Homework 2

Used libraries are:

- rethinking
- ggplot2

Question 1

The weights listed below were recorded in the !Kung census, but heights were not recorded for these individuals. Provide predicted heights and 89% compatibility intervals for each of these individuals. That is, fill in the table below, using model-based predictions

```
data(Howell1)

d <- Howell1
d2 <- d[d$age >= 18, ]
```

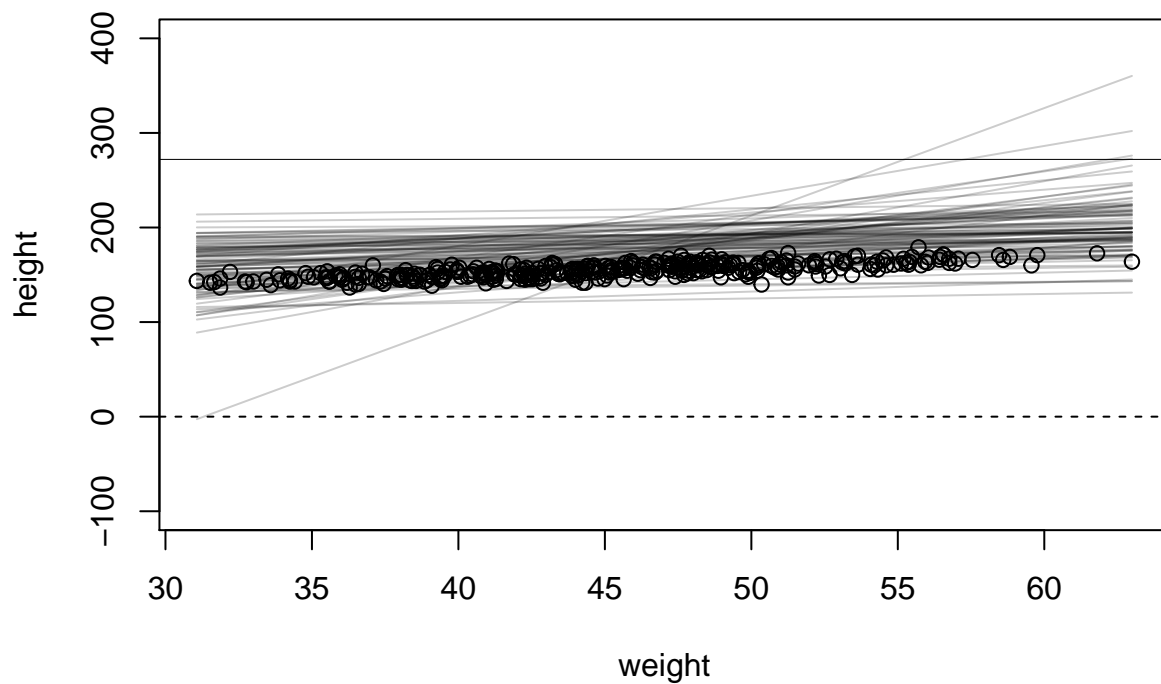Lets first visualize the prior predictive simulation

```
N <- 100
a <- rnorm(N, 178, 20)
b <- rlnorm(N, 0, 1)

plot(height~weight, data=d2, xlim=range(d2$weight), ylim=c(-100, 400), xlab="weight", ylab="height")

abline(h=0, lty=2)
abline(h=272, lty=1, lwd=0.5)

xbar <- mean(d2$weight)

for(i in 1:N) curve(a[i] + b[i] * (x-xbar), from=min(d2$weight), to=max(d2$weight), add=TRUE, col=col.a
```

Prior predictive simulation makes sense, lets get posterior distribution

```
flist <- alist(
  height ~ dnorm(mu, sigma),
  mu ~ a + b * (weight - xbar),
  a ~ dnorm(178, 20),
  b ~ dlnorm(0, 1),
  sigma ~ dunif(0, 50)
)


homework_1_model <- quap(flist, data=d2)
```

Now we have the model, lets simulate posterior predictive distributions

```
new_data <- data.frame( weight=c(45,40,65,31,53) )

simulated <- sim(homework_1_model, data=new_data, n=1e4)

df <- data.frame(apply(simulated, 2, PI))

knitr::kable(df, floating.environment="sidewaystable")
```

	X1	X2	X3	X4	X5
5%	146.5911	141.9087	164.3589	133.9336	153.6533
94%	162.6974	158.3250	181.0770	150.1613	169.9411

Question 2 and 3

Model the relationship between height (cm) and the natural logarithm of weight (log-kg): log(weight). Use the entire Howell1 data frame, all 544 rows, adults and non-adults. Use any model type from Chapter 4 that you think useful: an ordinary linear regression, a polynomial or a spline. Plot the posterior predictions against the raw data.

