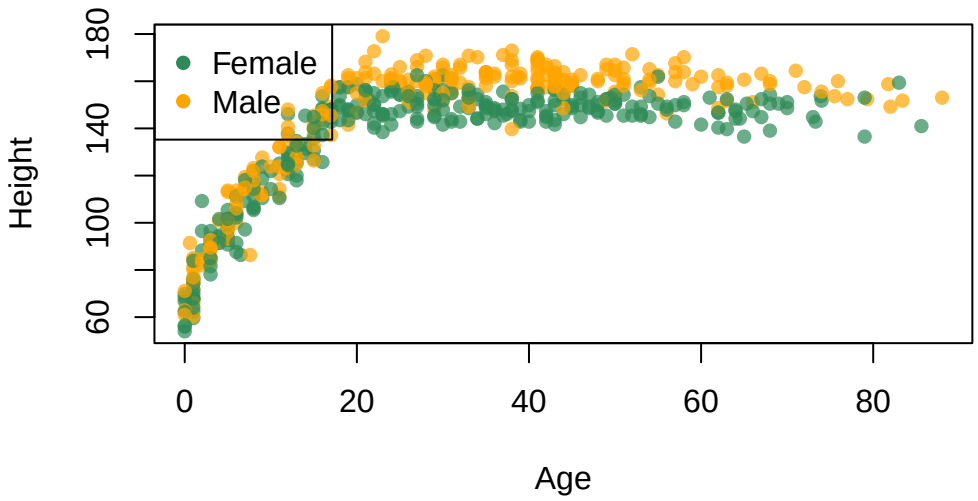
SCI 2025: Homework 6

Setup

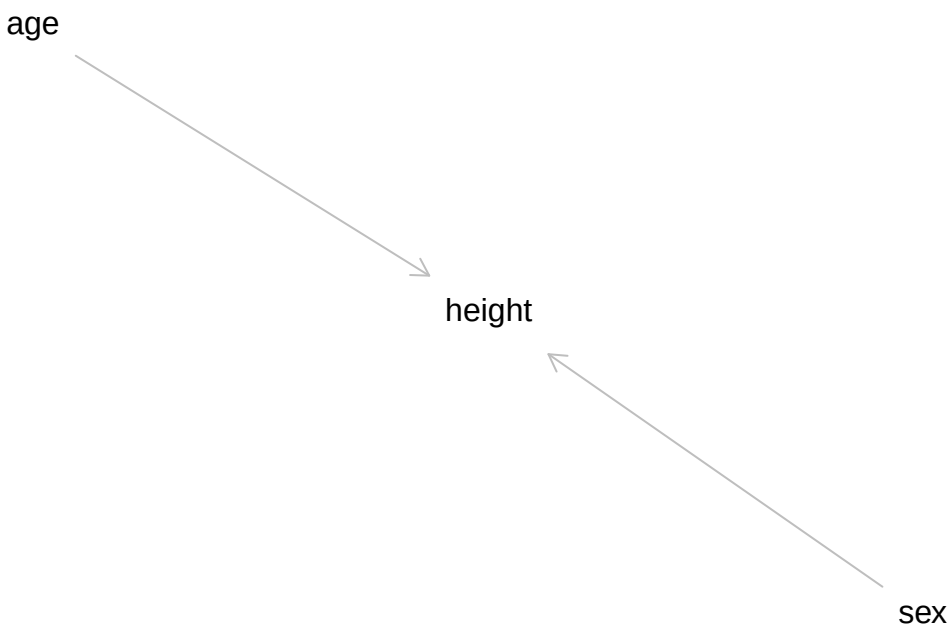
In this homework, we'll revisit the Howell data from Chapter 4, this time with an emphasis on interactions.

```
library(rethinking)
data(Howell1)

d <- Howell1

plot(d$height ~ d$age, col = ifelse(d$male == 0, col.alpha("seagreen", 0.7),
 col.alpha("orange", 0.7)),
     pch = 16,  # solid circle points
     xlab = "Age", ylab = "Height",
     main = "")
legend("topleft", legend = c("Female", "Male"),
       col = c("seagreen", "orange"), pch = 16)
```



It seems that there is an interaction between age and sex (male/female) in determining height. Let's fit a model to this data, assuming the following DAG:

```
library(dagitty)
dag <- dagitty(
  "dag {
    age -> height
    sex -> height
    }"
)

plot(dag)
```



Question 1:

First, fit a model that predicts height using only age, and not sex. You may use a linear or non-linear function, and you may wish to perform some data transformations such as standardization. As always, be thoughtful about your priors.

