EEEB UN3005/GR5005
Lab - Week 06 - 04 and 06 March 2019

Xun Zhao, xz2827

Gaussian Regression Models

This week we'll be working with simulated data that's based on a real ecological research scenario. Tree age (in years) can be evaluated using tree core data, while tree height (in centimeters) can be remotely sensed using LIDAR technology. Given some initial data relating tree age and tree height, a researcher interested in tree population demography might reasonably want to estimate tree age based on tree height, avoiding the time, expense, and labor associated with the field work needed to collect tree core data.

Exercise 1: Fitting a Gaussian Model of Tree Age

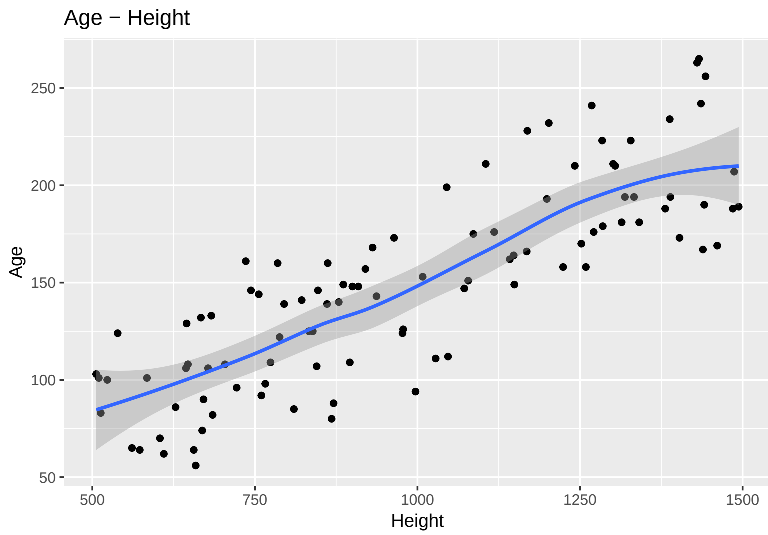
First, import the `simulated_trees.csv` dataset into R. This data contains tree height and age variables. Visualize the tree age variable using a plot type of your choice.

Now, construct a model of tree age using a Gaussian distribution, and use `map()` to fit the model. Assume priors of `dnorm(0, 50)` for the mean parameter and `dcauchy(0, 5)` for the standard deviation parameter. Visualize these priors before you fit your model. To ensure a good model fit, you may need to explicitly define your starting values: 0 for the mean parameter and 50 for the standard deviation parameter.

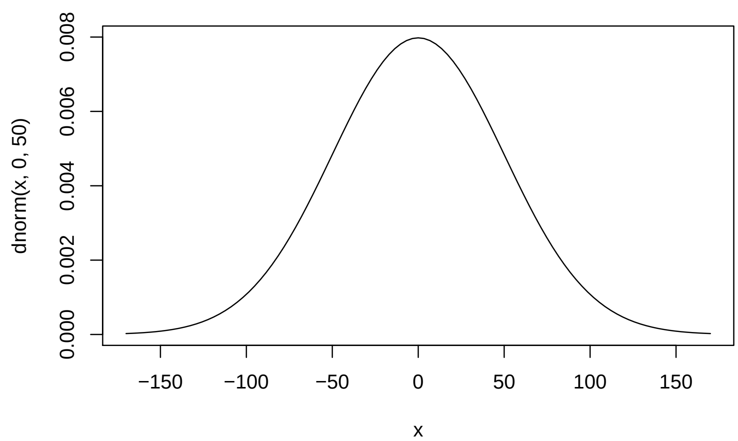
After fitting the model, use `precis()` to display the 99% PIs for all model parameters. How do you interpret the fit model parameters?

```r
d = read.csv('simulated_trees.csv')
ggplot(d, aes(x = height, y = age)) +
    geom_point() +
    geom_smooth() +
    xlab('Height') +
    ylab('Age') +
    ggtitle('Age - Height')
```