Plot the prior predictive distribution for the polynomial regression model in Chapter 4. You can modify the the code that plots the linear regression prior predictive distribution. 20 or 30 parabolas from the prior should suffice to show where the prior probability resides. Can you modify the prior distributions of , 1, and 2 so that the prior predictions stay within the biologically reasonable outcome space? That is to say: Do not try to fit the data by hand. But do try to keep the curves consistent with what you know about height and weight, before seeing these exact data

Lets assume the relationship between weight and height is weight = height^3

```
d3 <- as.data.frame(scale(d))
d3$height <- d3$height - min(d3$height)
```

```
N <- 100
a <- rnorm(N, -2, 0.1)
b <- rnorm(N, 0.1, 0.1)
c <- rnorm(N, 0.05, 0.05)
e <- rnorm(N, 0.05, 0.05)

plot(weight~height, data=d3, ylim=range(d3$weight), xlim=range(d3$height), ylab="weight", xlab="height")

x_min <- min(d3$height)

for(i in 1:N) curve(a[i] + + b[i] * x + c[i] * x^2 + e[i] * x^3, from=min(d3$height), to=max(d3$height)
```

3
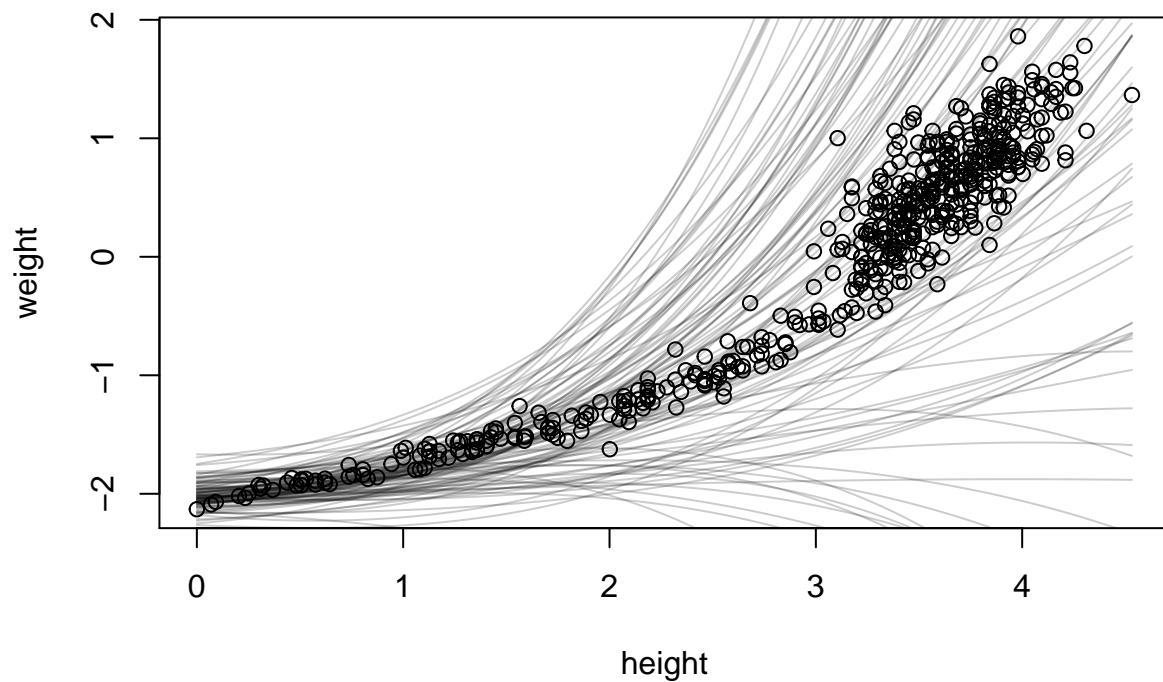
```
flist <- alist(
  weight ~ dnorm(mu, sigma),
  mu ~ a + b *  height + c * height ^ 2 + e * height ^ 3,
  a ~ dnorm(-2, 1),
  b ~ dnorm(0.1, 0.1),
  c ~ dnorm(0.05, 0.05),
  e ~ dnorm(0.05, 0.05),
  sigma ~ dunif(0, 100)
)

homework_model_2 <- quap(flist, data=d3)
```
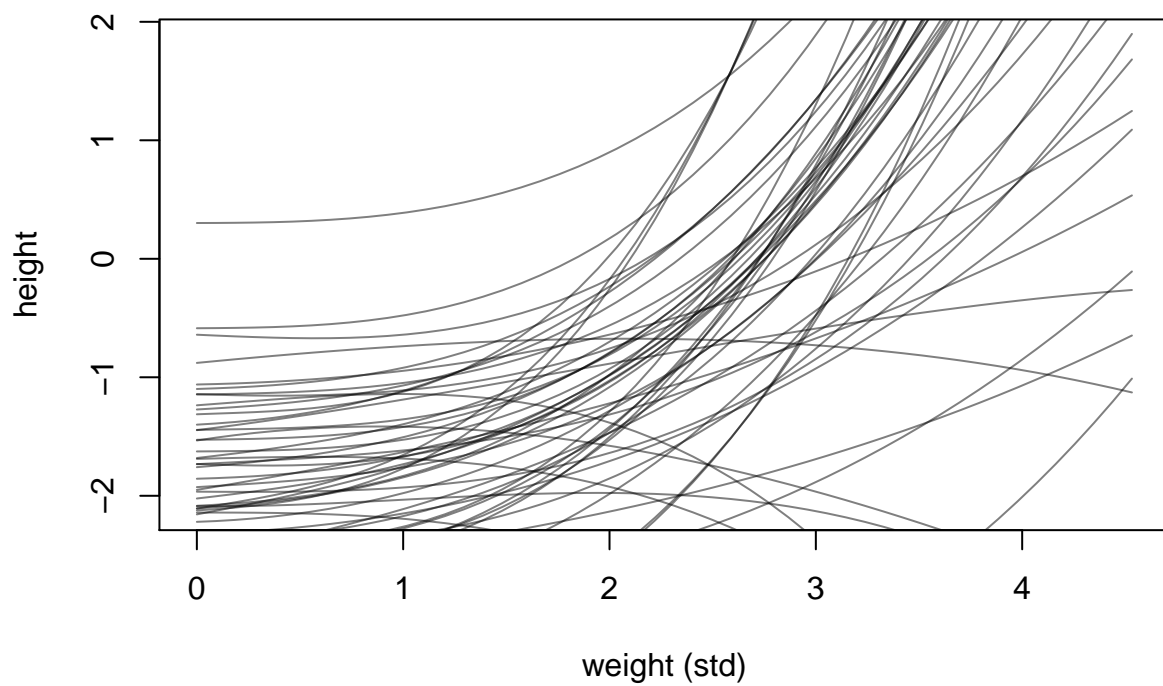
```
prior <- extract.prior( homework_model_2 )

w_seq <- seq( from=min(d3$height) , to=max(d3$height), length.out=50 )

mu <- link( homework_model_2 , post=prior, data=list( height= w_seq))

plot( NULL , xlim=range(w_seq) , ylim=range(d3$weight), xlab="weight (std)" , ylab="height" )
for ( i in 1:50 ) lines( w_seq , mu[i,] , col=col.alpha("black",0.5) )
```
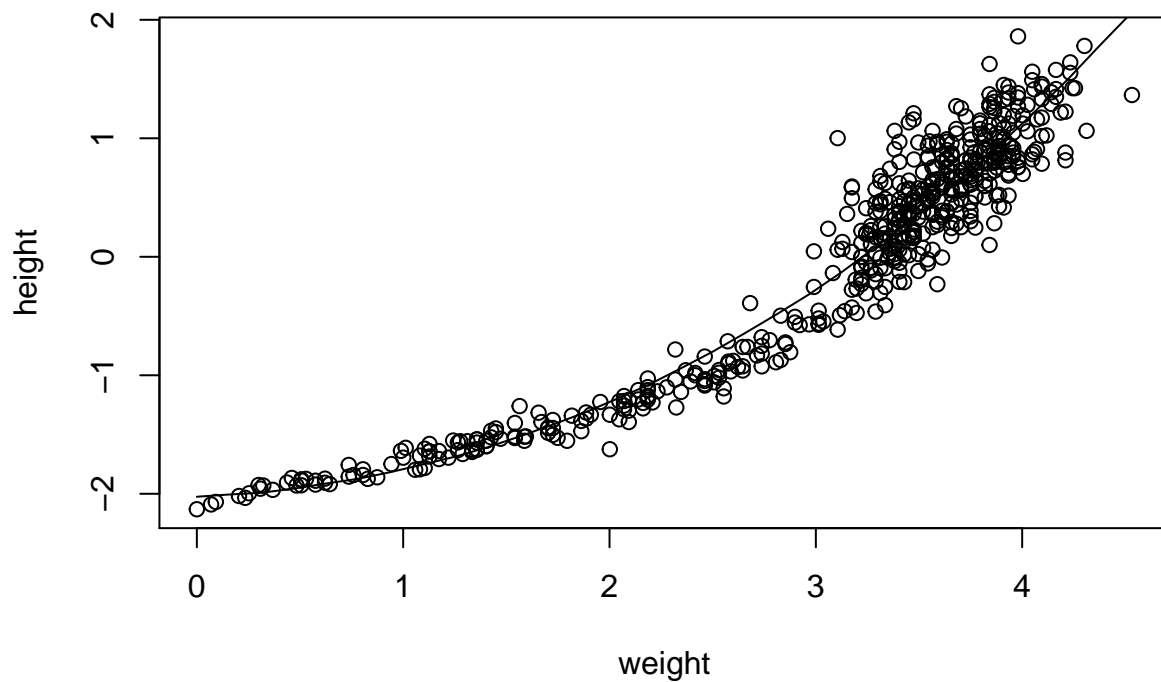
```
height_seq <- seq(min(d3$height), max(d3$height), length.out=30)
simulated <- sim(homework_model_2, data=data.frame(height=height_seq), n=1e4)

