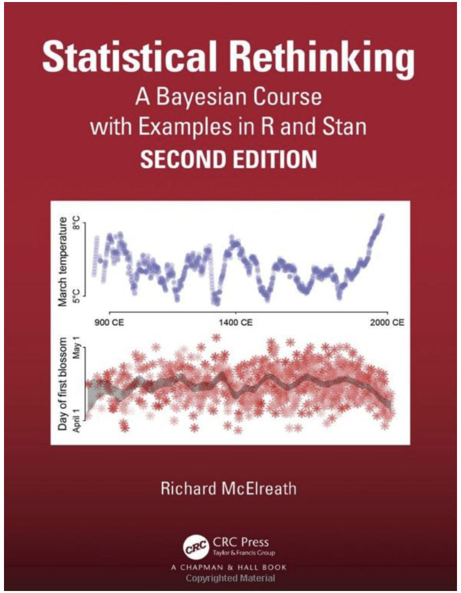
Introduction to Data Science

The Linear Model

Prof. Dr. Ralf Lämmel & M.Sc. **Johannes Härtel**
(johanneshaertel@uni-koblenz.de)



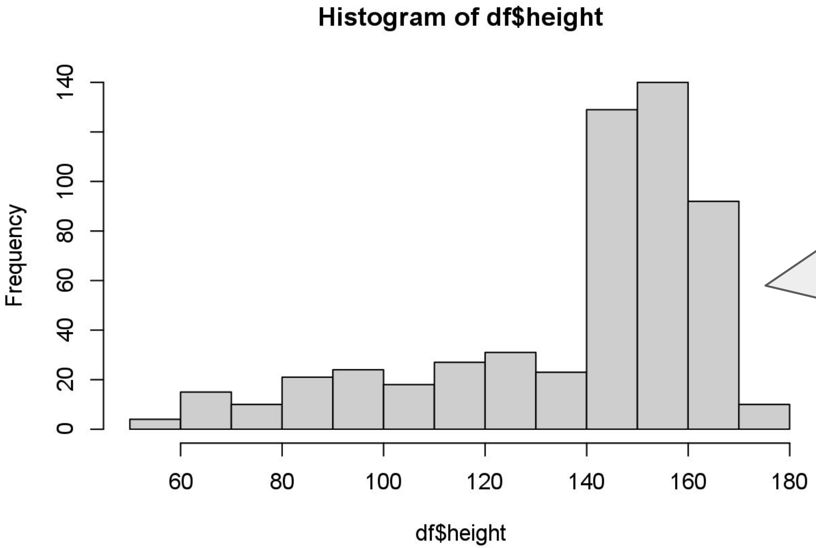
[McElreath20]

The major source for this lecture.

Preprocessing the data

Example

We will again be working with the *!Kung* data set (the height), but we will also consider other columns of the original data set.



This height data does **not seem to be distributed normally.** This is a problem, if we model it as normal distributed.

Inspecting all columns of the original data set

	height	weight	age	male
1	151.765	47.8256	63.0000	1
2	139.700	36.4858	63.0000	0
3	136.525	31.8648	65.0000	0
4	156.845	53.0419	41.0000	1
5	145.415	41.2769	51.0000	0
6	163.830	62.9926	35.0000	1
7	149.225	38.2435	32.0000	0
8	168.910	55.4800	27.0000	1

The data set includes more than just the height.

Do you have **any suggestions what causes the strange distribution** of height?

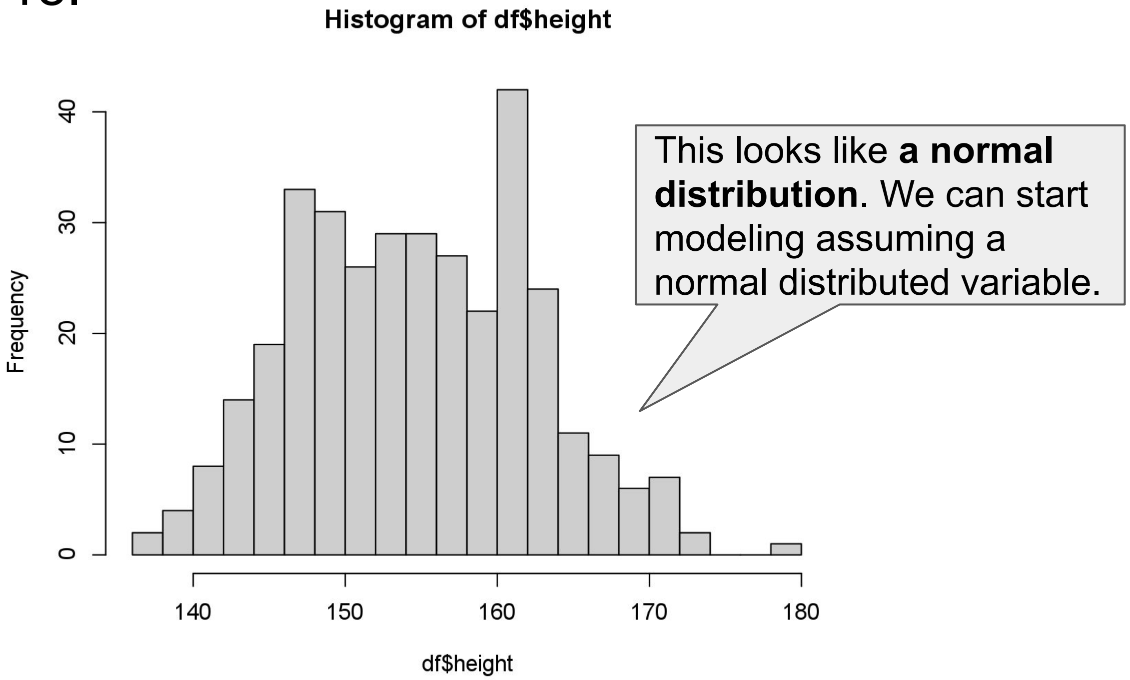
Getting rid of the children

Sorted by age; let's get rid of the children.

height	weight	age	male
69.8500	7.31417	0.00000	0
67.9450	7.82446	0.00000	1
68.5800	8.02291	0.00000	0
66.6750	8.13631	0.00000	0
62.8650	7.20077	0.00000	1
62.2300	7.25747	0.00000	0
55.8800	4.84776	0.00000	0
60.9600	6.23689	0.00000	1

Approximately normal distributed

We face an approximate normal distribution after filtering our entries where the age ≤ 18.



