# Quantitive Methods 2, ZHAW

Jürgen Degenfellner

2025-01-17

# Contents

# Chapter 1

# Introduction

This script is a continuation of the first one for Quantitative Methods 1 at ZHAW.

In the first part, we learned about the basics of probability theory, descriptive statistics, Bayesian statistics, and hypothesis testing.

In this script, we will dive into the basics of statistical modeling - a world of aesthetic wonder and surprises.

This script is a first draft as you are the first group to be working with it.

Please feel free to send me suggestions for improvements or corrections.

This **should be a collaborative effort** and will (hopefully) never be finished as our insight grows over time.

The script can also be seen as a pointer to great sources which are fit to deepen your understanding of the topics. Knowledge is decentralized, and there are many great ressources out there.

For the working setup with R, please see this and the following sections in the first script.

The complete code for this script can be found here.

## 1.1 Books we will heavily borrow from are:

- (Free) Statistical Rethinking, YouTube-Playlist: Statistical Rethinking 2023
- (Free) Understanding Regression Analysis: A Conditional Distribution Approach
- Data Analysis Using Regression and Multilevel/Hierarchical Models

- (Free) Doing Bayesian Data Analysis

# Chapter 2

# Introduction

## 2.1 What is statistical modeling and what do we need this for?

Typically, one simplifies the complex reality (and loses information) in order to make it better understandable, mathematically treatable and to make predictions.

Underlying our models, there are theories which should be falsifiable and testable. For instance, I would be really surprised if I pull up my multimeter and measure the voltage (V) and electric current (I) at a resistence (R) in a circuit and find that Ohm's law $V = IR$ is not true. This **law** can be tested over and over again and if one would find a single valid counterexample, the law would be falsified. It is also true that the law is probably not 100% accularate, but an extremely good approximation of reality. Real-world measurements carry measurement errors and when plotting the data, one would see that the data points might not lie exactly on a straight line. This is not a problem.

A statistical model is a mathematical framework that represents the relationships between variables, helping us understand, infer, and predict patterns in data. It acts as a bridge between observed data and the real-world processes that generated them. In health research, where variability and uncertainty are inherent, statistical models are valuable tools for making sense of complex phenomena. You can watch this as short intro.

Depending on the task at hand, we would use different models. In any case, logical reasoning and critical thinking comes first, then comes the model. **It makes no sense to estimate statistical models just for the sake of it**.

**All models are wrong, but some are useful**. Or to quote George Box:

> "Since all models are wrong the scientist cannot obtain a 'correct' one by excessive elaboration. On the contrary following William of Occam he should seek an economical description of natural phenomena. Just as the ability to devise simple but evocative models is the signature of the great scientist so overelaboration and overparameterization is often the mark of mediocrity."

In my opinion, statistical modeling is an art form: difficult and beautiful.

**One goal of this course** is to improve interpretation and limitations of statistical models. They are not magical turning data into truth. Firstly, the rule gargabe in, garbage out (GABA) applies. Secondly, statistical models are based on data and their variability and have inherent limitations one cannot overcome even with the most sophisticated models. This is expressed for instance in the so-called bias-variance trade-off. You can't have it all.

### 2.1.1 Explanatory vs. Predictive Models

I can recommend reading this article by Shmueli et al. (2010) on this topic.

Statistical models serve different purposes depending on the research question. Two primary goals are **explanation** and **prediction**, and each requires a different approach:

**Explanatory Models** focus on understanding causal relationships. These models aim to uncover mechanisms and answer **"why"** questions. For example:

- Does smoking increase the risk of lung cancer? **Yes**. (If you want to see what a large effect-size looks like, check out this study.)
- How large is the "effect" of smoking on lung cancer? **Large**.
- Does pain education and graded sensorimotor relearning improve disability (a question we ask in our Resolve Swiss project)?

Explanatory models are **theory-driven**, designed to test hypotheses. Here, one wants to understand the underlying mechanisms and the relationships between variables and hence often uses (parsimonious) models that are more interpretable, like linear regression.

**Predictive Models** prioritize forecasting future outcomes based on patterns in the data. These models aim to answer **"what will happen?"** For instance:

- Gait analysis using Machine Learning (ML)?
- Skin cancer detection using neural networks?

Predictive models are **data-driven**, often using complex algorithms to achieve high accuracy. Their success is measured using metrics like Root Means Square

Error (RMSE), Area Unter the Curve (AUC), or **prediction error on new, unseen data**. Any amount of model complexity is allowed. One could for instance estimate a neural network ("just" another statistical model) with many hidden layers and neurons in order to improve prediction quality. Interpretability of the model weights is not a priority here.

While explanatory and predictive goals often complement each other, their differences highlight the importance of clearly defining the purpose of your analysis. In applied health research, explanatory models help identify causal mechanisms, while predictive models can guide real-world decisions by providing actionable forecasts. Together, they enhance both our understanding of phenomena and our ability to make informed decisions in complex environments.

### 2.1.2 Individual vs. Population Prediction

Another important distinction is between **individual vs. population** prediction. In the smoking example above, we can be very sure about the mean effects that smoking has on lung cancer. On an individual level, it is harder to predict the outcome. Nevertheless, individual predictions will be (notably) better than random guessing. We will discuss this in greater detail.

### 2.1.3 Practical Use of Statistical Models

In my optinion, we should never be afraid to test our statistical models (as honestly as possible) against reality. We could for instance ask ourselves:

- "How much better does this model classify than the arithmetic mean? (i.e., the linear model with just an intercept)"
- "How much better does this model classify than random guessing?"
- Is it worth the effort to collect data and estimate this model by using hundreds of hours of our time?

In some cases, these questions can be answered straightforwardly.

- In advertising (Google, Facebook, …), a couple of percentage points in prediction quality might make a difference of millions of dollars in revenue offsetting the statistitians salary.
- Improved forecasts of a few percentage points in the stock market or just being slightly better than the average, will make you faboulously rich.
- Improved cancer forecasting might save lives, money and pain and is not only measured in money.

### 2.1.4 Start at the beginning

What do we actually want to do in general? Very broadly speaking we want to: **describe** the association of variables to each other that carry variability. Hence, the relationship is not deterministic like

$$y = 2x + 3$$

but rather we need to "loosen up" the relationship to account for variability (in $x$ and $y$). So, the values 2 and 3 are not fixed but aflicted with uncertainty. Depending on your philosophical view, you might say you want to find the "true" but unknown relationship between variables. This is what we do in simulation studies all the time: We know the true relationship, simulate data by adding variability and then try to estimate the true relationship we assumed in the first place. For some practical applications, we can get a really nice and complete answer to our question (for instance sample size for proportions).

So we are looking for a function $f$ such that

$$Y = f(X)$$

where

- $Y$ is the "outcome", "dependent variable" or "response".
- $X$ are the "predictors". $X$ can be a single Variable $x$ or many variables $x_1, x_2, ..., x_p$.

It is important to be aware of the notation here: "Predict" does **not necessarily** mean that we can predict the value in the future. It merely means we estimate the value (or mean) of $Y$ given $X$.

- This can be done at the same time points, known as **cross-sectional** analysis ("What is the maximum jumping height of a person given their age at a certain point in time, whereas both variables are measured at the same time?");
- or at different time points, known as **longitudinal analysis** ("What is the maximum jumping height of a person 10 years later ($t_2$) given their baseline health status at time $t_1$?").

The **simplest statistical model** would be the mean model where $Y$ is "predicted" by a constant: $Y = c$ which (at least in the classical linear regression) turns out to be $c = \bar{x}$. This simple model is often surprisingly good, or, to put it in other words, models with more complexity are often not that much better with regards to multiple metrics.

## 2.2 A (simple) model for adult body heights in the Bayesian framework

As repetition, read the parts about Bayes statistics from QM1 again to refresh your memory about the Bayesian framework.

It's recommendable to read the beginning of the book Statistical rethinking up until page 39 as well. We are not completely new to the topic of Bayes due to QM1.

We want to **start building our first model** right away.

Let's begin with the example in Statistical rethinking using data from the !Kung San people starting on page 79.