After fitting the model, plot the residuals as a function of age. Indicate which residuals are males and which are females, either by color, symbol, or some other visual indicator. What do you notice?

```
mean(d$height)
```

```
[1] 138.2636
```

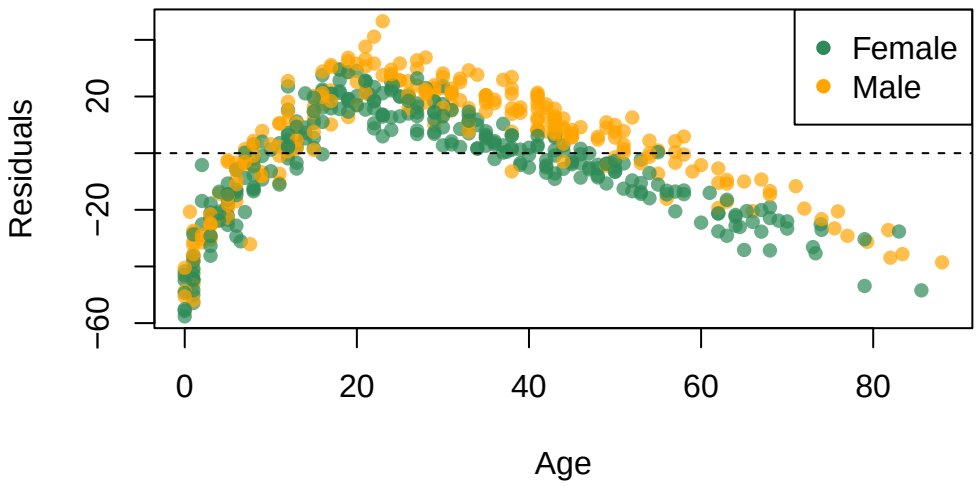
```
sd(d$height)
```

```
[1] 27.60245
```

```r
d$age_c <- (d$age - mean(d$age))

model1 <- quap(
  alist(
    height ~ dnorm(mu, sigma),
    mu <- a + b*age_c,
    a ~ dnorm(138, 40),
    b ~ dnorm(0, 5),
    sigma ~ dexp(0.1)
  ), data = d
)

precis(model1)
```

```
              mean         sd        5.5%       94.5%
a     138.2631048 0.86105518 136.8869724 139.6392373
b       0.9095424 0.04154928   0.8431386   0.9759462
sigma  20.0877540 0.60731006  19.1171552  21.0583528
```

```r
epred_model1 <- link(model1)
resid <- d$height - apply(epred_model1, 2, mean)

plot(resid ~ d$age, col = ifelse(d$male == 0, col.alpha("seagreen", 0.7),
 col.alpha("orange", 0.7)),
     pch = 16,
     xlab = "Age", ylab = "Residuals")
abline(h = 0, lty = 2)
legend("topright", legend = c("Female", "Male"),
       col = c("seagreen", "orange"), pch = 16)
```



Question 2:

Now, fit a model that predicts height using both age and sex. You should also include an interaction between age and sex. Once again, plot the residuals as a function of age with an indicator of sex. What do you notice?

You should also plot the predicted height as a function of age for males and females separately.

```
model2 <- quap(
  alist(
    height ~ dnorm(mu, sigma),
    mu <- a + b*age_c + b_male*male + b_age_male*age_c*male,
    a ~ dnorm(138, 40),
    b ~ dnorm(0, 2),
    b_male ~ dnorm(0, 3),
    b_age_male ~ dnorm(0, 0.5),
    sigma ~ dexp(0.1)
```