Gaussian/normal model of height

(a brain-dead model)

Definition

We start defining the height (h_i) of the !Kung people, as a normal distributed (observed) variable, with (unobserved) parameter mean (μ, Greek mu) and standard deviation (σ, Greek sigma).

Priors should be defined based on **pre-data knowledge**.

$$h_i \sim \text{Normal}(\mu, \sigma)$$ [likelihood]

$$\mu \sim \text{Normal}(178, 20)$$ [μ prior]

$$\sigma \sim \text{Uniform}(0, 50)$$ [σ prior]

New: Prior predictive simulation

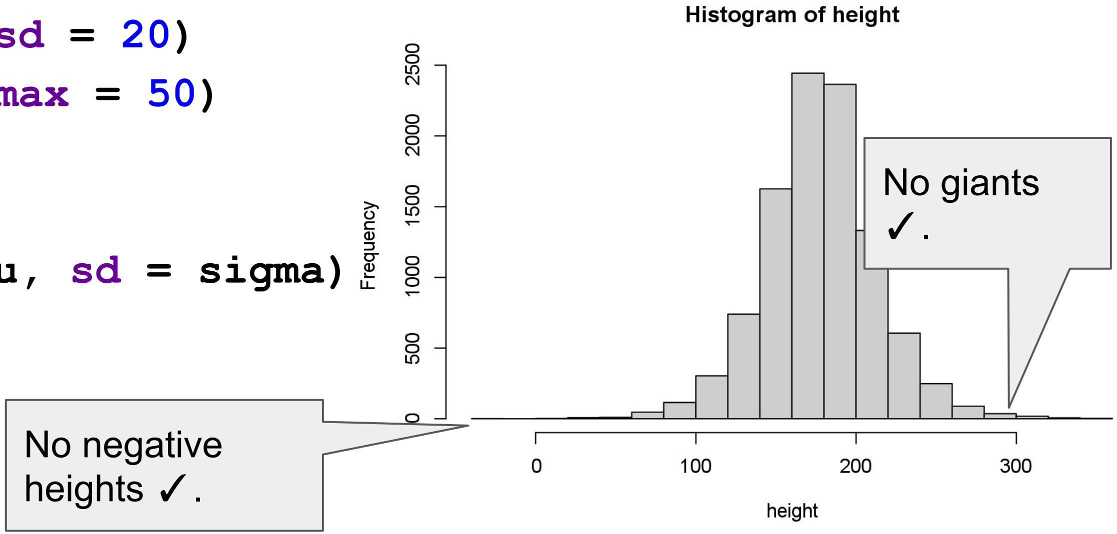
- What does our model think **before it sees the data**?
- **Pre-data knowledge:** We know that there are **no giants or negative heights**, how can we assure this to be impossible in our model?
- We use another kind of simulation, making the **prior assumptions** of the model **explicit**.
  - We simulate possible parameters using the prior.
  - We produce synthetic height data accordingly.

Demo

(prior predictive simulation)

Demo Backup (prior predictive simulation)

```
# Prior predictive simulation
n <- 1e4
# Possible parameters according to the prior.
mu <- rnorm(n, mean = 178, sd = 20)
sigma <- runif(n, min = 0, max = 50)

# Simulated heights.
height <- rnorm(n, mean = mu, sd = sigma)

hist(height)
```



Implementing and running the model on the data

$h_i \sim \text{Normal}(\mu, \sigma)$ [likelihood]

$\mu \sim \text{Normal}(178, 20)$ [μ prior]
$\sigma \sim \text{Uniform}(0, 50)$ [σ prior]

Math

correspondsTo

```r
model <- ulam(alist(
 h ~ dnorm(mu, sigma),
 mu ~ dnorm(178, 20),
 sigma ~ dunif(0, 50)
), data = ...)
```