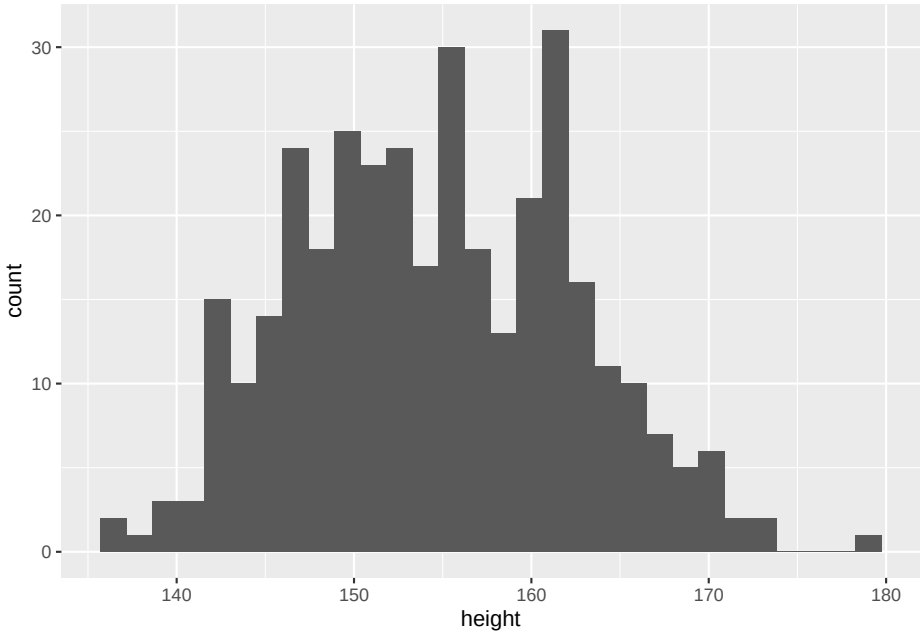
```
library(rethinking)
data("Howell1")
d <- Howell1
str(d)
```

```
## 'data.frame':    544 obs. of  4 variables:
##  $ height: num  152 140 137 157 145 ...
##  $ weight: num  47.8 36.5 31.9 53 41.3 ...
##  $ age   : num  63 63 65 41 51 35 32 27 19 54 ...
##  $ male  : int  1 0 0 1 0 1 0 1 0 1 ...
```

```
d2 <- d[d$age >= 18, ] # only adults
```

We want to model the adult height of the !Kun San people using prior knowledge (about the Swiss population) and data.

```
library(tidyverse)
d2 %>% ggplot(aes(x = height)) + geom_histogram()
```




Since we already have domain knowledge in this area, we can say that heights are usually normally distributed, or at least a mixture of normal distrubutions (female/male). We assume the following model:

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

As in QM1, we want to start with a Bayesian model and hence, we need some priors.

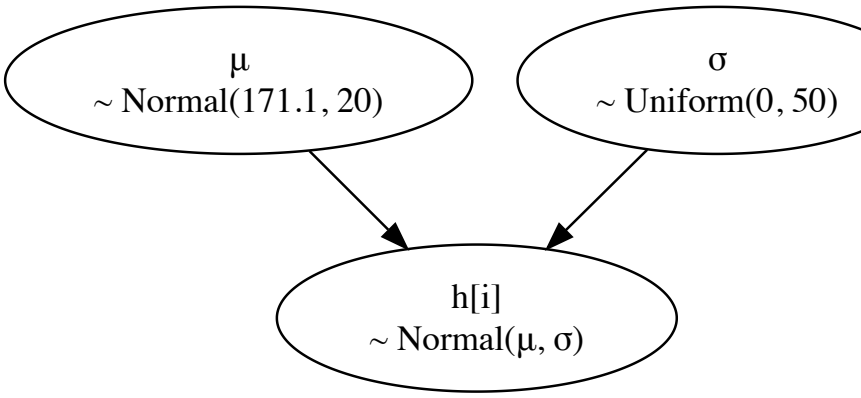
Since we are in Switzerland and just for fun, we use the mean of Swiss body heights as expected value for the **prior for the mean**. According to the link (Bundesamt für Statistik), the mean height of $n = 21,873$ people in the Swiss sample is 171.1 cm. We choose the same $\sigma$ for the prior of the normal as in the book not to deviate too much from the example at hand.

Next comes our **model definition in the Bayesian framework**, which I often find more intuitive than the frequentist approach:

$$\begin{aligned} h_i &\sim \text{Normal}(\mu, \sigma) \\ \mu &\sim \text{Normal}(171.1, 20) \\ \sigma &\sim \text{Uniform}(0, 50) \end{aligned}$$

**Description of the model definition**: The heights are normally distributed with unknown mean and standard deviation. As our current knowledge about the mean height, we use a prior distribution for the mean (we do not know but

want to estimate) by assuming the mean of a population we know and a standard deviation of 20 cm which allows are rather large range of possible values for $\mu$. $\sigma$ is also unknown and a priori we restrict ourselves to values between 0 and 50 cm, whereas we assign equal plausibility to all values in this range (which can and should be critically discussed).




**Vizualisation of the model structure**:

Mind that there is a conceptual difference between the normal distribution of the heights and the normal prior distribution of the mean. The latter expresses our prior knowledge/insecurity about the unobserved mean. The normal distribution says we expect the heights to be normally distributed but we do not know the parameters ($\mu$ and $\sigma$) yet.

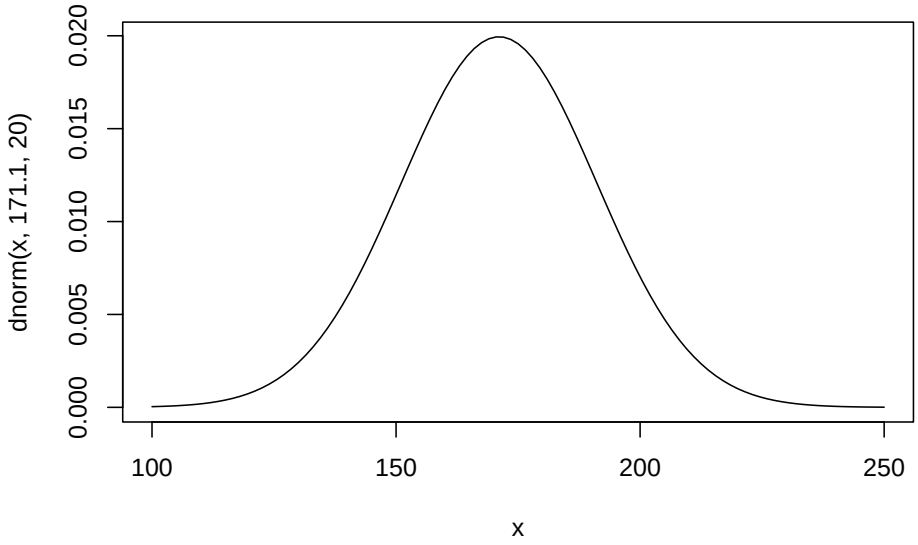
Of course we would not need the prior here due to the large sample size, but let's do it anyways for demonstration purposes. We are not completely uninformed about body heights and express our knowledge with the prior for $\mu$. The 20 in the prior for the mean expresses our range of possible true mean values and aknowledge that there are a variety of different subpopulations with different means.

Using the Swiss data in the link one could estimate that the standard deviation of the heights from $21,873$ Swiss people is around is 25.6553 cm (Exercise 1).

Remember, in the Baysian world, there is no **fixed but unknown** parameter, but instead we define a distribution over the unobserved parameter.

We **visualize the prior for** $\mu$.

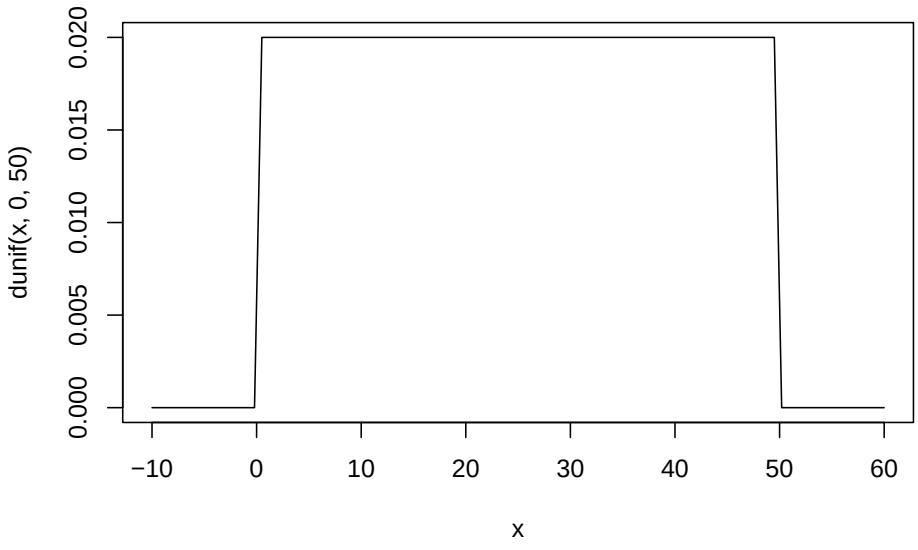
```
curve(dnorm(x, 171.1, 20), from = 100, to = 250)
```




The **prior for** $\sigma$ is uniform between 0 and 50 cm. This is a very wide prior and just constrains the values to be positive and below 50 cm. This could be stronger of course.