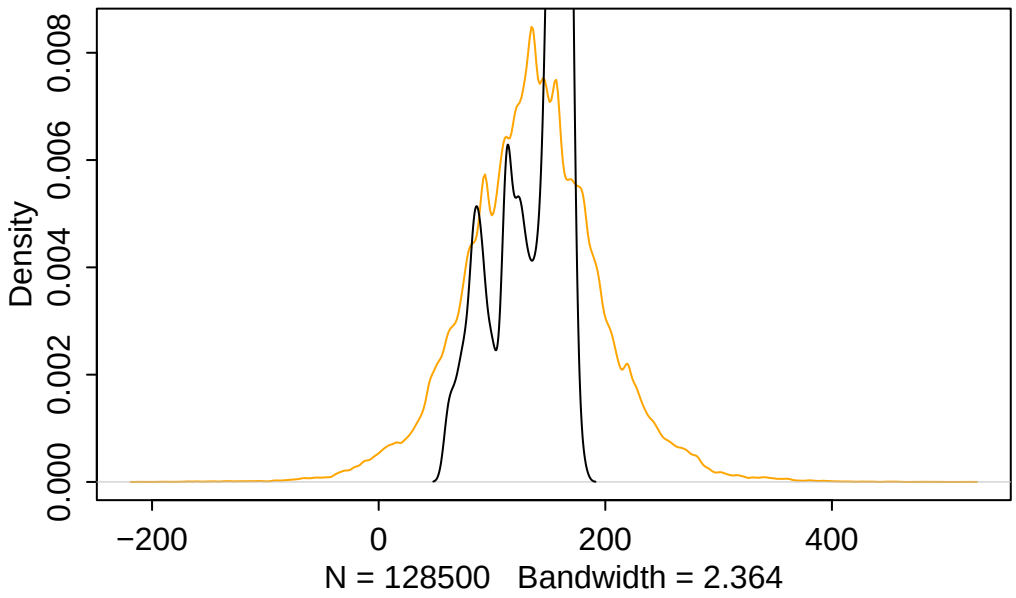
```
  ), data = d
)

precis(model2)
```

	mean	sd	5.5%	94.5%
a	135.5794061	1.09461875	133.82999394	137.3288183
b	0.8355574	0.05586826	0.74626912	0.9248457
b_male	5.6641948	1.47548082	3.30609151	8.0222982
b_age_male	0.1533913	0.08049849	0.02473918	0.2820435
sigma	19.6948593	0.59641487	18.74167309	20.6480454

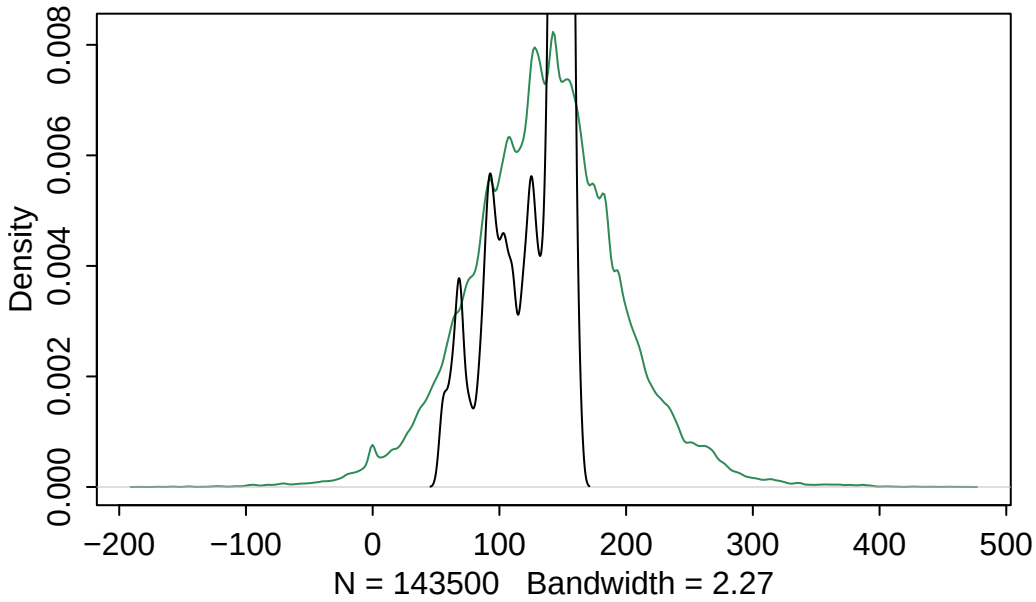
```
prior <- extract.prior(model2)

prior_mu <- sapply(1:500, function(i) prior$a[i] + prior$b[i]*d$age_c + prior$b_male[i]*

str(prior_mu)
```

```
 num [1:544, 1:500] 192 210 216 138 175 ...
```