ULAM

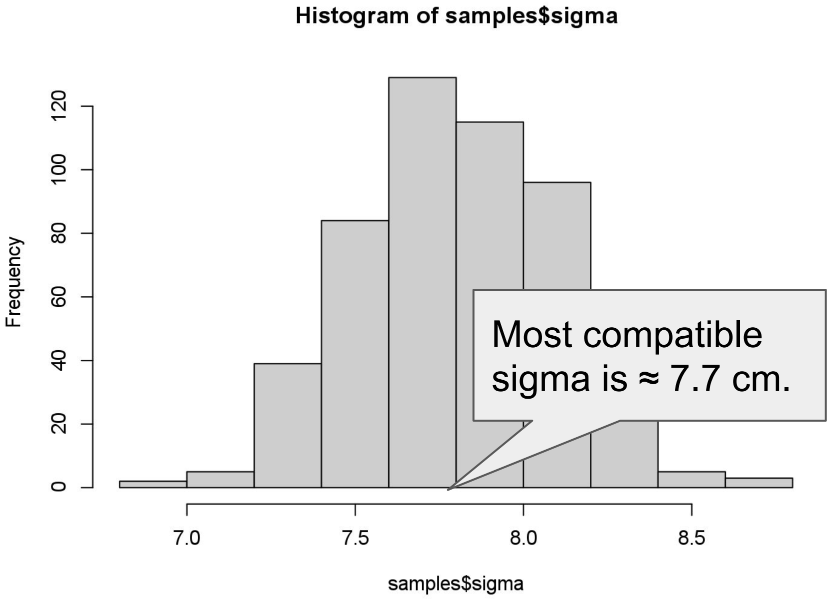
compilesTo

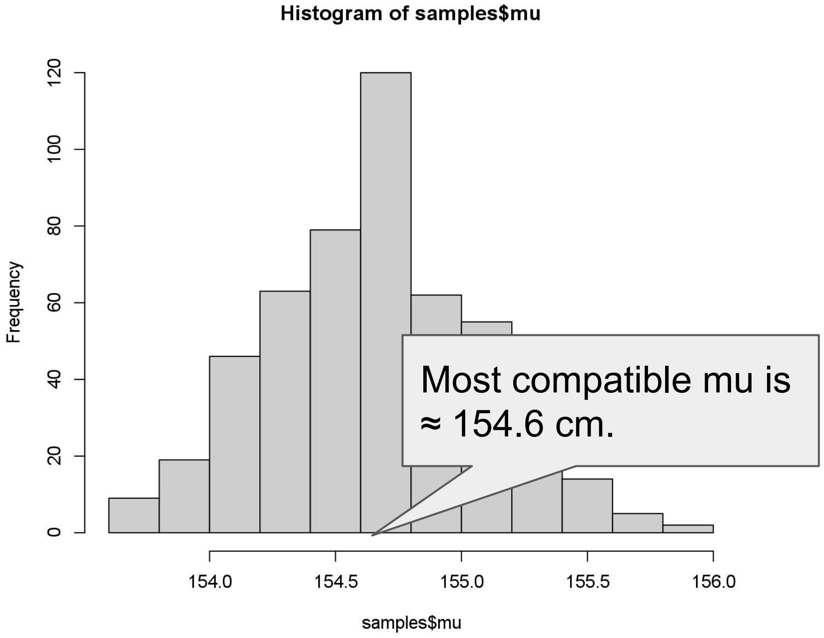
```stan
data{
	vector[346] h;
}
parameters{
	real mu;
	real<lower=0,upper=50> sigma;
}
model{
	sigma ~ uniform( 0 , 50 );
	mu ~ normal( 178 , 20 );
	h ~ normal( mu , sigma );
}
```

STAN

Results: Summarizing the posterior

The results when fitting the model on the real data in terms of the (marginal) posterior for the mean (mu, μ) and standard deviation (sigma, σ) parameters.



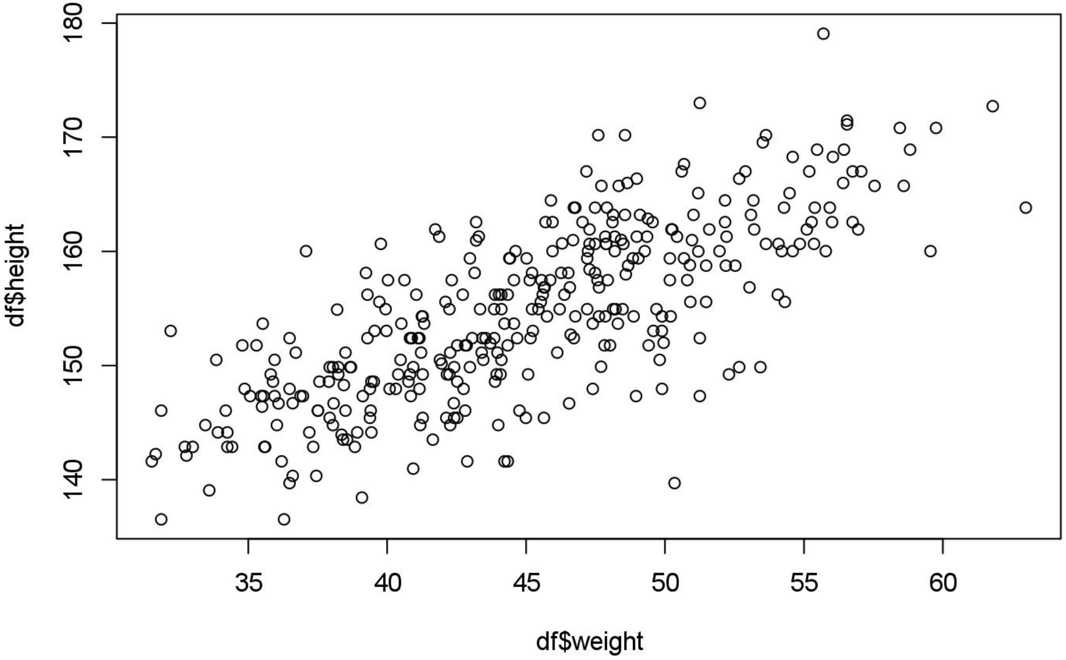


Linear model

How does the height variable relate to other predictor variables?

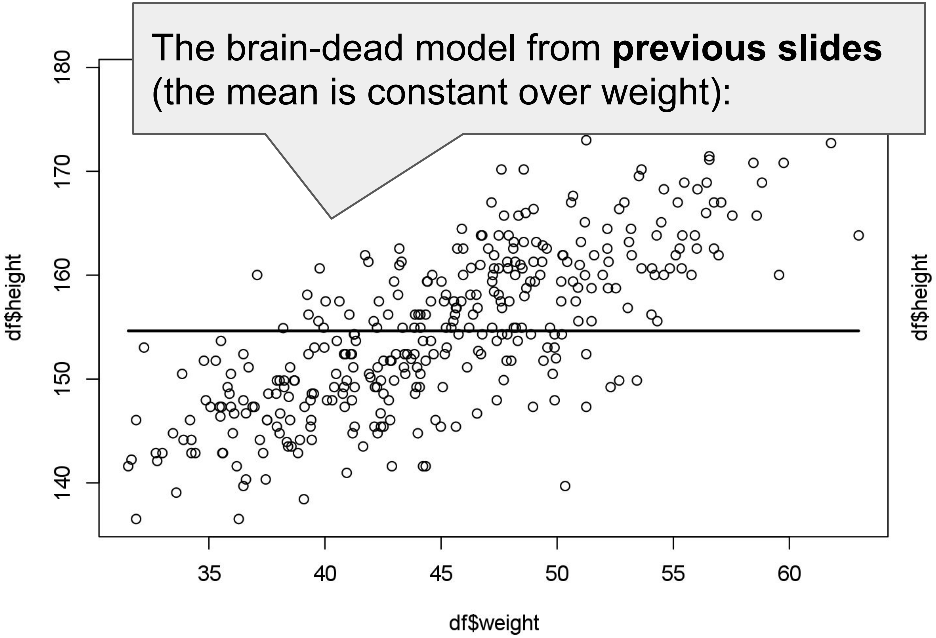
Example: How does **height relate to weight**?