**Visualization of the prior for** $\sigma$:

```
curve(dunif(x, 0, 50), from = -10, to = 60)
```




Note, we didn't specify a prior probability distribution of heights directly, but once we've chosen priors for $\mu$ and $\sigma$, these imply a prior distribution of individual heights.

**Without** even having seen the **new data**, we can check what our prior (model) for heights would predict. This is important. If the prior already predicts impossible values, we should reconsider our priors and/or model.

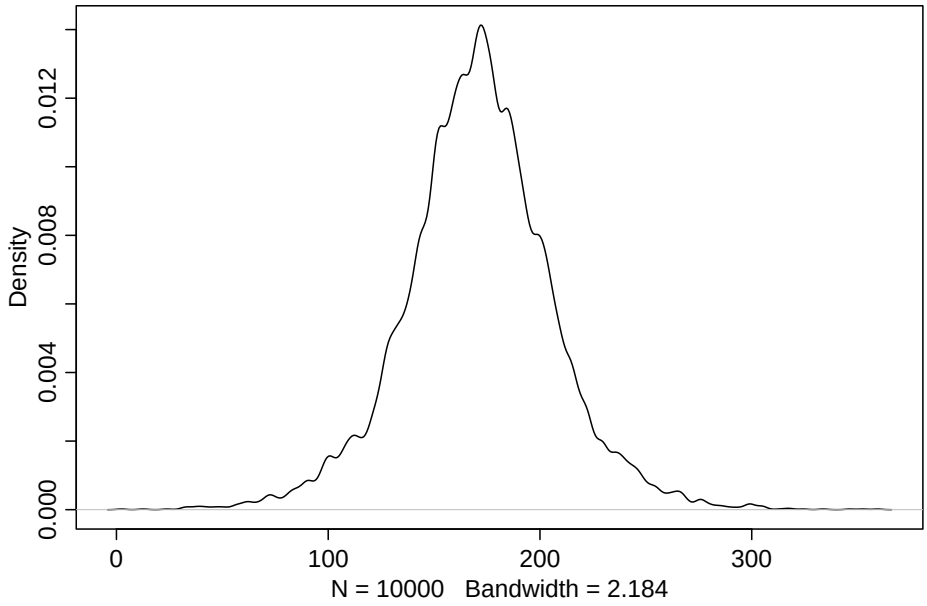
So, we simply draw $\mu$ and $\sigma$ from the priors and then draw heights from the normal distribution using the drawn parameters.

**Vizualisation of the prior for heights**:

```
sample_mu <- rnorm(10^4, 171.1, 20)
sample_sigma <- runif(10^4, 0, 50)
prior_h <- rnorm(10^4, sample_mu, sample_sigma)
length(prior_h)
```

```
## [1] 10000
```

```
dens(prior_h)
```




The prior is not itself a Gaussian distribution, but a distribution of relative plausibilities of different heights, before seeing the data.

Now, there are a couple of different ways to estimtate the model incorporating the new data. For didactic reasons, grid approximation is often used (as in the book). For many parameters, grid approximation becomes more and more infeasible (due to combinatorial explosion).

We will skip that for now and use quadratic approximation instead which works well for many common procedures in applied statistics (like linear regression).

Later, you'll probably use (or the software in the background) mostly Markov chain Monte Carlo (MCMC) sampling to get the posterior. Pages 39 and the following explain the 3 concepts grid approximation, quadratic approximation and MCMC.

In short, **quadratic approximation** assumes that our posterior distribution of body heights can be approximated well by a normal distribution, at least near the peak.

Please read the addendum to get a clearer picture of what a bivariate normal distribution is.

Using the library `rethinking` we can estimate the model using quadratic approximation. First, we define the model in the `rethinking` syntax (see R code 4.25 in the book).

```
library(rethinking)
flist <- alist(
  height ~ dnorm(mu, sigma),
  mu ~ dnorm(171.1, 20),
  sigma ~ dunif(0, 50)
)
```

Then we estimate/fit the model using quadratic approximation.

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

Now let's take a look at the fitted model: (Note: In the online-version of the book, they used the command `map` instead of `quap`.)

```
precis(m_heights)
```

```
##             mean        sd       5.5%      94.5%
## mu    154.606387 0.4119860 153.947954 155.264820
## sigma   7.731172 0.2913708   7.265505   8.196838
```

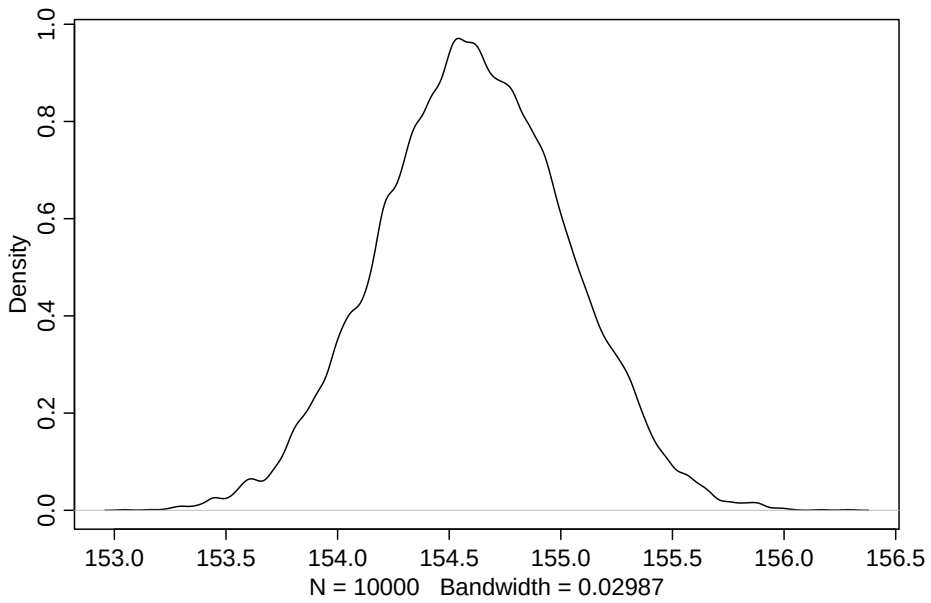
Above, we see the mean of the posterior for $\mu$ **and** $\sigma$; and a 89% credible interval for those parameters.

We can now plot the posterior distribution of the mean and the standard deviation.

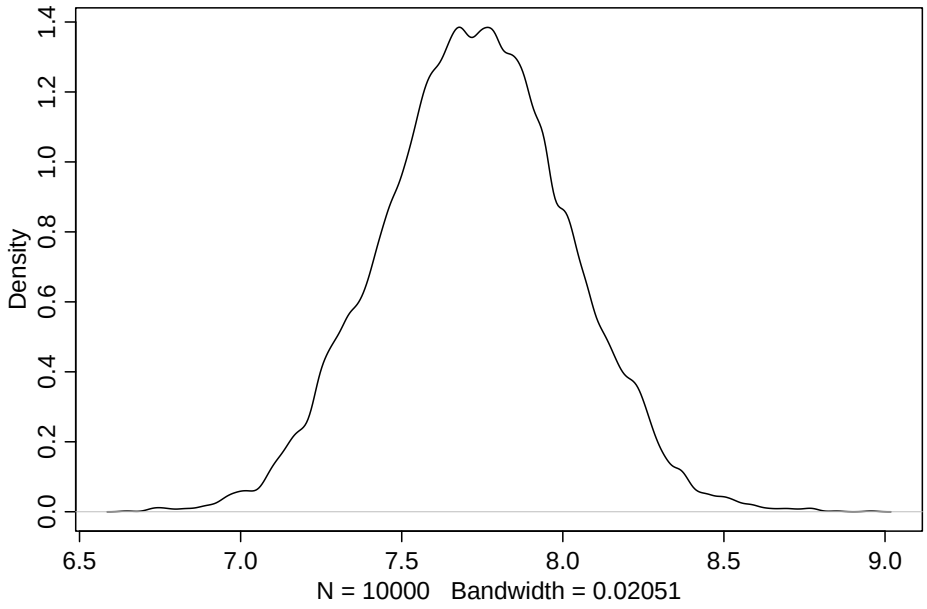
```
post <- extract.samples(m_heights, n = 10^4)
head(post)
```

```
##         mu    sigma
## 1 154.3658 8.455891
## 2 154.3274 7.903180
## 3 154.5510 7.704259
## 4 154.3533 7.778473
## 5 154.2720 8.143232
## 6 154.2503 7.849285
```

```
dens(post$mu)
```




```
dens(post$sigma)
```