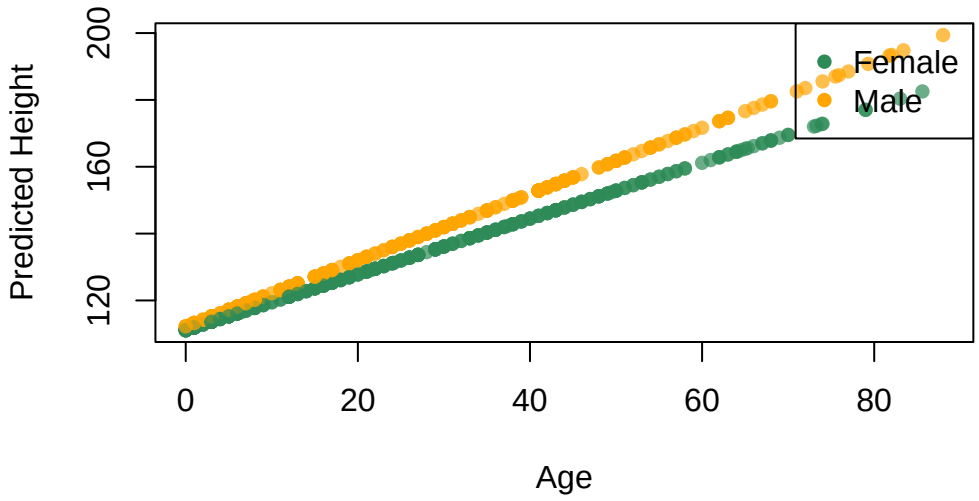
```
# prior predictive check
dens(prior_mu[d$male == 1, ], col = "orange")
dens(d$height[d$male == 1], add = TRUE, col = "black")
```




```
dens(prior_mu[d$male == 0, ], col = "seagreen")
dens(d$height[d$male == 0], add = TRUE, col = "black")
```

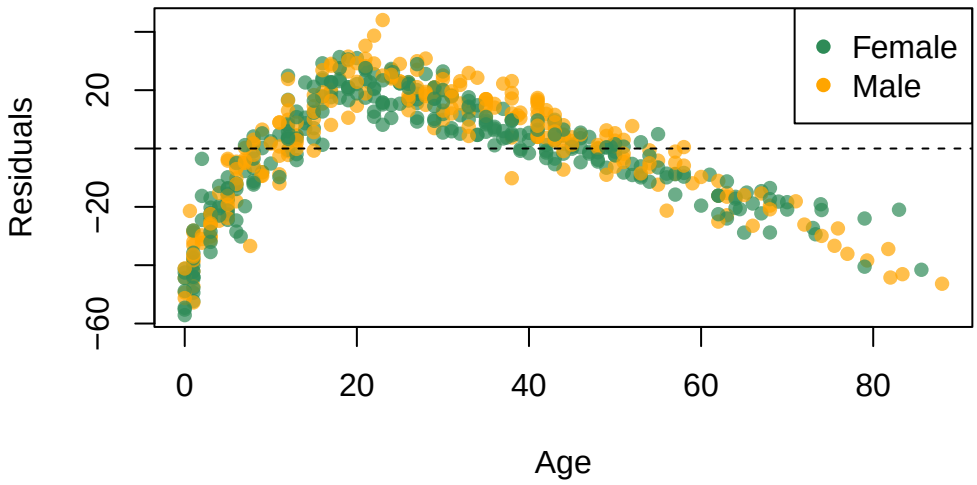


```r
epred_model2 <- link(model2)
resid <- d$height - apply(epred_model2, 2, mean)


plot(apply(epred_model2, 2, mean) ~ d$age, col = ifelse(d$male == 0, col.alpha("seagreen
 col.alpha("orange", 0.7)),
     pch = 16,
     xlab = "Age", ylab = "Predicted Height")
abline(h = 0, lty = 2)
legend("topright", legend = c("Female", "Male"),
       col = c("seagreen", "orange"), pch = 16)
```




```r
plot(resid ~ d$age, col = ifelse(d$male == 0, col.alpha("seagreen", 0.7),
 col.alpha("orange", 0.7)),
     pch = 16,
     xlab = "Age", ylab = "Residuals")
abline(h = 0, lty = 2)
legend("topright", legend = c("Female", "Male"),
```

```
       col = c("seagreen", "orange"), pch = 16)
```