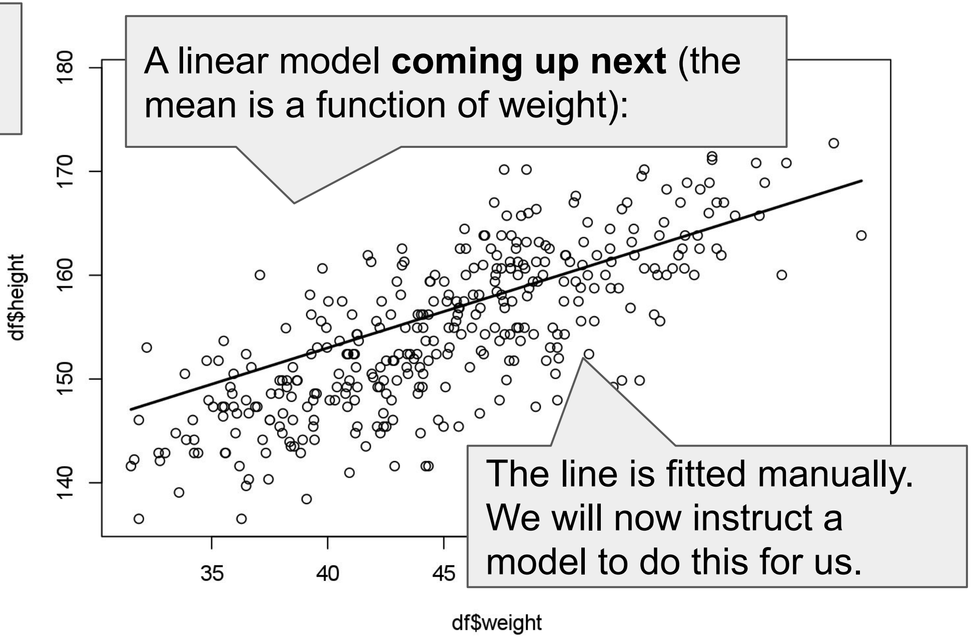
We can see a linear relationship between weight and height if plotting both variables in a scatter plot.



The **linear modeling** strategy

Defining the mean height (mu, μ) **as a function of predictor variables** (weight).





Recap: brain-dead model:

$h_i \sim \text{Normal}(\mu, \sigma)$ [likelihood]
$\mu \sim \text{Normal}(178, 20)$ [μ prior]
$\sigma \sim \text{Uniform}(0, 50)$ [σ prior]

Definition

Defining the linear model.

Height of person i.

Mean weight in the data set.

$$h_i \sim \text{Normal}(\mu_i, \sigma) \quad \text{[likelihood]}$$

$$\mu_i = \alpha + \beta\,(w_i - w\_bar) \quad \text{[linear model]}$$

No stochastic, but a **functional relationship** to define the mean (μ, mu) (written not '~' but '=').

Weight of person i.

$$\alpha \sim \text{Normal}(178, 20) \quad [\alpha \text{ prior}]$$

$$\beta \sim \text{Normal}(0, 10) \quad [\beta \text{ prior}]$$

$$\sigma \sim \text{Uniform}(0, 50) \quad [\sigma \text{ prior}]$$

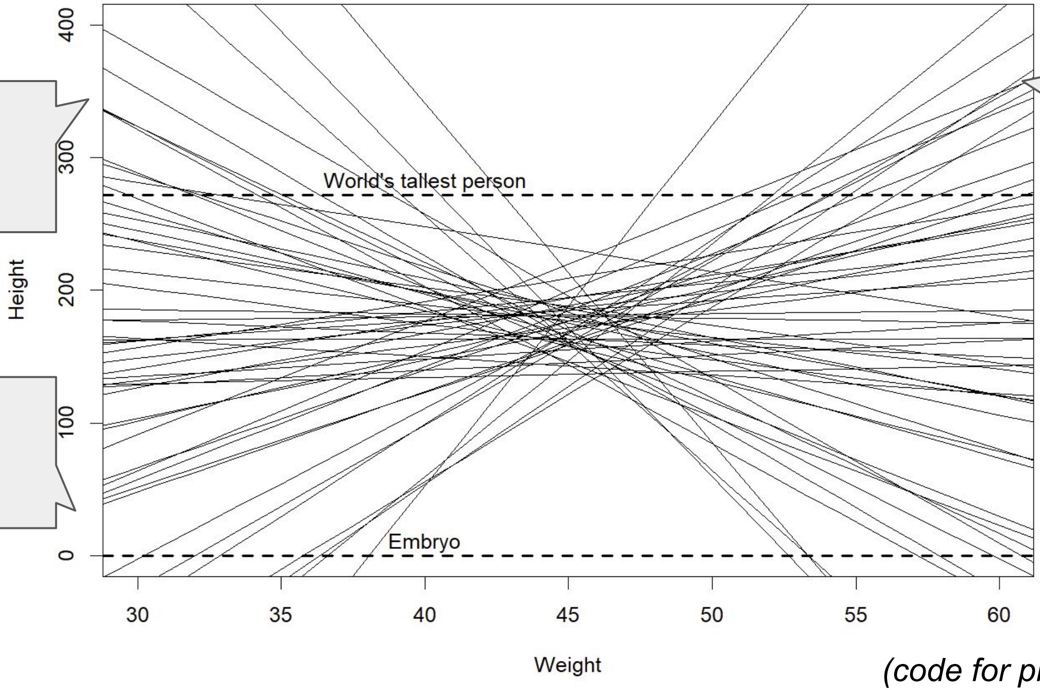
New **priors** for parameter Greek alpha (α) and Greek beta (β).