Note, that **these samples come from a multi-dimensional posterior distribution**. In our case, we approximated the posterior with a bivariate normal distribution. They are not necessarily independent from each other, but in this case they are. We know this from the model definition above. $\mu$ and $\sigma$ are both defined as normal respectively uniform distributions and by definition do not influence each other. This is also visible in the vizualisation of the model structure: There is no confounding variable or connection between those priors. One could think of a common variable $Z$ that influences both $\mu$ and $\sigma$. This could be genetic similarity which could influence both $\mu$ and $\sigma$.

Let's verify that $\mu$ and $\sigma$ are uncorrelated:

```
vcov(m_heights)
```

```
##                mu        sigma
## mu    0.1697324421 0.0002040443
## sigma 0.0002040443 0.0848969479
```

gives you the variance-covariance matrix of the parameters of the posterior distribution. In the diagonal you see the variance of the parameters.

```
diag(vcov(m_heights))
```

```
##         mu      sigma
## 0.16973244 0.08489695
```

And we can compute the correlation matrix easily:

```
cov2cor(vcov(m_heights))
```

```
##                mu       sigma
## mu    1.000000000 0.001699793
## sigma 0.001699793 1.000000000
```

Let's plot the posterior in 3D, because we **can**:

WebGL is not supported by your browser - visit https://get.webgl.org for more info

**How beautiful ist that?**

We see in the 3D plot, that the "mountain" is not rotated, indicating graphically that the parameters are independent from each other.

We also see in the correlation matrix, the correlation of the parameters is $\sim 0$. In the context of a joint normal distribution, this means that the parameters are independent.

And, it is not an accident that the posterior looks like this. Using quadratic approximation, we used the bivariate normal distribution to **approximate** the posterior.

## 2.3 Classical approach for the simplest model

We have seen, how we could use domain and prior knowledge to fit a very simple model for body heights of a population (!Kung San) in the Bayesian framework.

Now, let's start at the same point in the classical framework. Here, we do not use any prior knowledge, at least not that explicitely.

The classical approach to fit a regression line is the so-called **least squares method**.

The **(simple mean-) model** is:

$$Y_i = height_i = c + \varepsilon_i$$

- for some $c \in \mathbb{R}$ and
- normally distributed errors $\varepsilon_i \sim \text{Normal}(0, \sigma)$.

These are on average zero and have a constant standard deviation of $\sigma$. So, we assume there is a fixed, but unknown, constant $c$ that we want to estimate and we assume that there is a special sort of error in our model that is normally distributed. Sometimes there is a large deviation from the true $c$, sometimes there is a small deviation. On average, the deviations are zero and the **errors should also be independent from each other**:

$$\varepsilon_i \perp \varepsilon_j \text{ for } i \neq j$$

This means that just because I have just observed a large deviation from the true $c$ does not mean, that the probability of a large deviation in the next observation is higher/lower. Note, that we cannot readily define different types of errors in the classical framework.

But what is $c$? We determine the shape of the model ourselves (constant model, or mean model) and then estimate the parameter $c$. By defining the shape of the model ourselves and imposing a distribution where we want to estimate the parameter of said distribution, we are in **parametric statistics**.

We choose the $c$ which minimizes the sum of squared errors from the actual heights. This has the advantage that deviations upper and lower from the actual height are equally weighted. The larger the deviation the (quadratically) larger the penalty.

**Why do we do that?** Because, if the model assumptions (more on that later) are correct, the least squares estimator is a really good estimator. How good? Later…

We want to miminize the following function:

$$SSE \text{ (Sum of Squared Errors) } (c) = (height_1 - c)^2 + (height_2 - c)^2 + ... + (height_n - c)^2 =$$

$$= \sum_{i=1}^{n} (height_i - c)^2$$

The SSE is a function of $c$ and we want to find the $c$ that minimizes the function. Since it is a quadratic function, we can always find the minimum. We have learnt

in school how to do this (hopefully): Take the derivative of the function and set it to zero. Solve for $c$ and you have the $c$ which yields the minimum of SSE(c).

Let's do that:

$$\frac{d}{dc}SSE(c) = 2(height_1 - c)(-1) + 2(height_2 - c)(-1) + ... + 2(height_n - c)(-1) =$$

$$= -2\sum_{i=1}^{n}(height_i - c)$$

This should be zero for the minimum:

$$-2\sum_{i=1}^{n}(height_i - c) = 0$$

$$\sum_{i=1}^{n}(height_i - c) = 0$$

$$\sum_{i=1}^{n}(height_i - c) = 0$$

$$\sum_{i=1}^{n} height_i - n \cdot c = 0$$

$$\hat{c} = \frac{1}{n}\sum_{i=1}^{n} height_i = \overline{height_i}$$

The hat over the $c$ indicates that this is the estimated value of $c$. Everytime we estimate a parameter, we put a hat over it.

And voilà, we have estimated the parameter $c$ of the model, which is just the sample mean of all the heights. In contrast to before, we did not put in a lot of prior knowledge, but just estimated the parameter from the data.

In R, we can do this easily:

```
mod <- lm(height ~ 1, data = d2)
summary(mod)
```

```
##
## Call:
## lm(formula = height ~ 1, data = d2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -18.0721  -6.0071  -0.2921   6.0579  24.4729
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 154.5971     0.4127   374.6   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.742 on 351 degrees of freedom
```

```
dim(d2)
```

```
## [1] 352   4
```

```
mean(d2$height) # same as the intercept
```

```
## [1] 154.5971
```

```
sd(d2$height) / sqrt(nrow(d2)) # standard error of the estimator
```

```
## [1] 0.4126677
```

```
# test-statistic for the intercept:
mean(d2$height) / (sd(d2$height) / sqrt(nrow(d2)))
```

```
## [1] 374.6285
```

```
# residual standard error:
sqrt(sum(mod$residuals^2) / (nrow(d2) - 1))
```

```
## [1] 7.742332
```

the `~1` means that there is just a so-called **intercept** in the model. There are **no covariates**, just the constant $c$. This is the simplest we can do. `lm` stands for linear model and with this base command in R we ask the software to do the least squares estimation for us.

Let's look at the **R-output** of the model estimation:

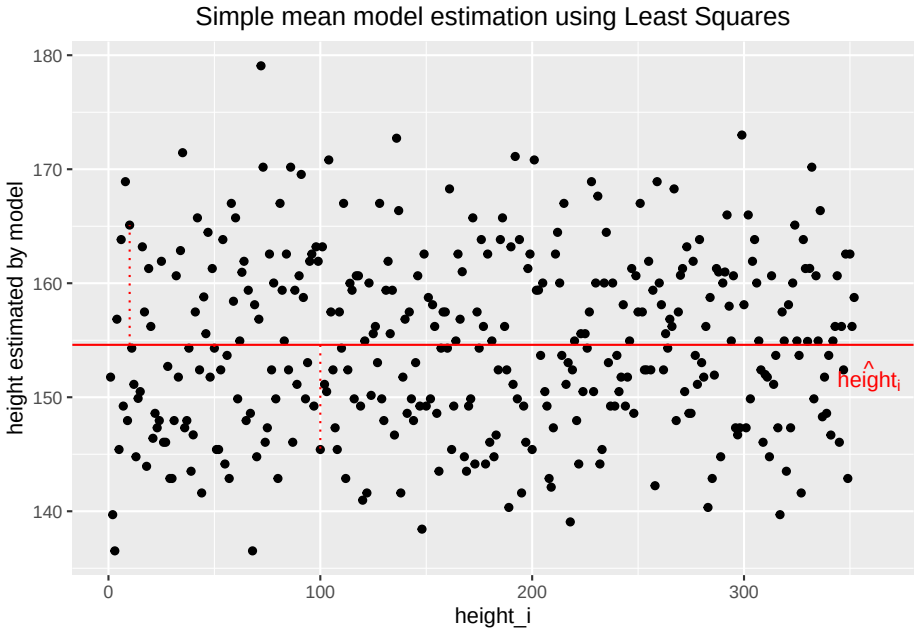
- `lm(formula = height ~ 1, data = d2)`: This is the model we estimated.