Question 3

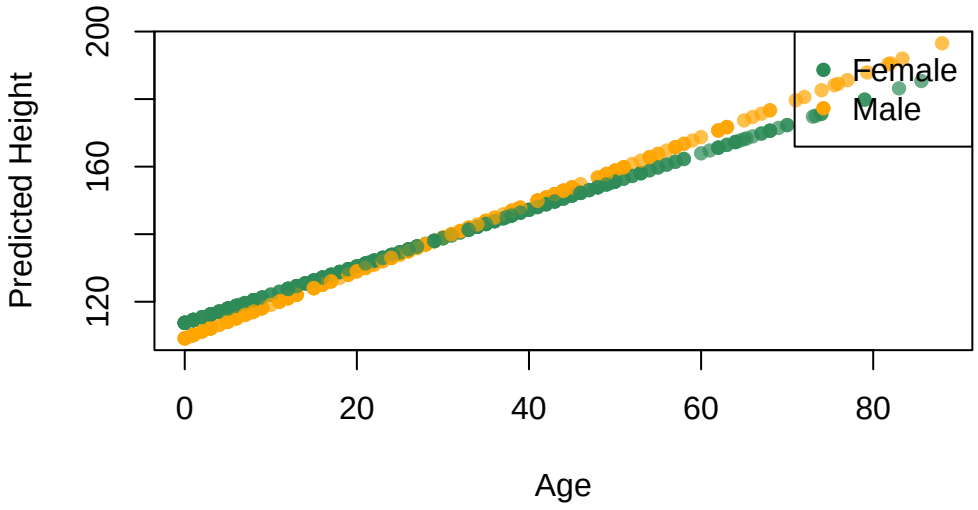
It is standard advice not to include an interaction term in a model unless you also include the "main effects" (i.e., the variables without the interaction term). Try refitting your model from Question 2, retaining the parameters that capture the interaction between age and sex, but removing an main effect of sex. Compare the model predictions to your model from Question 2. What happens? Can you explain it?

```
model3 <- quap(
  alist(
    height ~ dnorm(mu, sigma),
    mu <- a + b*age_c + b_age_male*age_c*male,
    a ~ dnorm(138, 40),
    b ~ dnorm(0, 2),
    b_age_male ~ dnorm(0, 3),
```

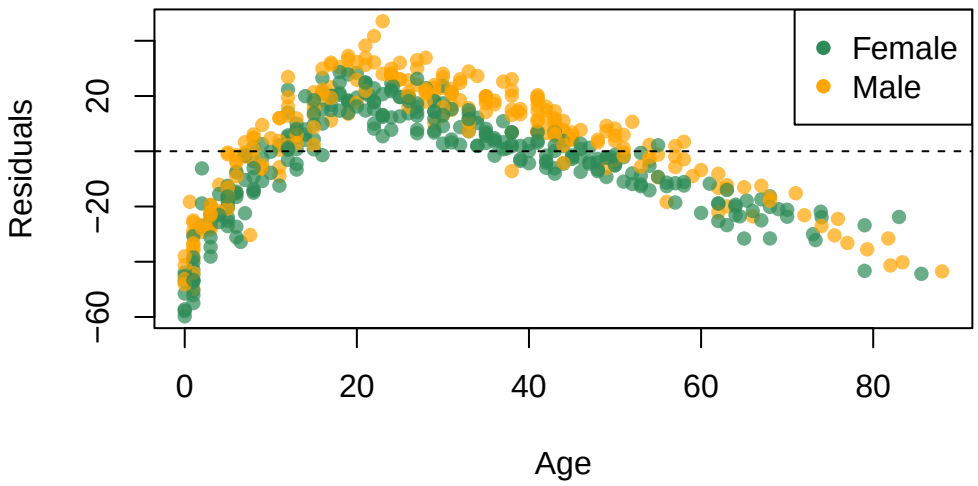
```
    sigma ~ dexp(0.1)
  ), data = d
)


precis(model3)
```

```
                  mean         sd         5.5%       94.5%
a           138.2538667 0.85827256 136.88218140 139.6255521
b             0.8343727 0.05713899   0.74305354   0.9256918
b_age_male    0.1575503 0.08288670   0.02508133   0.2900192
sigma        20.0224517 0.60534884  19.05498734  20.9899161
```

```
epred_model3 <- link(model3)
resid <- d$height - apply(epred_model3, 2, mean)


plot(apply(epred_model3, 2, mean) ~ d$age, col = ifelse(d$male == 0, col.alpha("seagreen
 col.alpha("orange", 0.7)),
     pch = 16,
     xlab = "Age", ylab = "Predicted Height")
abline(h = 0, lty = 2)
legend("topright", legend = c("Female", "Male"),
       col = c("seagreen", "orange"), pch = 16)
```




```
plot(resid ~ d$age, col = ifelse(d$male == 0, col.alpha("seagreen", 0.7),
 col.alpha("orange", 0.7)),
     pch = 16,
     xlab = "Age", ylab = "Residuals")
abline(h = 0, lty = 2)
legend("topright", legend = c("Female", "Male"),
       col = c("seagreen", "orange"), pch = 16)
```