New: Prior predictive simulation:

What does this model think before seeing the data?

All these lines are **possible according to our priors**.

One of these lines is the **line we are searching** for.



There are some **crazy** lines, we could eventually improve the prior (if we like to).

(code for producing the plot on next slide)

New: Prior predictive simulation

Corresponding code to produce the previous plot.

```
w_bar <- mean(df$weight)
w <- seq(20, 70, length.out = 40)
# Draw 50 possible lines.
for (i in 1:50) {
 # Possible parameters according to the prior.
 a <- rnorm(1, mean = 178, sd = 20)
 b <- rnorm(1, mean = 0, sd = 10)
 # Simulated mu of heights.
 h <- a + b * (w - w_bar)
 # Plotting the line.
 lines(w, h)
}
```

Implementing and running the model on the data

h_i	~	$\text{Normal}(\mu_i, \sigma)$	[likelihood]
μ_i	=	$\alpha + \beta\,(w_i - w\_bar)$	[linear model]
α	~	$\text{Normal}(178, 20)$	[α prior]
β	~	$\text{Normal}(0, 10)$	[β prior]
σ	~	$\text{Uniform}(0, 50)$	[σ prior]

Math

correspondsTo

```
model <- ulam(alist(
 h ~ dnorm(mu, sigma),
 mu <- a + b * (w - w_bar),
 a ~ dnorm(178, 20),
 b ~ dnorm(0, 10),
 sigma ~ dunif(0, 50)
), data = ...)
```

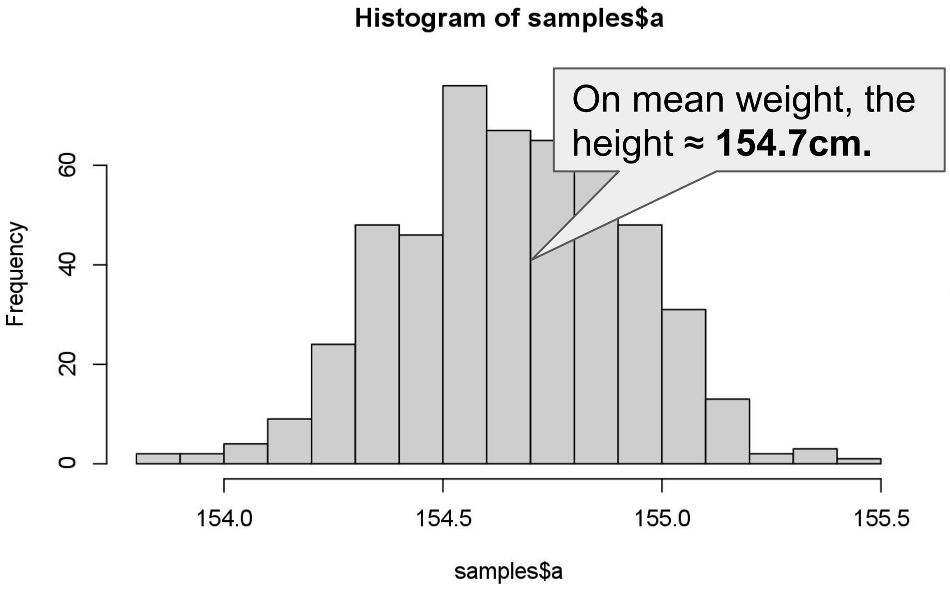
ULAM

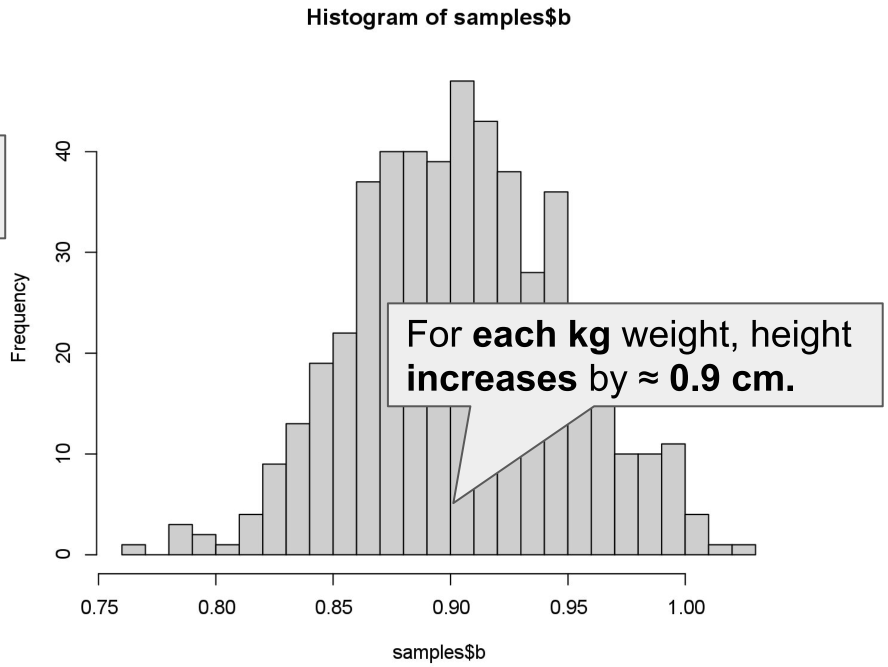
compilesTo

```
data{
        vector[346] h;
        real w_bar;
        vector[346] w;
}
parameters{
        real a;
        real b;
        real<lower=0,upper=50> sigma;
}
model{
        vector[346] mu;
        sigma ~ uniform( 0 , 50 );
        b ~ normal( 0 , 10 );
        a ~ normal( 178 , 20 );
        for ( i in 1:346 ) {
          mu[i] = a + b * (w[i] - w_bar);
        }
        h ~ normal( mu , sigma );
}
```

STAN

Results: Summarizing the (marginal*) posterior of alpha (α), and beta (β).