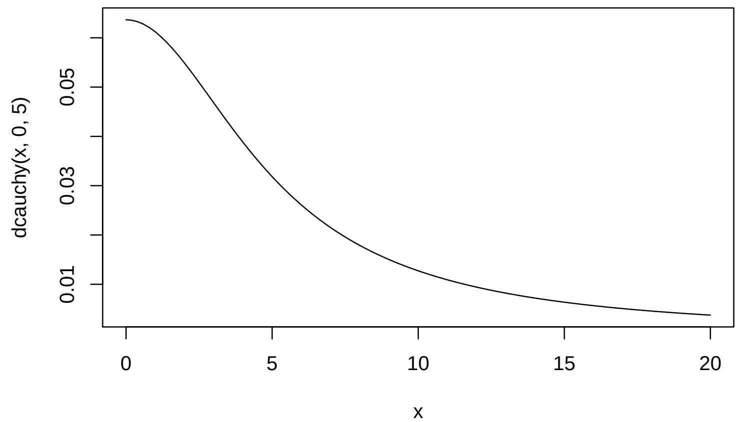
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```




```r
curve(dnorm(x, 0, 50), from = -170, to = 170)
```




```r
curve(dcauchy(x, 0, 5), from = 0, to = 20)
```




```
model = map(
    alist(
        age ~ dnorm(mu, sigma),
        mu ~ dnorm(0, 50),
        sigma ~ dcauchy(0, 5)),
    start = list(mu = 0, sigma = 50),
    data = d)
precis(model, prob = 0.99)
```

```
##          Mean StdDev   0.5%  99.5%
## mu     147.09   4.93 134.39 159.80
## sigma   49.53   3.47  40.59  58.47
```

[**Answer**]

Given a tree, its age is very likely to be between about 135 and 160. And the mean age of all trees is 147, with standard deviation about 5.

Exercise 2: Fitting a Linear Regression of Tree Age

Now, use tree `height` as a predictor of tree `age` in a linear regression model. Assume priors of `dnorm(0, 50)` for the intercept parameter, `dnorm(0, 50)` for the slope parameter, and `dcauchy(0, 5)` for the standard deviation parameter. To ensure a good model fit, you may need to explicitly define your starting values: 0 for the intercept and slope parameters and 50 for the standard deviation parameter. Again, use `map()` to fit the model.

After fitting the model, use `precis()` to summarize the 99% PIs for all model parameters. What does the estimated value of the slope parameter suggest about the relationship between tree height and tree age? What does the intercept parameter correspond to in this model?

```
model2 = map(
    alist(
        age ~ dnorm(mu, sigma),
        mu <- a + b * height,
        a ~ dnorm(0, 50),
        b ~ dnorm(0, 50),
        sigma ~ dnorm(0, 5)),
    start = list(a = 0, b = 0, sigma = 50),
    data = d)
precis(model2, prob = 0.99)
```

```
##        Mean StdDev   0.5% 99.5%
## a      7.86   8.61 -14.32 30.04
## b      0.14   0.01   0.12  0.16
## sigma 24.85   1.44  21.14 28.56
```

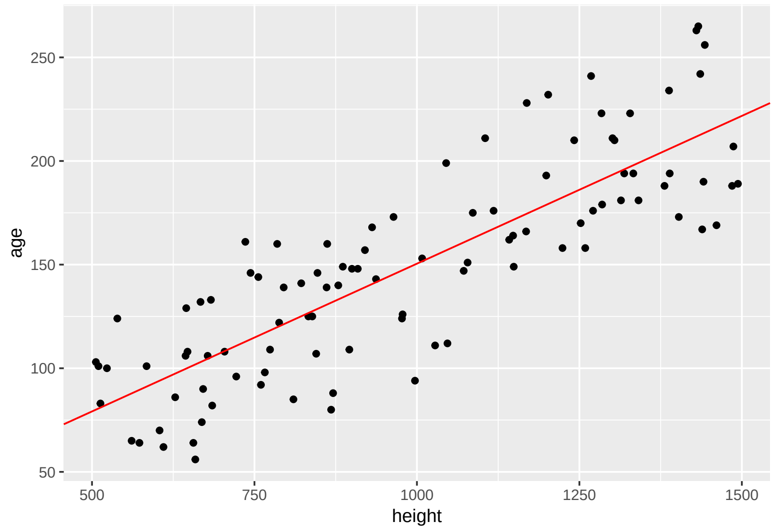
[**Answer**]

The estimated intercept is `a` in the table above, and estimated slope is `b` above. Both of them is distributed as Gaussian distribution, so we can think their `Mean` value is the final estimated value.

