Steps in a Bayesian Data Analysis

Bayesian Data Analysis

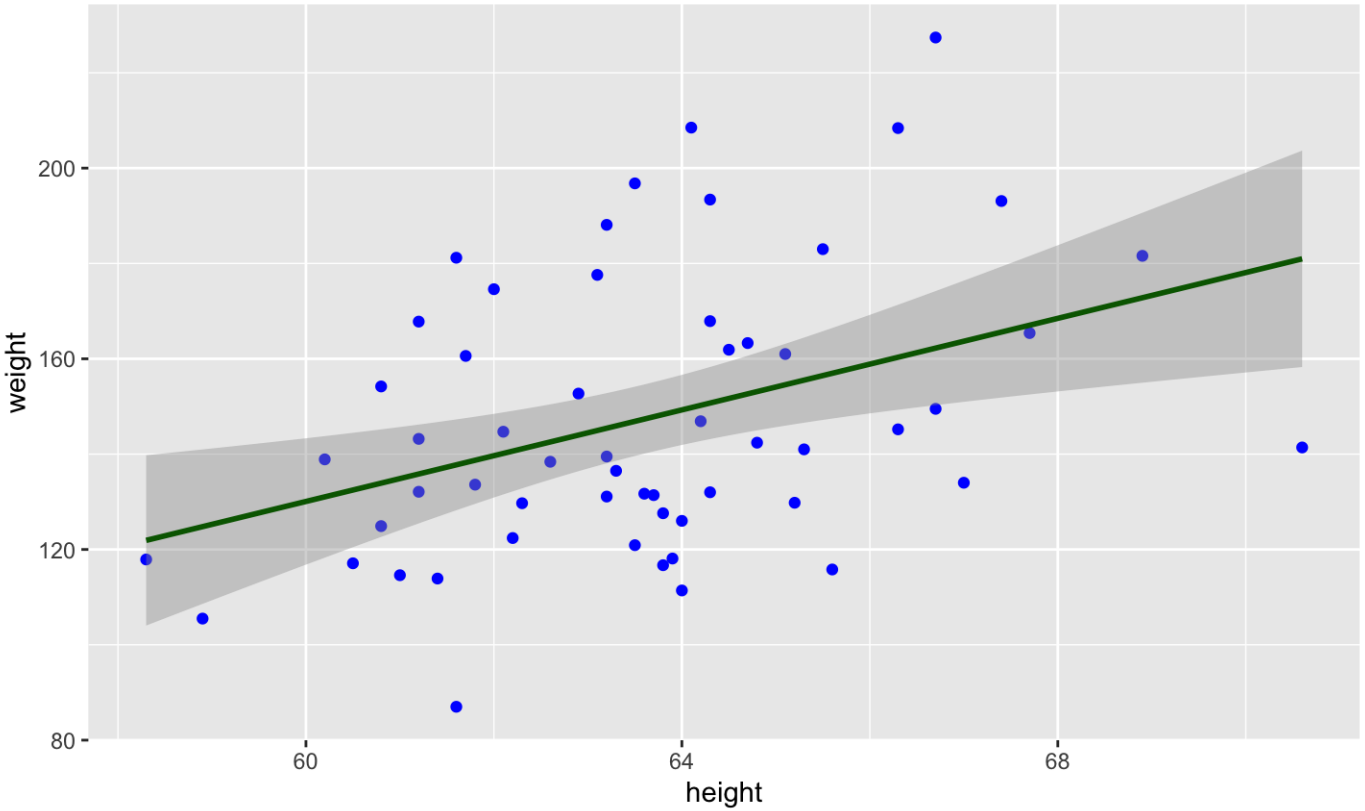
Steve Buyske

Steps in a Bayesian data analysis

1. Identify the relevant data, including which variables are outcomes and which are predictors.
2. Define a descriptive mathematical model.
3. Specify prior distributions on the parameters.
4. Use Bayesian inference to combine the prior and the data to determine the posterior distribution
5. Check that the posterior gives predictions consistent with the data.
6. Pull whatever summary information or conclusions we need from the posterior distribution.

A regression example

The figure below shows height and weight for 57 women, along with a least squares line and a 95% confidence band for the regression line.



1. Our data is a table of 57 pairs of heights and weights, and we will consider weight the outcome and height the predictor.

2. Our mathematical model will be just a linear relationship,

$$\widehat{\text{weight}} = \beta_0 + \beta_1 \text{height},$$

where $\widehat{\text{weight}}$ is the predicted weight and β_0 and β_1 are the unknown intercept and slope.

- We know that weight is not determined entirely by height, so as part of our model we will add that

  $$\text{weight} \sim \text{Normal}(\widehat{\text{weight}}, \sigma),$$

  where the $\sim$ means that weight has a distribution given, in this case, by a normal distribution with mean equal to $\widehat{\text{weight}}$ and standard deviation equal to σ.