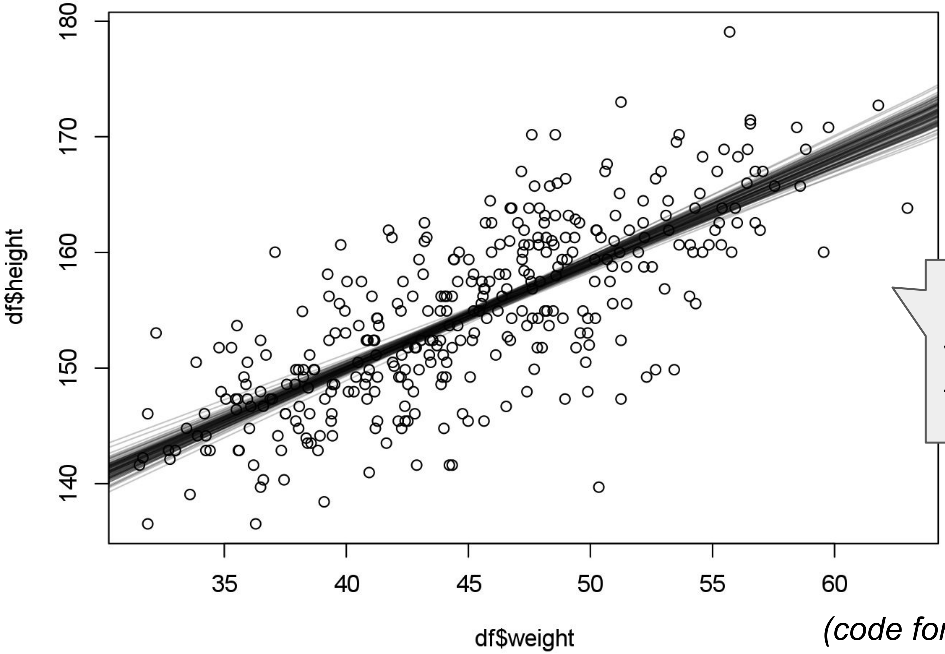


We are summarizing based on **samples** (see the previous lecture on details why we use samples)

**averaged over the other parameters*

Results: Summarizing the posterior of the mean (μ, mu).

Everything that depends upon parameters has a posterior distribution; hence, also **mean (μ, mu)** has a posterior. See the following plot for a smart visualization of this posterior.



Mu is defined as a function of weight, so we can draw a line for each posterior sample.

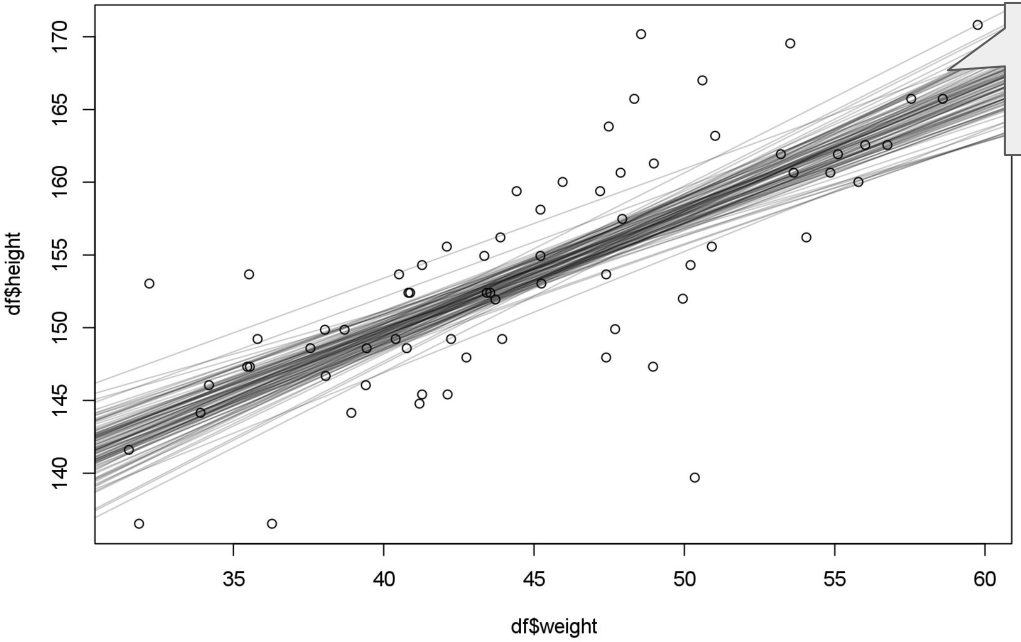
(code for producing the plot on next slide)

Code for producing the previous plot

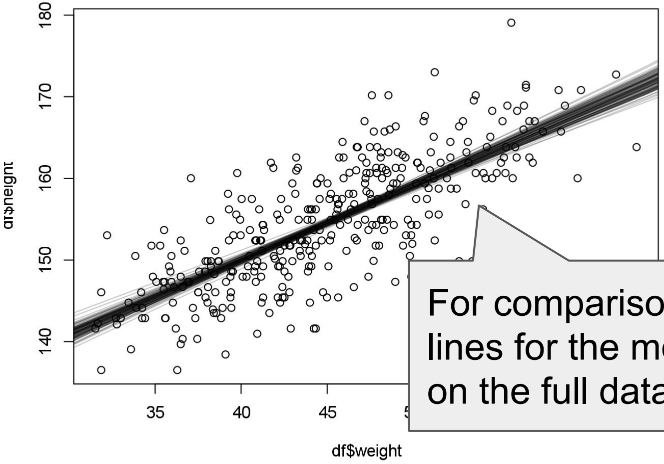
```
w <- seq(20, 70, length.out = 40)# Different possible weights (w)

for (j in 1:100) {
 # Parameters from the posterior (described by samples).
 a <- samples$a[j] # alpha
 b <- samples$b[j] # beta
 # Implementing the functional relationship.
 mu <- a + b * (w - w_bar)
 # Plotting the line (as overlay).
 lines(w, mu, col = rgb(0,0,0, alpha = 0.2))
}
```

Increasing uncertainty for of the mean (μ, mu)

We can see how the uncertainty increases **when dropping some data** (in this example we dropped **80%** of the data entries).