- `Residuals`: The difference between the actual height and the estimated height: $r_i = height_i - \hat{c}$. A univariate 5-point summary is given.
- `Coefficients`: The estimated coefficients of the model. In this case, there is just the intercept. We get the
  - `Std. Error` of the estimate, i.e. the standard error of the mean, which is (according to the Central Limit Theorem)

$$\frac{\sigma}{\sqrt{n}}$$

    and can be estimated by the sample standard deviation divided by the square root of the sample size.
  - the `t value` and the `Pr(>|t|)` which is the $p$-value of the (Wald-)test of the null hypothesis that the coefficient is zero ($H_0$ : intercept = 0). This is a perfect example of an absolutely useless $t$-test. Why? Because obviously (exercise 2) the population mean of body heights is not zero.
- `Residual standard error`: The standard deviation of the residuals $r_i = height_i - \hat{c}$. In this case identical with the sample standard deviation of heights (exercise 3). 351 degrees of freedom. There are 352 observations and 1 parameter estimated (intercept/mean). Hence, there are $352 - 1 = 351$ freely movable variables in the statistic of the sample standard deviation.

Let's look at the situation graphically:



Above, the heights are plotted against the index of the observation. The variability of heights around the regression line (constant in this case) seems to stay constant, which is a good sign. We will call this **homoscedasticity** later. The dashed vertical red lines show two residuals, the difference between the actual height and the estimated height. The model-estimated heights ($\widehat{heights}_i$) are all identical and nothing but the mean of all heights.

Peter Westfall explains in his excellent book a conditional distribution approach to regression. I highly recommend reading the first chapters.

What does this mean in this context?

## 2.4 Exercises

### 2.4.1 [E] Exercise 1

Use the Swiss body heights data to determine - the 95% "Vertrauensintervall" for $\mu$ and - calculate the standard deviation of the heights from $21,873$ Swiss people.

### 2.4.2 [E] Exercise 2

Why do we not need a hypothesis test to know that the population mean of body heights is not zero? Give 2 reasons.

### 2.4.3 [M] Exercise 3

Verify analytically that the `Residual standard error` is identical with the sample standard deviation of the heights.

### 2.4.4 [M] Exercise 4

Repeat the estimation of the simple model using a different data set about chicken weights, which is included in R.

- Set useful priors for the mean and standard deviation of the model for the Baysian and the frequentist version considering your a priori knowledge about chicken weights.

## 2.5 Addendum

### 2.5.1 The bivariate normal distribution

As a refresher, you can look into the old QM1 script and read the chapter "4.7 Gemeinsame Verteilungen". Maybe this video also helps.

The bivariate normal distribution is a generalization of the normal distribution to two dimensions. Now, we look at the distribution of two random variables $X$ and $Y$ **at the same time**.

Instead of one Gaussian bell curve, we have a 3D bell curve. This curve defines how plausible different combinations of $X$ and $Y$ are.

Single points (like (3,6)) still have probability zero, because now the **volume** over a single point ($x$, $y$) is zero. The probability of a certain area is now the **volume** under the curve compared to the **area** under the density curve in the one-dimensional case.

**Example**: The following plot shows the density of a bivariate normal distribution of two variables $X$ and $Y$ with $\mu_X = 0$, $\mu_Y = 0$, $\sigma_X = 1$, $\sigma_Y = 1$ and $\rho = \frac{2}{3}$.

Below is the correlation matrix of the bivariate normal distribution.

```
##           [,1]      [,2]
## [1,] 1.0000000 0.6666667
## [2,] 0.6666667 1.0000000
```

WebGL is not supported by your browser - visit https://get.webgl.org for more info

If you move the plot around with your mouse, you see that there is a positive correlation between $X$ and $Y$ ($\rho = \frac{2}{3}$). This means that if $X$ is above its mean, $Y$ is also more likely to be above its mean. The variances of $X$ and $Y$ are both 1. That means, that if you cut through the plot in $X = 0$ or $Y = 0$, you see the same form of normal distribution. If you look at if from above, we have

hihglighted the section on the surface over the area $X \in [0.5, 2]$ and $Y \in [0.5, 2]$. The volume over this area under the density curve is the probability of this area: $P(X \in [0.5, 2] \text{ and } Y \in [0.5, 2])$

# Chapter 3

# Simple Linear Regression

## 3.1 Simple Linear Regression in the Bayesian Framework

We will now add one covariate/explanatory variable to the model. Refer to Statistical Rethinking "4.4 Linear prediction" or "4.4 Adding a predictor" as it's called in the online version of the book.

So far, our "regression" did not do much to be honest. The mean of a list of values was already calculated in the descriptive statistics section before and we have mentioned how great this statistic is as measure of location and where its weaknesses are.

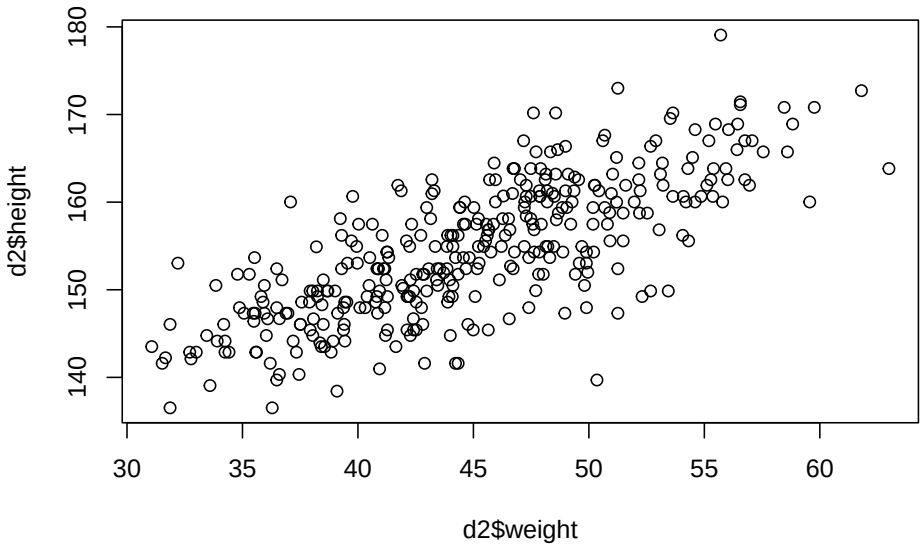
Now, we want to model how body height and weight are related. Formally, one wants to *predict* body heights from body weights.

Here and in the frequentist framework, we will see that it is **not the same** problem (and therefore results in a different statistical model) **to predict body weights from body heights or vice versa**.

The word "predictor" is important here. It is a technical term and describes a variable that we know (in our case weight) and with which we want to "guess as good as possible" the value of the dependent variable (in our case height).

We **always** visualize the data first to improve our understanding.

```
plot(d2$height ~ d2$weight)
```




The scatterplot indicates a linear relationship between the two variables. The higher the weight, the higher the height; with some deviations of course. This relationsip is neither causal, not deterministic.

- It is not causal since an increase in weight does not necessarily lead to an increase in height, especially in grown-ups.
- It is not deterministic since there are deviations from the line. It if was deterministic, we would not need statistical modeling.

For simpler notation, we will call `d2$weight` $x$. $\bar{x}$ is the mean of $x$.