height.PI <- apply(simulated, 2, mean)
plot(weight~height, data=d3,  xlab="weight", ylab="height")
lines(height_seq, height.PI)
```

```r
d$log_weight <- log(d$weight)

m2 <- quap(
alist(
height ~ dnorm( mu , sigma ) ,
mu <- a + b* log_weight ,
a ~ dnorm( 178 , 20 ) ,
b ~ dlnorm( 0 , 1 ) ,
sigma ~ dunif( 0 , 50 )
) ,
data=d )
plot( d$weight , d$height , col=col.alpha(rangi2,0.7) )
x_seq <- log(1:60)
mu <- sim( m2 , data=list(log_weight=x_seq) )
mu_mean <- apply(mu,2,mean)
mu_ci <- apply(mu,2,PI,0.99)
lines( exp(x_seq) , mu_mean )
shade( mu_ci , exp(x_seq) )
```
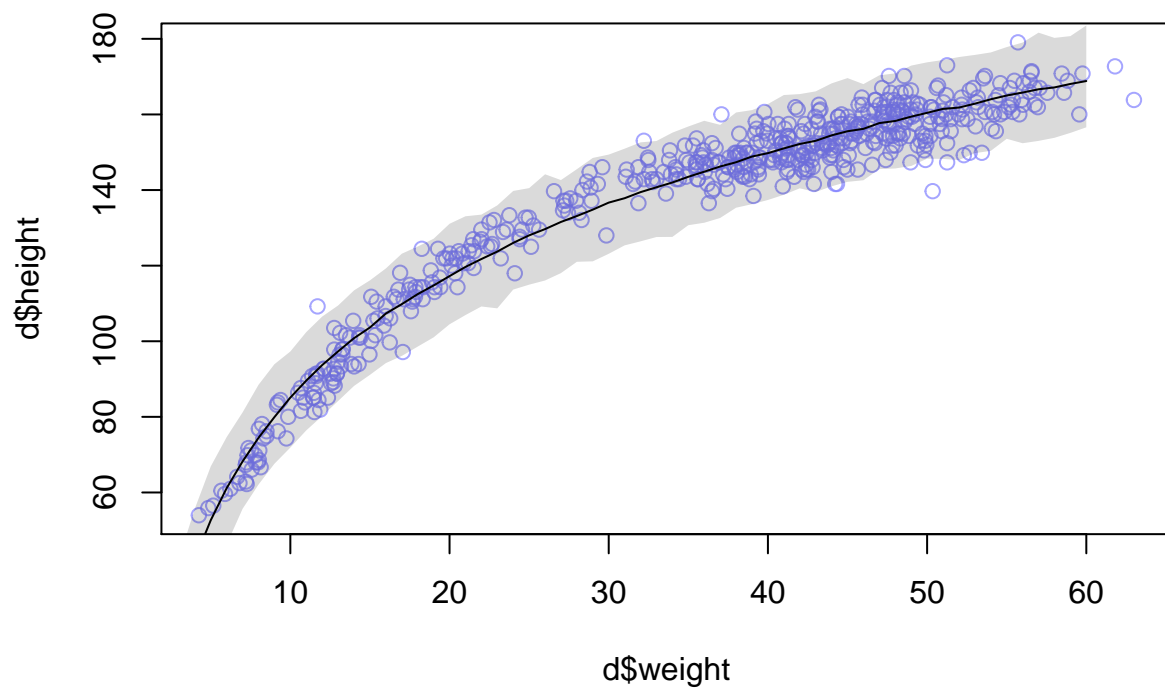
```
posterior <- extract.samples( m2 )

w_seq <- seq( from=min(d$weight) , to=max(d$weight), length.out=50 )

mu <- link( m2, post=prior, data=list( weight= w_seq, log_weight=log(w_seq)))

plot( NULL , xlim=range(w_seq) , ylim=range(d$height), xlab="weight (std)" , ylab="height" )
for ( i in 1:50 ) lines( w_seq , mu[i,] , col=col.alpha("black",0.5) )
```