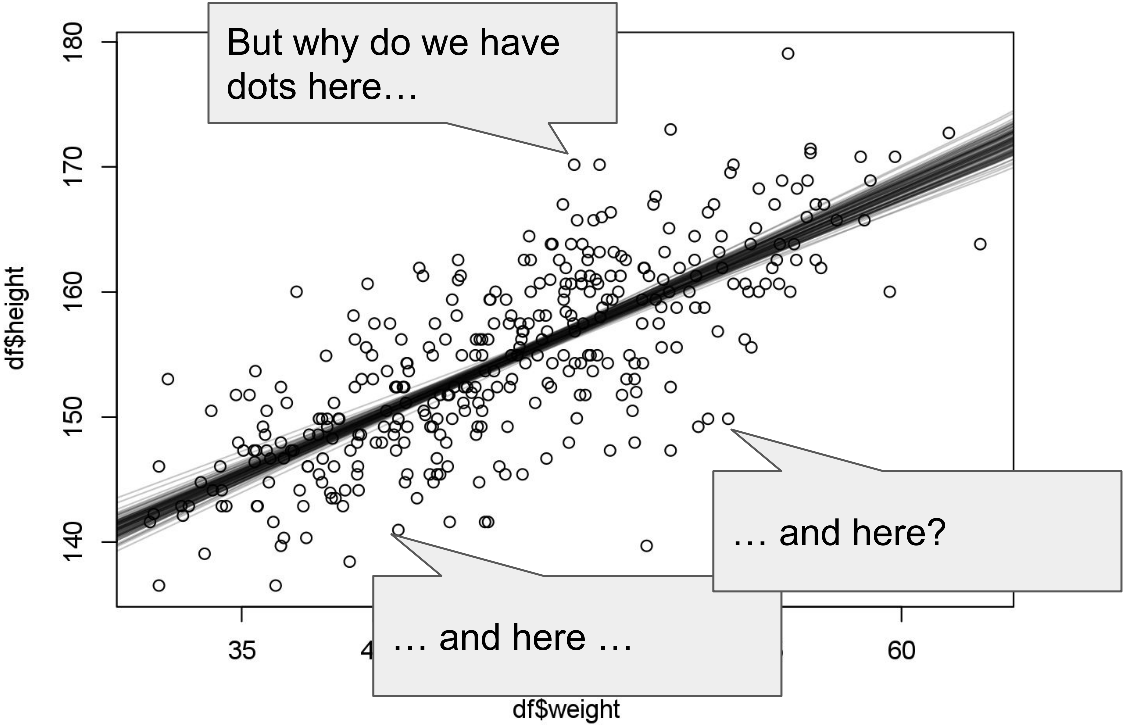
We see that the uncertainty about our mean (μ, mu) increases a lot.



For comparison, the lines for the model fit on the full data.

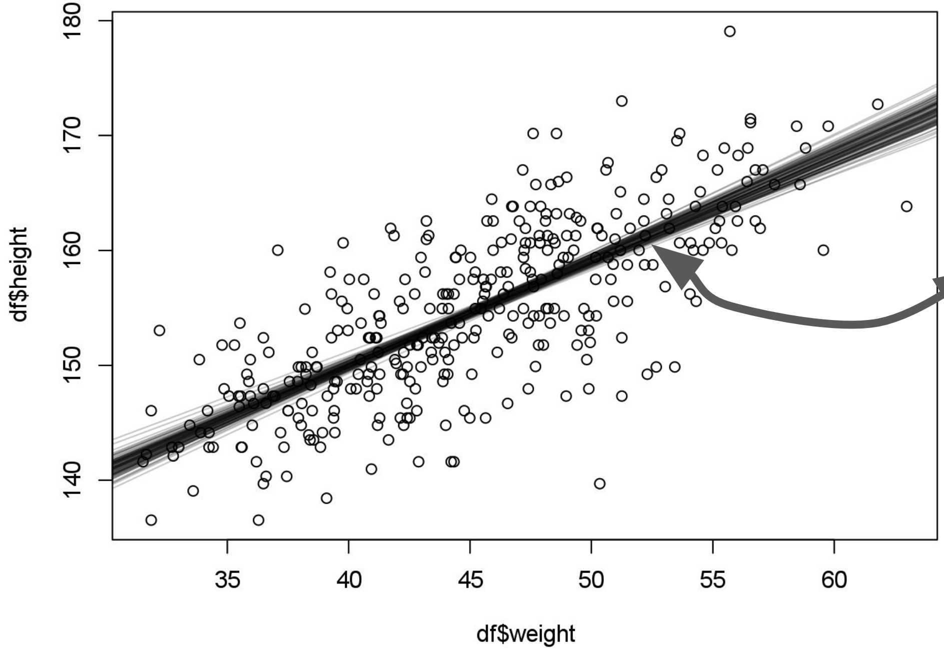
Two types of uncertainty

This is the uncertainty of the mean (μ, mu), but not of the height (h).



Two types of uncertainty

The lines depict the uncertainty of mean (μ, mu) over weight.



