are masked from 'package:stats':
##
##     IQR, mad, sd, var, xtabs
```

```
## The following objects are masked from 'package:base':
##
##     anyDuplicated, aperm, append, as.data.frame, basename, cbind,
##     colnames, dirname, do.call, duplicated, eval, evalq, Filter, Find,
##     get, grep, grepl, intersect, is.unsorted, lapply, Map, mapply,
##     match, mget, order, paste, pmax, pmax.int, pmin, pmin.int,
##     Position, rank, rbind, Reduce, rownames, sapply, setdiff, sort,
##     table, tapply, union, unique, unsplit, which.max, which.min
```

```
##
## Attaching package: 'graph'
```

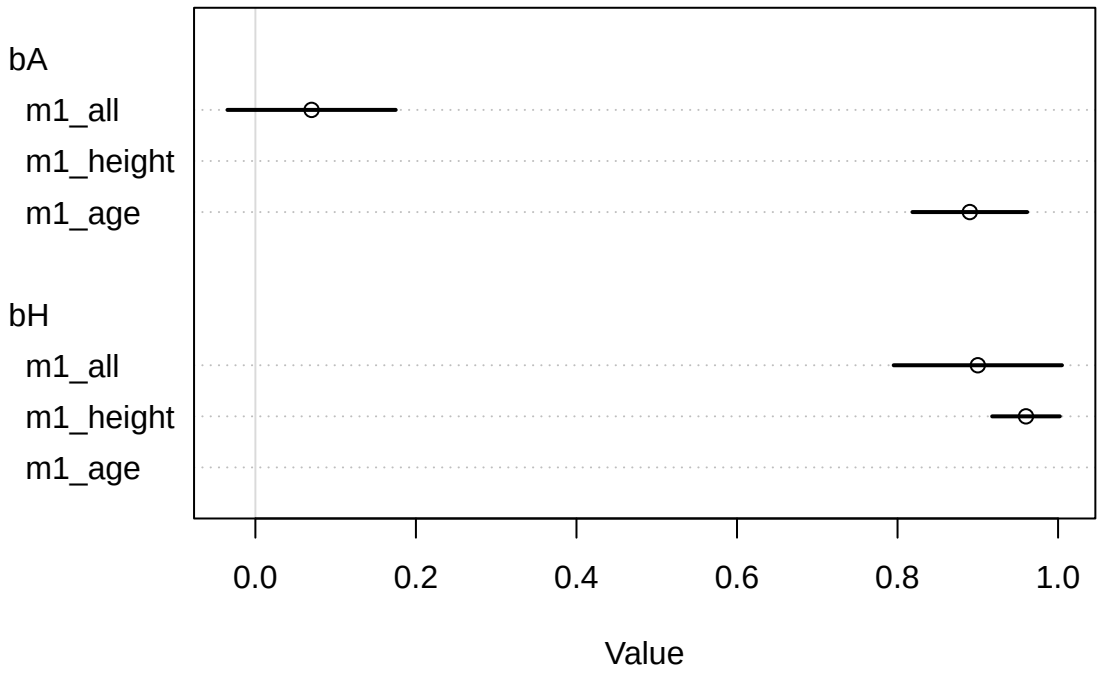
```
## The following object is masked from 'package:dagitty':
##
##     edges
```

```
## Loading required package: grid
```

```
##
## Attaching package: 'Rgraphviz'
```

```
## The following object is masked from 'package:dagitty':
##
##     graphLayout
```

```
plot( coeftab(m1_age,m1_height,m1_all), par=c("bA","bH") )
```




```
# from the above plot it is evident that height has a bigger stronger
# overall influence on the weight of the children. Thus the causal relationship
# is stronger between height-weight
# Thus, one can conclude that after knowing the height of a !Kung child,
# there is almost little additional value in also knowing the child's age?
```

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.