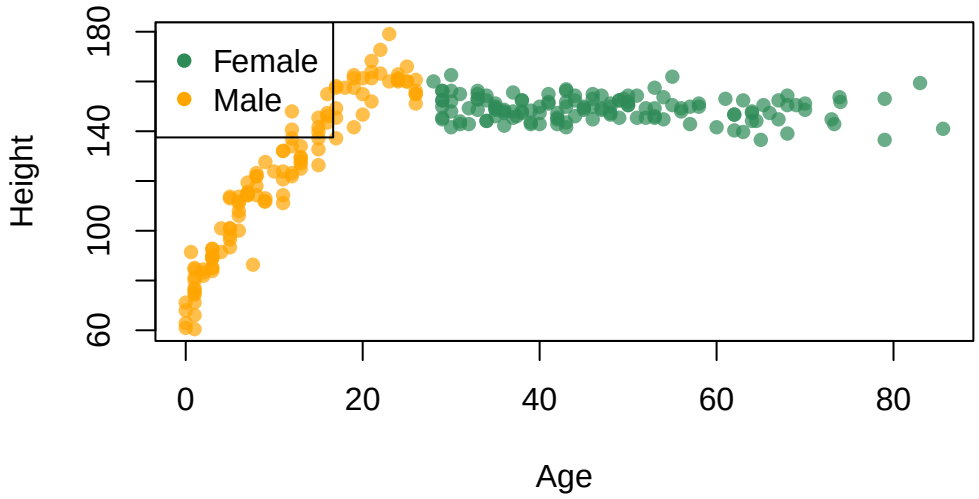


Question 4

An unintuitive aspect of interactions is that they can appear in statistical models as non-linearity in a "main effect"–and vice-versa! To see an example, imagine that there was an imbalance in the ages of males and females in the sample.

```
selection <- ifelse(d$male == 1 & d$age < median(d$age) |
d$male == 0 & d$age > median(d$age), 1, 0)

d2 <- d[selection == 1, ]

plot(d2$height ~ d2$age, col = ifelse(d2$male == 0, col.alpha("seagreen", 0.7),
 col.alpha("orange", 0.7)),
     pch = 16,  # solid circle points
     xlab = "Age", ylab = "Height",
     main = "")
legend("topleft", legend = c("Female", "Male"),
```

```
        col = c("seagreen", "orange"), pch = 16)
```



Now I want you to fit two models:

(1) A model that predicts height using age and sex, with a *linear* relationship between age and height for both males and females.

```
model4 <- quap(
  alist(
    height ~ dnorm(mu, sigma),
    mu <- a + b*age_c + b_male*male + b_age_male*age_c*male,
    a ~ dnorm(138, 40),
    b ~ dnorm(0, 2),
    b_male ~ dnorm(0, 3),
    b_age_male ~ dnorm(0, 0.5),
    sigma ~ dexp(0.1)
  ), data = d2
)
```

```
precis(model4)
```

```
                  mean         sd        5.5%        94.5%
a          153.2095285 1.12959141 151.4042232 155.01483370
b           -0.1529702 0.05024319  -0.2332686  -0.07267193
b_male      23.4305513 2.00912135  20.2195874  26.64151528
b_age_male   3.2594355 0.10730981   3.0879337   3.43093736
sigma        8.3580608 0.41612308   7.6930158   9.02310587
```

(2) A model that predicts height using age and sex, with a non-linear relationship between age and height for males and females.

```
model5 <- quap(
  alist(
    height ~ dnorm(mu, sigma),
    mu <- a + b*age_c + b2*age_c^2 + b_male*male + b_age_male*age_c*male + b_age_male2*a
    a ~ dnorm(138, 40),
    b ~ dnorm(0, 2),
    b2 ~ dnorm(0, 0.5),
    b_male ~ dnorm(0, 3),
    b_age_male ~ dnorm(0, 0.5),
    b_age_male2 ~ dnorm(0, 0.5),
    sigma ~ dexp(0.1)
  ), data = d2
)

precis(model5)
```

```
                    mean          sd          5.5%         94.5%
a           151.718804577 1.093949158 149.970462538 153.467146617
```