3. We have three unknown parameters, β_0, β_1, and σ, so we will need prior distributions for each of them.

The details don't matter right now, so let's just say that we choose as prior distributions

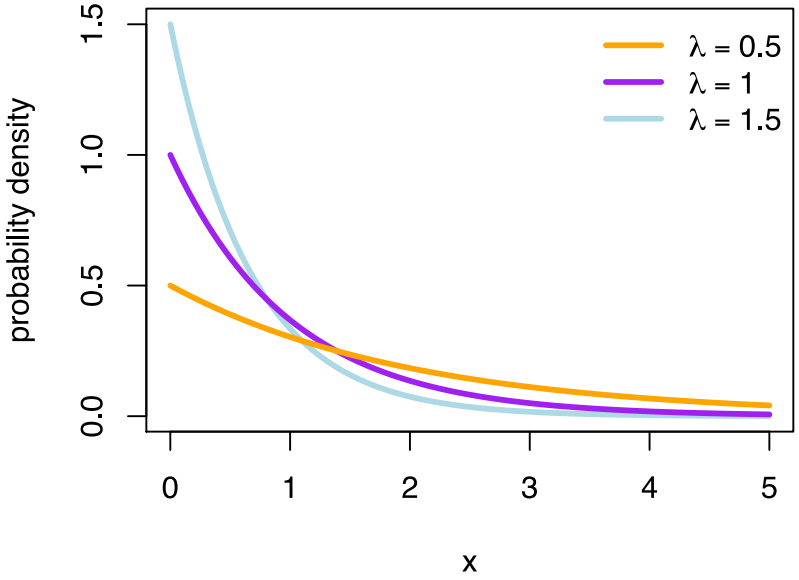
$$\beta_0 \sim \text{Normal}(\overline{\text{weight}}, \text{large number})),$$

$$\beta_1 \sim \text{Normal}(0, \text{another large number})),$$

$$\sigma \sim \text{Exponential}(\text{small number})),$$

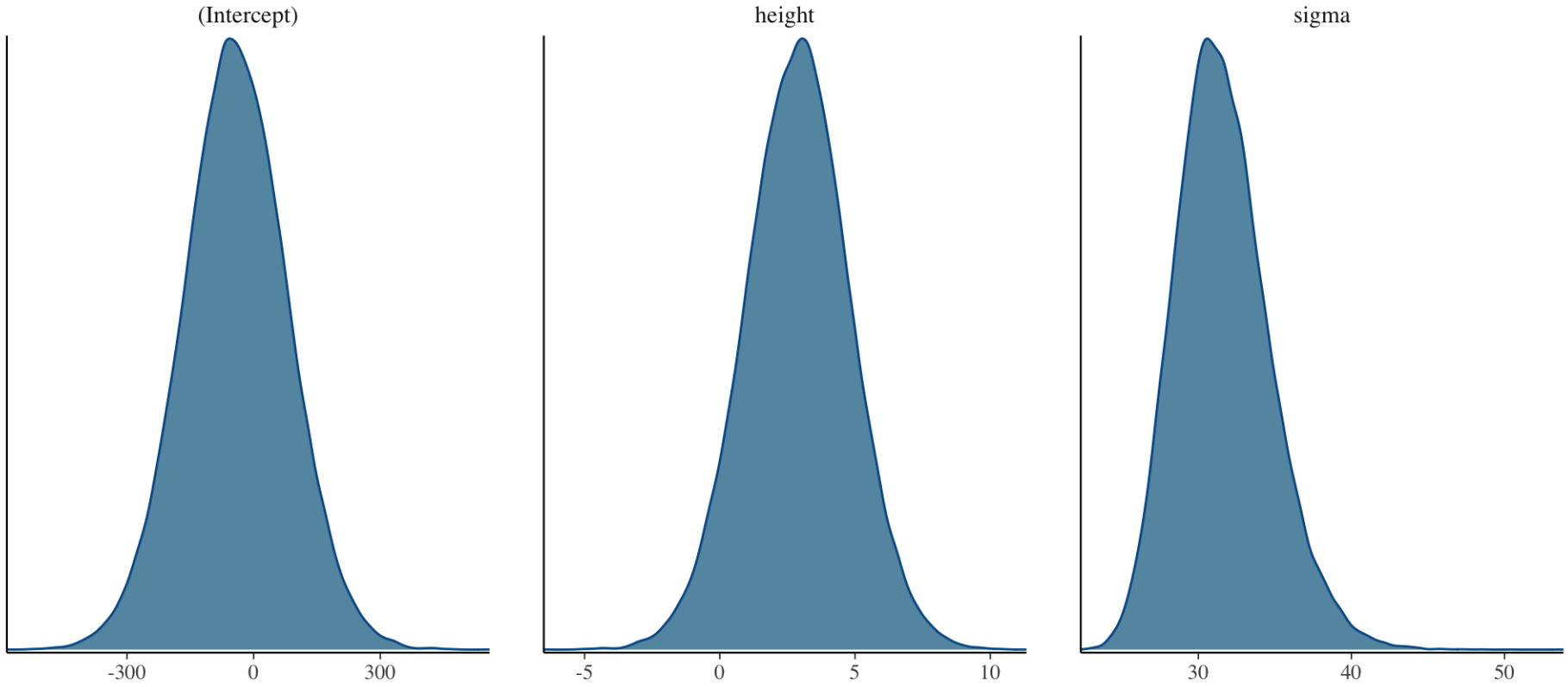
(I used $\overline{\text{weight}}$ for the mean of the prior for β_0 with the idea that if the slope is near zero then the predicted value of weight for a woman of height zero would just be the average height. We'll discuss better approaches later.)