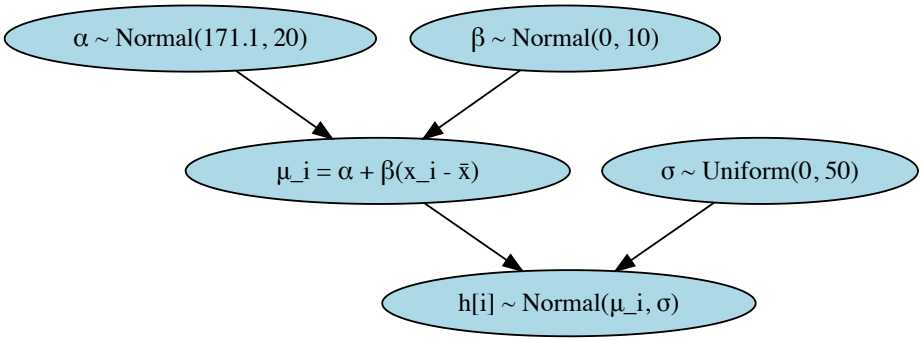
### 3.1.1 Model definition

Let's write down our **model** (again with the Swiss population prior mean):

$$
\begin{aligned}
h_i &\sim \text{Normal}(\mu_i, \sigma) \\
\mu_i &\sim \alpha + \beta(x_i - \bar{x}) \\
\alpha &\sim \text{Normal}(171.1, 20) \\
\beta &\sim \text{Normal}(0, 10) \\
\sigma &\sim \text{Uniform}(0, 50)
\end{aligned}
$$

Visualization of the **model structure**:

```
## file:////private/var/folders/pm/jd6n6gj10371_bml1gh8sc5w0000gn/T/Rtmpl22xBb/file371
```




There are now additional lines for the priors of $\alpha$ and $\beta$. The model structure also shows the way to simulate from the prior. One starts at the top and ends up with the heights.

- $h_i$ is the height of the $i$-th person and we assume it is normally distributed.
- $\mu_i$ is the mean of the height of the $i$-th person and we assume it is normally distributed. Compared to the intercept model, a different mean is assumed for each person. **The mean $\mu_i$ is linearly dependent on the weight** of the $i$-th person (second line).
- $\alpha$ is the intercept and we use the same prior as before.
- $\beta$ is the slope of the line and we use the normal distribution as prior for it, hence it can be positive or negative and how plausible each value is, is determined by that specific normal distribution. Note, that we could easily adapt the distribtion to any distribution we like.
- The prior for $\sigma$ is unchanged.
- $x_i - \bar{x}$ is the deviation of the weight from the mean weight, thereby **we center** the weight variable. This is a common practice in regression analysis.

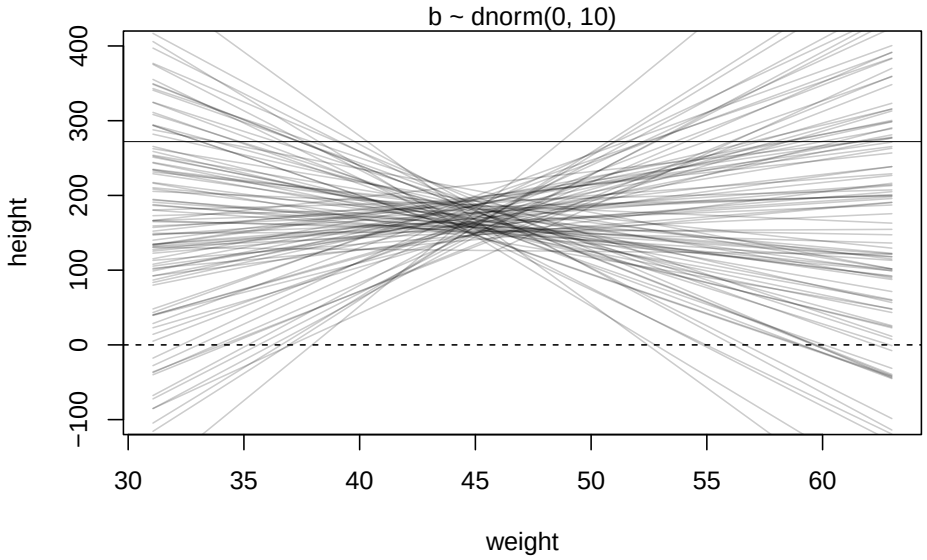
The linear model is quite popular in applied statistics and one reason is probably the rather straightforward interpretation of the coefficients.

### 3.1.2 Priors

We want to plot our priors to get a feeling what the model would predict without seeing the data. This is a kind of "sanity check" to see if the priors are reasonable.

```
set.seed(2971)
N <- 100  # 100 lines
a <- rnorm(N, 171.1, 20)
b <- rnorm(N, 0, 10)
```

```r
# Assume d2$weight is defined, e.g., using some dataset or simulation
xbar <- mean(d2$weight)

plot(NULL, xlim = range(d2$weight), ylim = c(-100, 400),
     xlab = "weight", ylab = "height")
abline(h = 0, lty = 2)  # horizontal line at 0
abline(h = 272, lty = 1, lwd = 0.5)  # horizontal line at 272
mtext("b ~ dnorm(0, 10)")

# Overlay the 100 lines
for (i in 1:N) {
  curve(a[i] + b[i] * (x - xbar),
        from = min(d2$weight), to = max(d2$weight),
        add = TRUE, col = col.alpha("black", 0.2))
}
```

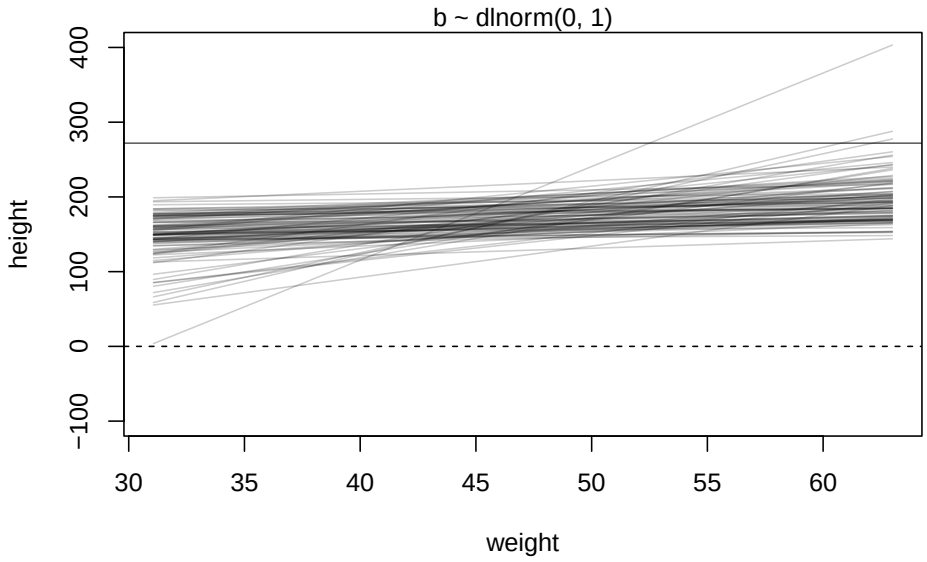



This relationship seems rather non-restrictive. According to our priors, one could see very steeply rising or falling lines. We could at least make the priors for the slope ($\beta$) non-negative. One possibility to do this is to use a log-normal distribution for the prior of $\beta$ which can only take non-negative values.

$$\beta \sim \text{Log-Normal}(0, 1)$$

Lets plot the priors again.

```r
set.seed(2971)
N <- 100  # 100 lines
a <- rnorm(N, 171.1, 20)
b <- rlnorm(N, 0, 1)

# Assume d2$weight is defined, e.g., using some dataset or simulation
xbar <- mean(d2$weight)

plot(NULL, xlim = range(d2$weight), ylim = c(-100, 400),
     xlab = "weight", ylab = "height")
abline(h = 0, lty = 2)  # horizontal line at 0
abline(h = 272, lty = 1, lwd = 0.5)  # horizontal line at 272
mtext("b ~ dlnorm(0, 1)")

# Overlay the 100 lines
for (i in 1:N) {
  curve(a[i] + b[i] * (x - xbar),
        from = min(d2$weight), to = max(d2$weight),
        add = TRUE, col = col.alpha("black", 0.2))
}
```




This seems definitely more realistic.

### 3.1.3 Fit model

Now, let's **estimate the posterior/fit the model** as before:

```
# load data again, since it's a long way back
library(rethinking)
data(Howell1)
d <- Howell1
d2 <- d[d$age >= 18, ]
xbar <- mean(d2$weight)
# fit model
mod <- quap(
    alist(
        height ~ dnorm(mu, sigma),
        mu <- a + b * (weight - xbar),
        a ~ dnorm(171.1, 100),
        b ~ dnorm(0, 10),
        sigma ~ dunif(0, 50)
    ) ,
data = d2)
```

Let's look at the **marginal distributions** of the parameters:

```
precis(mod)
```

```
##              mean         sd        5.5%       94.5%
## a     154.5972120 0.27033045 154.1651717 155.0292523
## b       0.9050131 0.04192754   0.8380048   0.9720214
## sigma   5.0718673 0.19115323   4.7663675   5.3773671
```

The analysis yields estimates for all our parameters of the model: $\alpha$, $\beta$ and $\sigma$. The estimates are the mean of the posterior distribution.