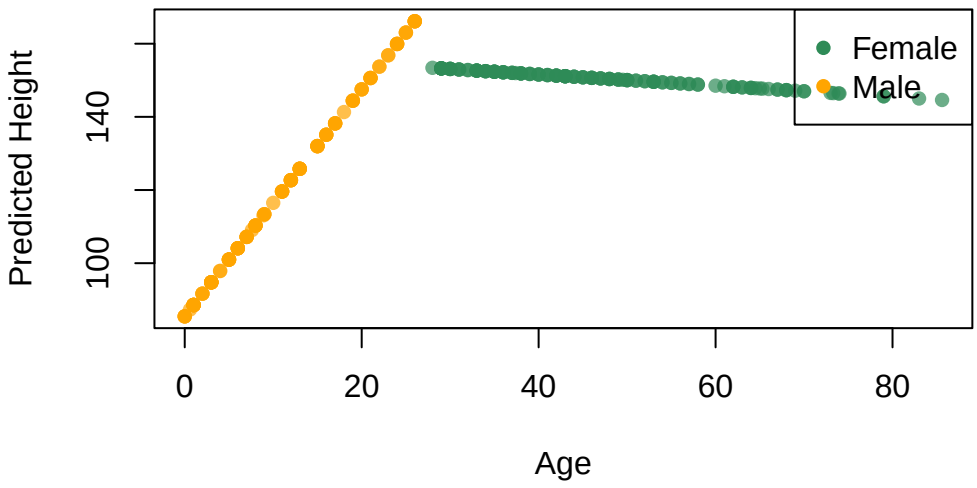
```
b            -0.210693897 0.115602889  -0.395449641  -0.025938153
b2            0.002783481 0.002489482  -0.001195191   0.006762154
b_male        7.373972325 1.997840164   4.181037880  10.566906770
b_age_male   -0.444810906 0.273886970  -0.882535183  -0.007086629
b_age_male2  -0.129073677 0.007960161  -0.141795553  -0.116351802
sigma         6.293414502 0.279320685   5.847006099   6.739822905
```

Compare the summaries of the two models, and compare them using PSIS LOOCV or WAIC. What do you notice? Can you explain it?

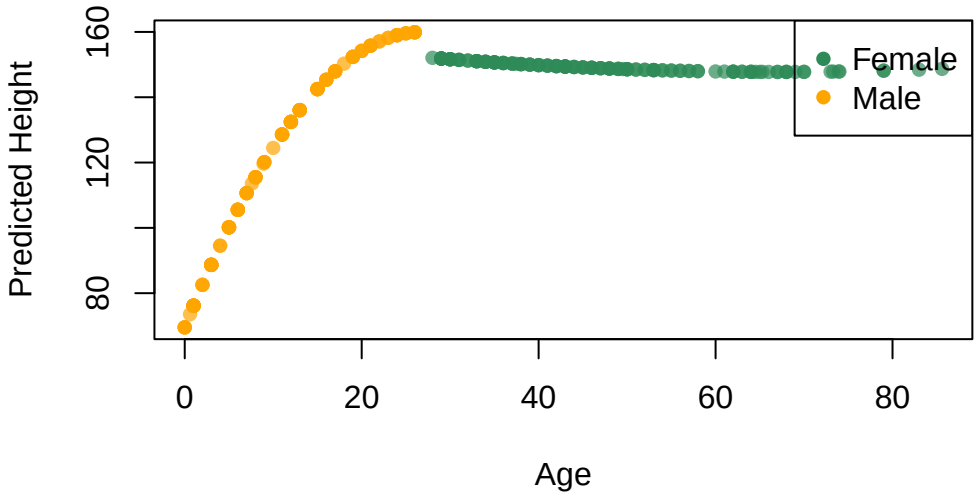
```
compare(model4, model5)
```

```
           WAIC       SE    dWAIC      dSE    pWAIC       weight
model5 1689.556 30.62926   0.0000       NA 7.528043 1.000000e+00
model4 1832.002 27.64254 142.4462 28.36425 5.454455 1.170055e-31
```

```
epred_model4 <- link(model4)
epred_model5 <- link(model5)

plot(apply(epred_model4, 2, mean) ~ d2$age, col = ifelse(d2$male == 0, col.alpha("seagre
 col.alpha("orange", 0.7)),
     pch = 16,
     xlab = "Age", ylab = "Predicted Height")
abline(h = 0, lty = 2)
legend("topright", legend = c("Female", "Male"),
       col = c("seagreen", "orange"), pch = 16)
```




```
plot(apply(epred_model5, 2, mean) ~ d2$age, col = ifelse(d2$male == 0, col.alpha("seagre
 col.alpha("orange", 0.7)),
     pch = 16,
     xlab = "Age", ylab = "Predicted Height")
abline(h = 0, lty = 2)
legend("topright", legend = c("Female", "Male"),
       col = c("seagreen", "orange"), pch = 16)
```