```
data(Howell1)
d <- Howell1
children <- d[ d$age < 13 , ]
```

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

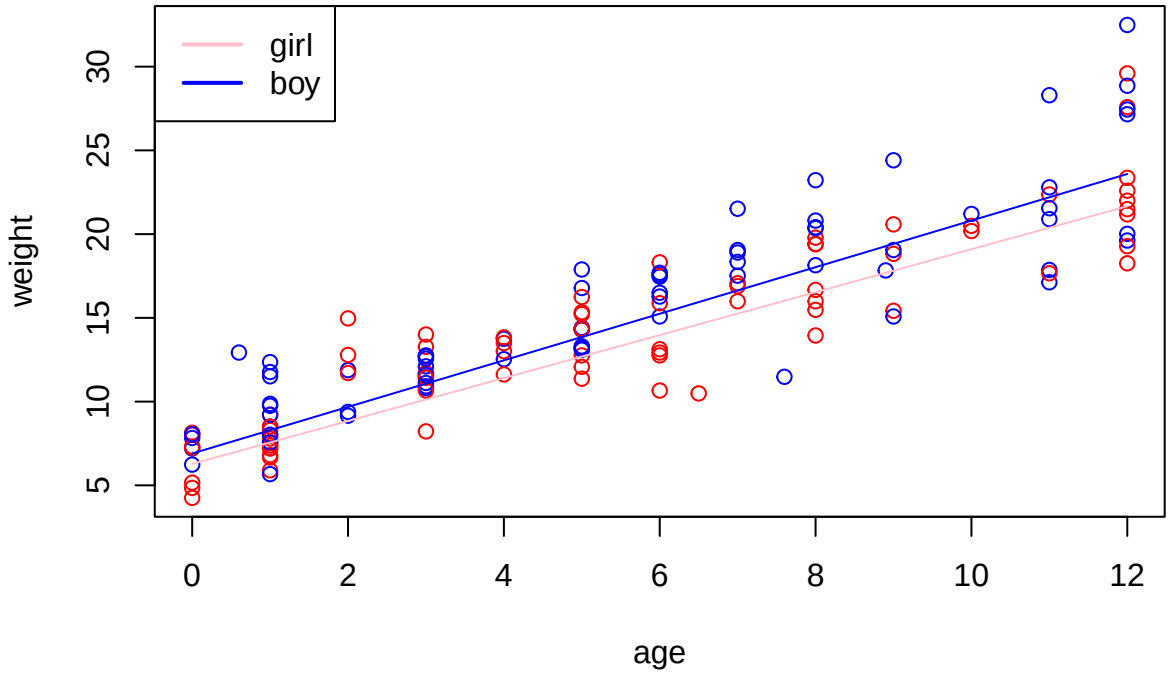
```
##  num [1:146] 2 1 1 1 2 2 1 1 1 1 ...
```

```
# indexing --> now "1" means female and "2" means male
```

```
xbar <- mean(children$age)
# fit model
m2a <- quap(
alist(
weight ~ dnorm( mu , sigma ) ,
mu <- a[sex] + b[sex]*( age - xbar ) ,
a[sex] ~ dnorm( 4 , 1 ) ,
b[sex] ~ dlnorm( 0.5 , 1 ) ,
sigma ~ dunif( 0 , 10 )
) , data=children )
```

```
# here both 'a' and 'b' are assumed to be dependent on the sex
# It is explainable because the weights of a boy child and a girl child
# are different at birth. Moreover, the rate of growth of weights are also
# different for both genders.
```

```
# plot data and both regression lines
plot( weight ~ age , data=children , col=c("red","blue")[children$sex])
post <- extract.samples( m2a )
a_map_girl <- mean(post$a[,1])
a_map_boy <- mean(post$a[,2])
b_map_girl <- mean(post$b[,1])
b_map_boy <- mean(post$b[,2])
curve( a_map_girl + b_map_girl*(x - xbar) , add=TRUE , col="pink")
curve( a_map_boy + b_map_boy*(x - xbar) , add=TRUE , col="blue")
legend(x = "topleft",
legend = c("girl", "boy"),
lty = c(1, 1),
col = c("pink", "blue"),
lwd = 2)
```




```
precis( m2a , depth=2 )
```

```
##            mean         sd      5.5%     94.5%
## a[1]  13.205823 0.30322253 12.721214 13.690431
## a[2]  14.406959 0.31668501 13.900835 14.913083
## b[1]   1.281660 0.07791211  1.157141  1.406178
## b[2]   1.391138 0.08098978  1.261701  1.520576
## sigma  2.586109 0.17168118  2.311729  2.860489
```

```
# the marginal posterior distributions
```

```
round( vcov( m2a ) , 3 )
```

```
##        a___1  a___2 b___1  b___2  sigma
```

```
## a___1  0.092  0.012 0.001  0.000 -0.017
## a___2  0.012  0.100 0.000 -0.001 -0.020
## b___1  0.001  0.000 0.006  0.000  0.000
## b___2  0.000 -0.001 0.000  0.007  0.000
## sigma -0.017 -0.020 0.000  0.000  0.029
```

```
# the variance-covariance matrix
```

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

```
seq <- 0:12
mu1 <- sim(m2a,data=list(age=seq,sex=rep(1,13)))
mu2 <- sim(m2a,data=list(age=seq,sex=rep(2,13)))
mu_contrast <- mu1
for ( i in 1:13 ) mu_contrast[,i] <- mu2[,i] - mu1[,i]
plot( x=seq, y=(colMeans(mu2)-colMeans(mu1)), type="l" , 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)
```