$h_i \sim \text{Normal}(\mu_i, \sigma)$ [likelihood]

$\mu_i = \alpha + \beta\,(w_i - w\_bar)$ [linear model]

$\alpha \sim \text{Normal}(178, 20)$ [α prior]

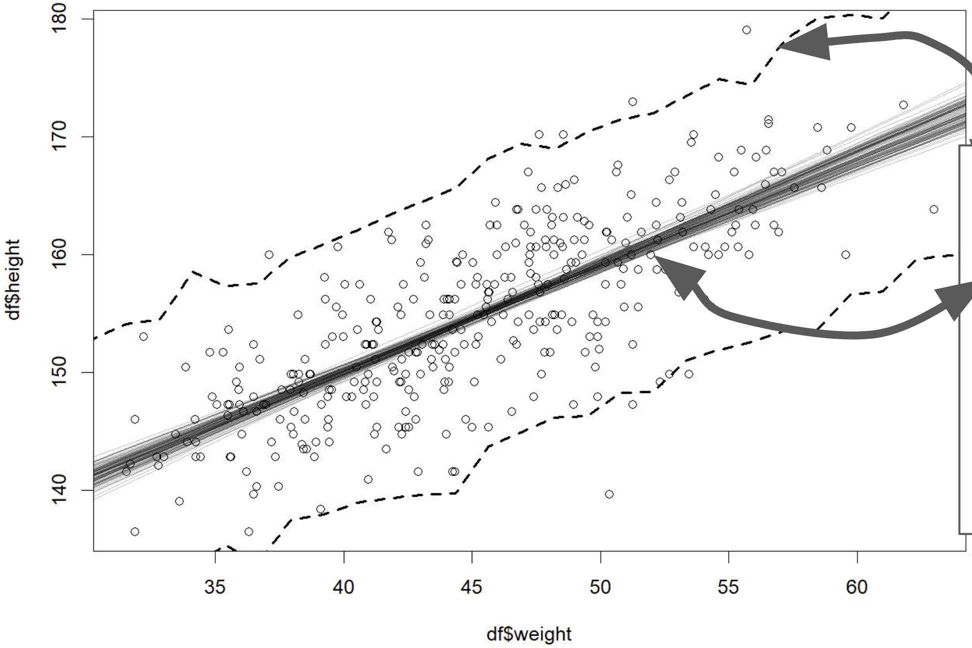
$\beta \sim \text{Normal}(0, 10)$ [β prior]

$\sigma \sim \text{Uniform}(0, 50)$ [σ prior]

Math

Two types of uncertainty

We can also depict the uncertainty of h (here is the range that should include 98% of the data entries in the model's world).



$h_i \sim \text{Normal}(\mu_i, \sigma)$ [likelihood]

$\mu_i = \alpha + \beta\,(w_i - \text{w\_bar})$ [linear model]

$\alpha \sim \text{Normal}(178, 20)$ [α prior]

$\beta \sim \text{Normal}(0, 10)$ [β prior]

$\sigma \sim \text{Uniform}(0, 50)$ [σ prior]

Math

(code for producing the plot on next slide)

Code for producing the plot

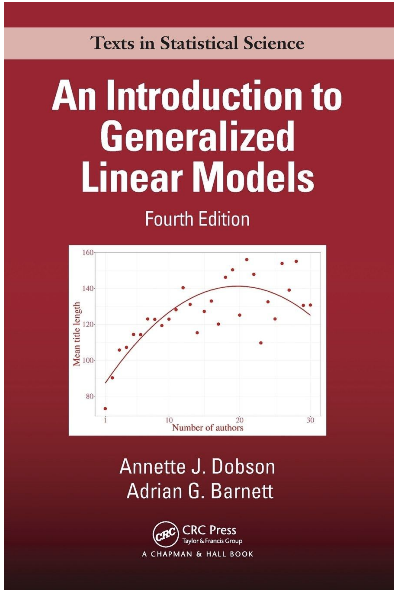
```
w <- seq(20, 70, length.out = 40)# Different possible weights (w).

# Quantiles of h for a particular w.
quantiles <- sapply(seq_along(w), function(j) {
 # Extract all parameter vectors from the posterior.
 a <- samples$a
 b <- samples$b
 sigma <- samples$sigma
 # Implement the functional relation on w.
 mu <- a + b * (w[j] - w_bar)
 # Simulating heights on weight w.
 h <- rnorm(length(mu), mu, sigma)
 # Return the quantiles of the simulated heights.
 return(quantile(h, probs = c(0.99, 0.01)))
})

# plot stuff
lines(w, quantiles[1,], lty = 2, lwd = 2)
lines(w, quantiles[2,], lty = 2, lwd = 2)
```

Summary

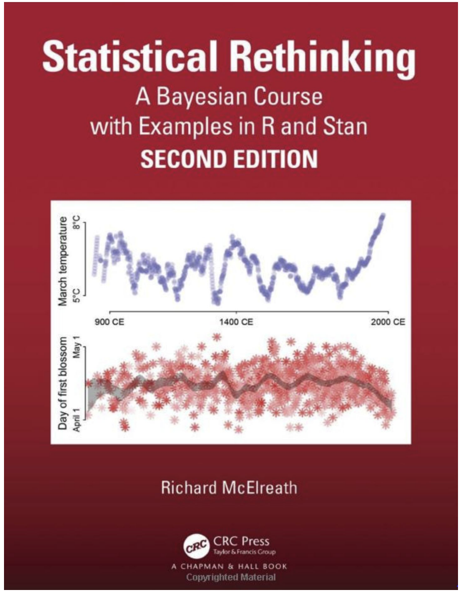
Summary

- A basic **Gaussian/normal model** of height.
- **Prior predictive simulations** to check the implications of the priors.
- A basic **linear model** relating height and weight.
- The difference between **stochastic and functional relation** connecting variables (~ or =).
- New methods to **depict the uncertainty** included in the posterior.

References:



[DobsonB18]



[McElreath20]