Question 5

An important consequence of interactions (as well as non-linearity) is that there will no longer be a 1:1 mapping between model parameters and the estimand (e.g., the expected increase in height for a one-year increase in age). Instead, we will need to compute "marginal effects".

To get the average marginal effect of age on height, follow these steps, using your fit model from Question 2:

1. For each individual in the sample, compute the model-predicted height at their actual age. I recommend using the `link` function to do this.

```
epred_model2 <- link(model2)
```

2. For each individual in the sample, compute the model-predicted height at their actual age plus one year. You can do this by modifying the data used in step 1, like so:

```r
newdata <- data.frame(age = d$age + 1, male = d$male)
pred_age_plus_one <- link(model, data = newdata)
```

```r
newdata <- data.frame(age_c = d$age_c + 1, male = d$male)
pred_age_plus_one <- link(model2, data = newdata)
```

3. Compute the difference between the predicted heights in step 2 and step 1. This is the marginal effect of age on height for each individual.

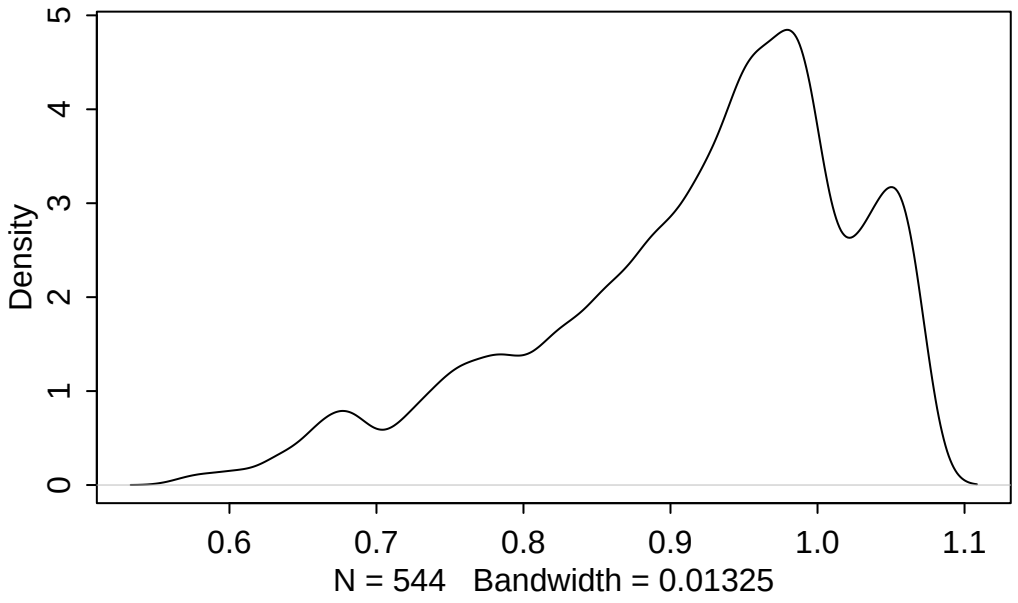
```r
marginal_effect <- pred_age_plus_one - epred_model2
```

4. Take the average of the marginal effects from step 3 across individuals.

```r
AME <- apply(marginal_effect, 2, mean)
```

Finally, summarize the average marginal effect of age on height. How does this estimate compare to the model parameters?

```r
dens(AME)
```



```
precis(AME)
```

```
         mean        sd      5.5%    94.5%   histogram
AME 0.9156713 0.1078439 0.7083266 1.055038
```