Aside: The exponential distribution looks like this



4. The figure below shows the densities of β_0, β_1, and σ.

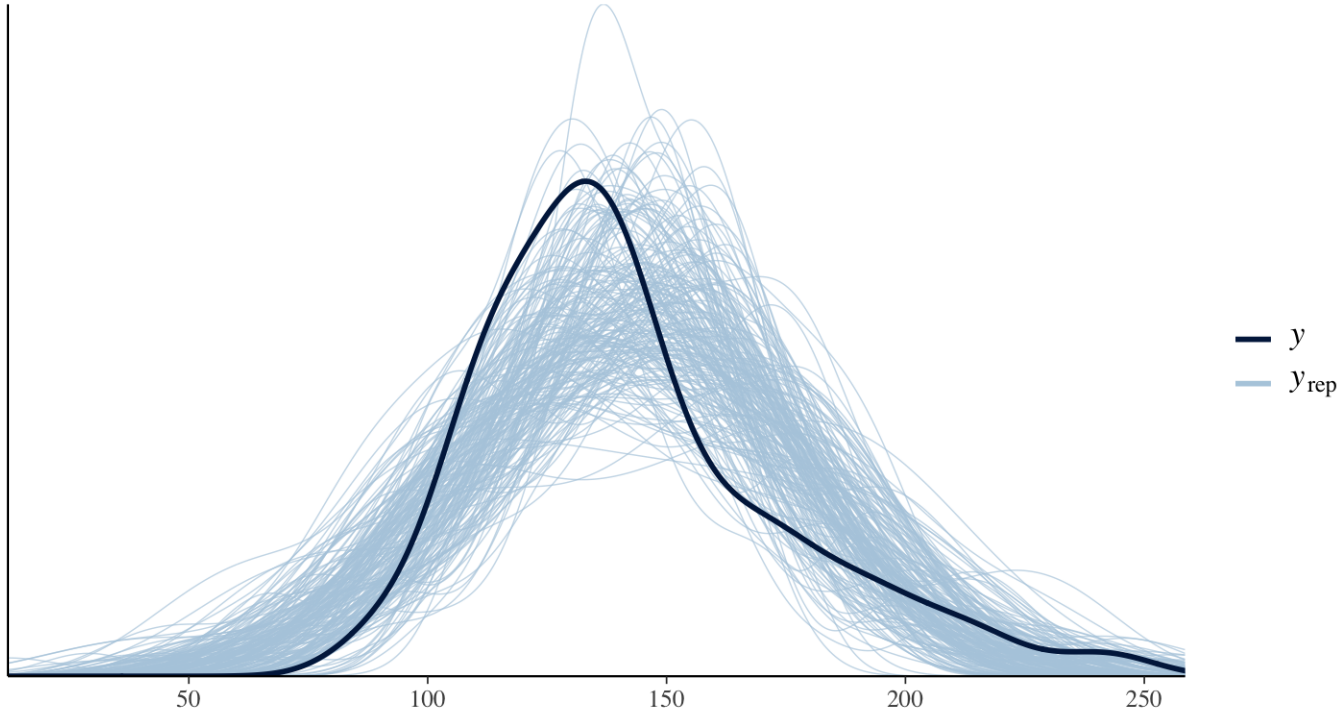
We'll learn how to find these later in the semester.



- Remember that we don’t think of the parameters as fixed but unknown, as in traditional statistics.
- Instead, each of the parameters is a random variable, which is why each of them has a (posterior) distribution, just as all random variables do.

5. If our fit is a good one, when we simulate data based on the model, we should get results that look similar to the actual data. Here's a plot with the empirical density—think of it as a smoothed histogram—of weights, shown in black, is compared to the the empirical densities of 200 simulated datasets in blue.

The fit is okay, but not very good, indicating that our model is not entirely adequate.



6. Finally, ignoring for now the problems in the last slide, let's get 90% credible intervals for β_0, β_1, and σ. "Credible intervals" are the Bayesian analogue of confidence intervals.

	5%	95%
(Intercept)	-243.85	161.38
height	-0.30	6.03
sigma	27.03	37.01

We believe that there is a 0.90 probability that the parameters lie within those intervals.

Steps in a Bayesian data analysis

1. Identify the relevant data, including which variables are outcomes and which are predictors.
2. Define a descriptive mathematical model.
3. Specify prior distributions on the parameters.
4. Use Bayesian inference to combine the prior and the data to determine the posterior distribution
5. Check that the posterior gives predictions consistent with the data.
6. Pull whatever summary information or conclusions we need from the posterior distribution