See exercise 2.

**Interpretation of** $\beta$: The mean of the posterior distribution of $\beta$ is 0.9. A person with a weight of 1 kg more weight can be expected to be 0.9 cm taller. A 96% credible interval for this estimate is $[0.83, 0.97]$. We can be quite sure that the slope is positive.

It might also be interesting to insept the variance-covariance matrix, respectively the correlation between the parameters as we did before in the intercept model.

```
diag(vcov(mod))
```

```
##           a           b       sigma
## 0.073078550 0.001757918 0.036539558
```

```
cov2cor(vcov(mod))
```

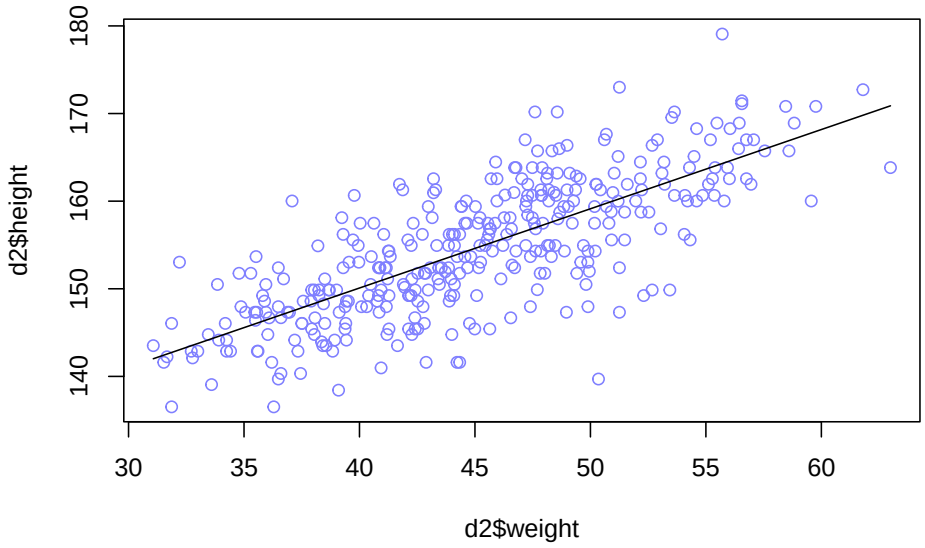
```
##                     a             b         sigma
## a       1.000000e+00 -9.591866e-10  3.330963e-05
## b      -9.591866e-10  1.000000e+00 -2.879607e-05
## sigma   3.330963e-05 -2.879607e-05  1.000000e+00
```

As we can see the correlations are near zero. Compare to the graphical display of the model structure. There is no connection.

### 3.1.4 Result

**Graphical end result** of fitting the model:

```
plot(d2$height ~ d2$weight, col = rangi2)
post <- extract.samples(mod)
a_quap <- mean(post$a)
b_quap <- mean(post$b)
curve(a_quap + b_quap * (x - xbar), add = TRUE)
```



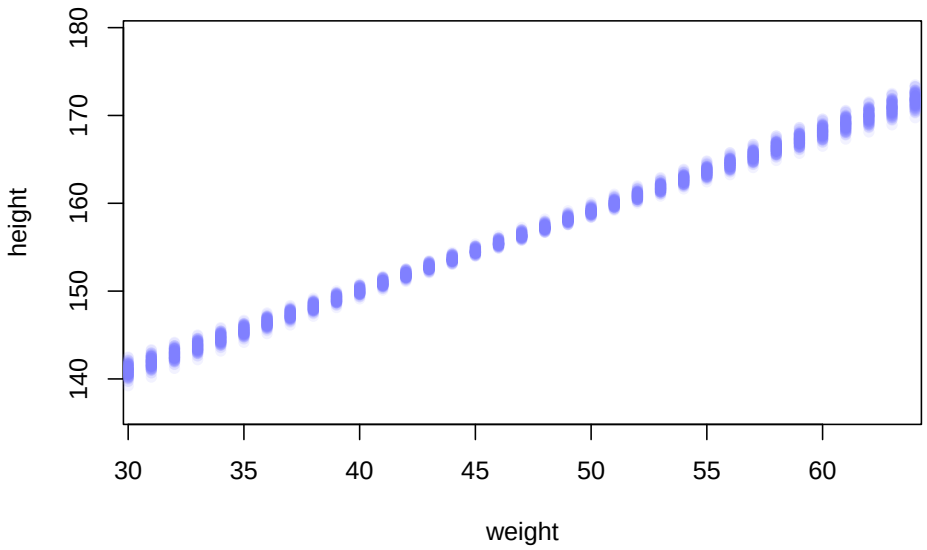

### 3.1.5 Credible bands

We could draw again and again from the posterior distribution and calculate the means like above. Plotting the regression lines with the respective parameters $\alpha$, $\beta$ would indicate the variability of the estimates.

```
# Define a sequence of weights for predictions
weight.seq <- seq(from = 25, to = 70, by = 1)

# Use the model to compute mu for each weight
mu <- link(mod, data = data.frame(weight = weight.seq))
str(mu)
```

```
##  num [1:1000, 1:46] 138 136 137 136 137 ...
```

```
# Visualize the distribution of mu values
plot(height ~ weight, d2, type = "n")  # Hide raw data with type = "n"

# Loop over samples and plot each mu value
for (i in 1:100) {
  points(weight.seq, mu[i, ], pch = 16, col = col.alpha(rangi2, 0.1))
}
```

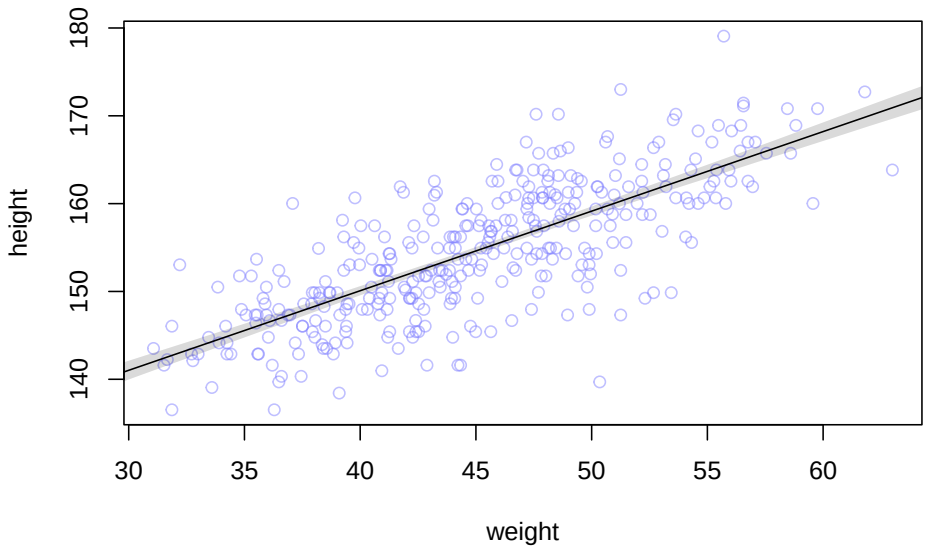



The `link` function fixes the weight at the values in `weight.seq` and draws samples from the posterior distribution of the parameters. We will do the analog thing in the frequentist framework.

We can also draw a nice shade for the regression line:

```
# Summarize the distribution of mu
mu.mean <- apply(mu, 2, mean)
mu.PI <- apply(mu, 2, PI, prob = 0.89)
plot(height ~ weight, d2, col = col.alpha(rangi2, 0.5))
```

```
lines(weight.seq, mu.mean)
shade(mu.PI, weight.seq)
```




As we can see, we are pretty sure about the mean of height which we wanted to model in the first place. Mean modeling is one thing, individual prediction is another. Given a certain weight of a person, what is the height of the same person? The first line in the model definition ($height_i \sim Normal(\mu_i, \sigma)$) tells us that a person's weight is distributed *around* the mean (which linearly depends on weight) and is not necessary the mean itself.