Exercise 3: Plotting Raw Data and the *Maximum a Posterior* Trend Line

Using either base R or `ggplot()`, plot tree age versus tree height, and add the *maximum a posterior* line for the estimated mean age at each tree height value. In other words, this line is the relationship between tree height and mean age implied by the *maximum a posterior* intercept and slope estimates from your fit model.

```
graph = ggplot(d, aes(x = height, y = age)) + geom_point()
graph = graph + geom_abline(intercept = coef(model2)["a"], slope = coef(model2)["b"], co
plot(graph)
```



Exercise 4: Making Age Predictions

Generate 10,000 posterior samples from your fit model with `extract.samples()`. Using these samples, make 10,000 age predictions for a tree with a height of 1,200 centimeters. What is the mean value of these age predictions? What is the 50% HPDI of these age predictions? Plot the density of these age predictions.

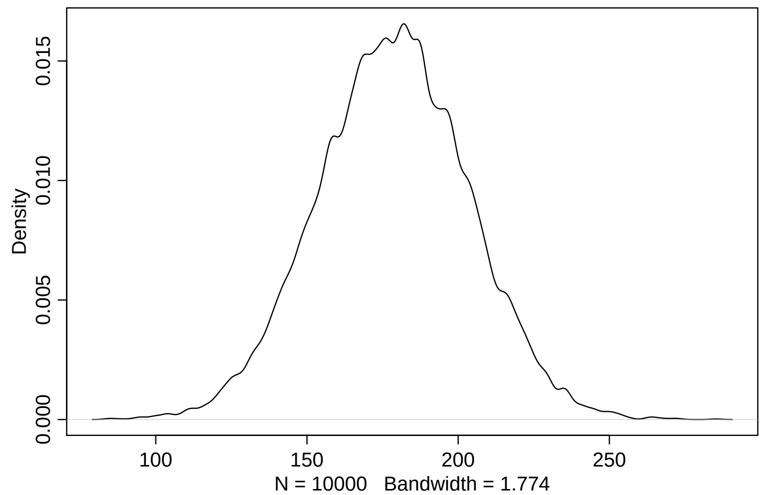
```
samples = extract.samples(model2, 10000)
preds = rnorm(10000, mean = samples$a + samples$b * 1200, sd = samples$sigma)
print(mean(preds))
```

```
## [1] 178.4979
```

```
HPDI(preds, 0.5)
```

```
##      |0.5     0.5|
## 163.6584 196.8909
```

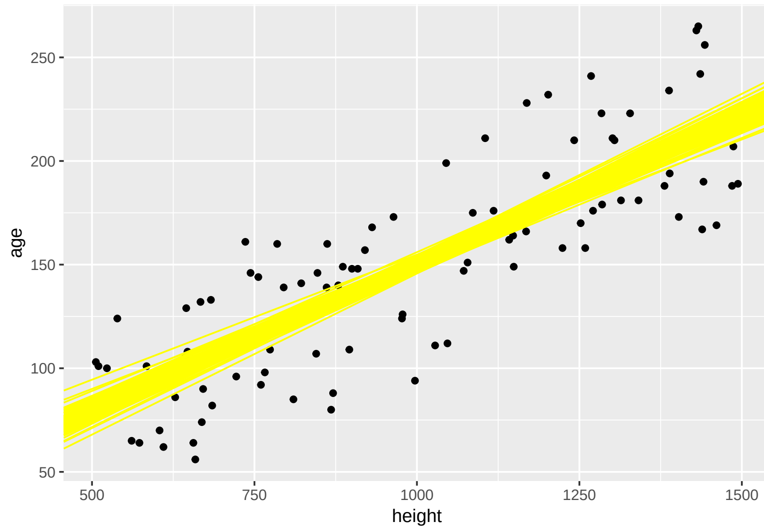
```
dens(preds)
```



Bonus Exercise: Visualizing Uncertainty in the Posterior

Recreate the plot from Exercise 3, but instead of overlaying the *maximum a posterior* trend line, overlay the trend lines implied by the first 100 posterior samples.

```
graph = ggplot(d, aes(x = height, y = age)) + geom_point()
for (i in 1:100){
    graph = graph +
        geom_abline(
            intercept = samples$a[i],
            slope = samples$b[i],
            col = "yellow"
        )
}
plot(graph)
```



age
height
250
200
150
100
50
500
750
1000
1250
1500