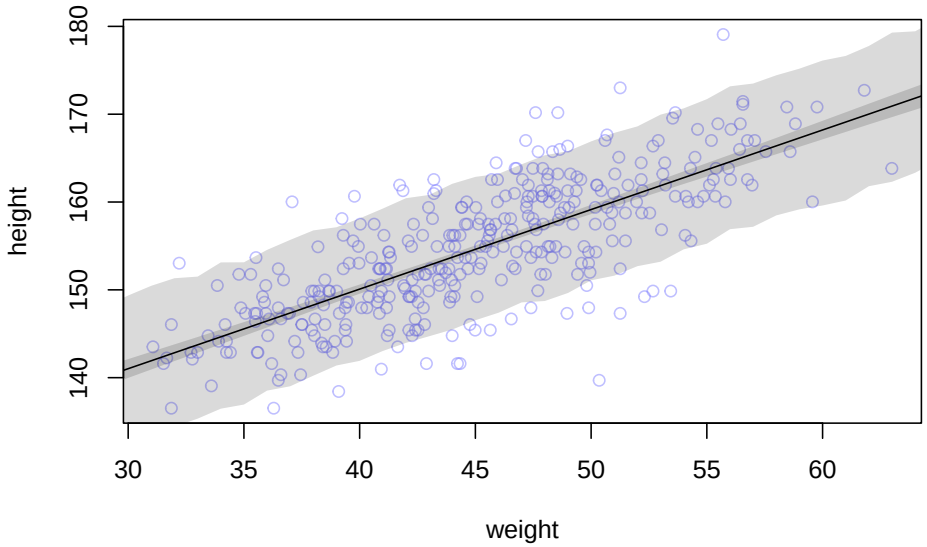
To get to an **individual prediction**, we need to consider the uncertainty of the parameter estimation *and* the uncertainty from the Gaussian distribution around the mean (at a certain weight). We do this with `sim`.

```
# Simulate heights from the posterior
sim.height <- sim(mod, data = list(weight = weight.seq))
str(sim.height)
```

```
##  num [1:1000, 1:46] 138 130 130 147 136 ...
```

```
# Compute the 89% prediction interval for simulated heights
height.PI <- apply(sim.height, 2, PI, prob = 0.89)

# Plot the raw data
plot(height ~ weight, d2, col = col.alpha(rangi2, 0.5))

# Draw MAP (mean a posteriori) line
```

```
lines(weight.seq, mu.mean)

# Draw HPDI (highest posterior density interval) region for mu
shade(mu.PI, weight.seq)

# Draw PI (prediction interval) region for simulated heights
shade(height.PI, weight.seq)
```




The lighter and wider shaded region is where the model expects to find 89% of the heights of a person with a certain weight.

This part is sometimes a bit desillusioning when seen for the first time: Draw a horizontal line at 150 cm and see how many weights (according to the individual prediction) are compatible with this height. Weights from 30 to 50 kg are compatible with this height according to the 89% prediction interval. The higher the credibility, the wider the interval, the wider the range of compatible weights (more than 60% of the weight-range).

```
(50 - 30) / (range(d2$weight)[2] - range(d2$weight)[1])
```

```
## [1] 0.6265362
```

### 3.1.6 Summary

- We have added a covariate (weight) to the simple mean model to predict height.

- We have centered the weight variable.
- We have defined and refined priors for the intercept and slope.
- We have estimated the posterior distribution of the parameters using quadratic approximation with `quap`.
- We have visualized the result.
- We have created credible bands for mean and individual predictions.

## 3.2 Simple Linear Regression in the Frequentist Framework

We will now do the same analysis in the frequentist framework while introducing some foundational theory along the way. I recommend reading the first couple of chapters from Westfall.

### 3.2.1 Model definition

Our linear model is defined as:

$$h_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

where

- $\varepsilon_i$ is the error term with $\varepsilon_i \sim N(0, \sigma), \forall i$
- $\beta_0$ is the unknown but fixed intercept
- $\beta_1$ is the unknown but fixed slope

#### 3.2.1.1 Model Assumptions of the Classical Regression Model (Westfall, 1.7):

The first and **most important assumption** is that the data are produced probabilistically, which is specifically stated as

$$Y|X = x \sim p(y|x)$$

What does this mean?

- $Y|X = x$ is the random variable Y **conditional** on X being equal to x, i.e. the distribution of $Y$ if we know the value of $X$ (in our example the weight in kg). This is a nice image of what is meant here.
- $p(y|x)$ is the distribution of potentially observable $Y$ given $X = x$. In our case above this was the normal distribution with mean $\mu_i$ and variance $\sigma$.

One always things about the so-called data generating process (Westfall, 1.2). How did the data come about? There is a process behind it and this process is attempted to be modeled.

Further assumptions:

- Correct functional specification: The conditional mean function $f(x) = \mathbb{E}(Y|X = x)$. In the case of the linear model, the assumption is $\mathbb{E}(Y|X = x) = \alpha+\beta x$. The expectation of $Y$ (height) depends linearly on $x$ (weight).
- The errors are homoscedastic (constant variance $\sigma$). This means the variances of all conditional distributions $p(y|x)$ are constant $(= \sigma^2)$.
- Normality. For the classical linear regression model all the conditional distributions $p(y|x)$ are normal distributions.
- The errors are independent of each other. The potentially observable $\varepsilon_i = Y_i - f(\mathbf{x_i}, )$ is uncorrelated with $\varepsilon_j = Y_j - f(\mathbf{x_j}, )$ for $i \neq j$.

These assumptions become clearer as we go along and should be checked for every model we fit. They are not connected, they can all be true or false. The question is not "Are the assumptions met?" since they never are exactly met. The question is **how** "badly" the assumptions are violated?

Remember, **all models are wrong, but some are useful**.

In full, the classical linear regression model can be written as:

$$Y_i|X_i = x_i \sim_{independent} N(\beta_0 + \beta_1 x_{i1} + ... \beta_k x_{ik}, \sigma^2)$$

for $i = 1, ..., n$.

### 3.2.2 Fit the model

Again, we fit the model using the least squares method. One has to minimize the sum of squared differences between the true heights and the model-predicted heights in order to find $\beta_0$ and $\beta_1$.

$$SSE(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2$$

We omit the technical details and give the results for $\beta_0$ and $\beta_1$:

$$\hat{\beta_0} = \bar{y} - (\hat{\beta_1}\bar{x}),$$
$$\hat{\beta_1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Let's use R again to solve the problem:

```
library(rethinking)
data(Howell1)
d <- Howell1
d2 <- d[d$age >= 18, ]
mod <- lm(height ~ weight, data = d2)
summary(mod)
```

```
##
## Call:
## lm(formula = height ~ weight, data = d2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.7464  -2.8835   0.0222   3.1424  14.7744
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 113.87939    1.91107   59.59   <2e-16 ***
## weight        0.90503    0.04205   21.52   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.086 on 350 degrees of freedom
## Multiple R-squared:  0.5696, Adjusted R-squared:  0.5684
## F-statistic: 463.3 on 1 and 350 DF,  p-value: < 2.2e-16
```

**Interpretation of R-output**:

- `Call`: The model that was fitted.
- `Residuals`: $r_i = height_i - \widehat{height}_i$. Difference between true heights and model-predicted heights.
- `Coefficients`: The estimated $\beta_0$ and $\beta_1$.
  - `Estimate`: The (least squares) estimated value of the coefficient.
  - `Std. Error`: The standard error of the estimate.
  - `t value`: The value of the $t$-statistic for the (Wald-) hypothesis test $H_0 : \beta_i = 0$.
  - `Pr(>|t|)`: The $p$-value of the hypothesis test.
- `Residual standard error`: The estimate of $\sigma$ which is also a model parameter (as in the Bayesian framework).
- `Multiple R-squared`: The proportion of the variance explained by the model (we will explain this below).

- `Adjusted R-squared`: A corrected version of the $R^2$ which takes into account the number of predictors in the model.
- `F-statistic`: The value of the $F$-statistic for the hypothesis test: $H_0$ : $\beta_1 = \beta_2 = \cdots = \beta_k = 0$. Note, the alternative hypotheses to this test is that *any* of the $\beta_i$ is not zero. If that Is the case, the model explains more than the mean model with just $\beta_0$.

### 3.2.3 ANOVA (Analysis of Variance)

A non-obvious and very useful finding is that the total variability in the data can be **decomposed** (or analysed) into two parts:

- The variability explained by the model (the regression line)
- The variability not explained by the model (the residuals)

Total sum of squares = Regression sum of squares + Residual sum of squares

$$SST = SSR + SSE$$

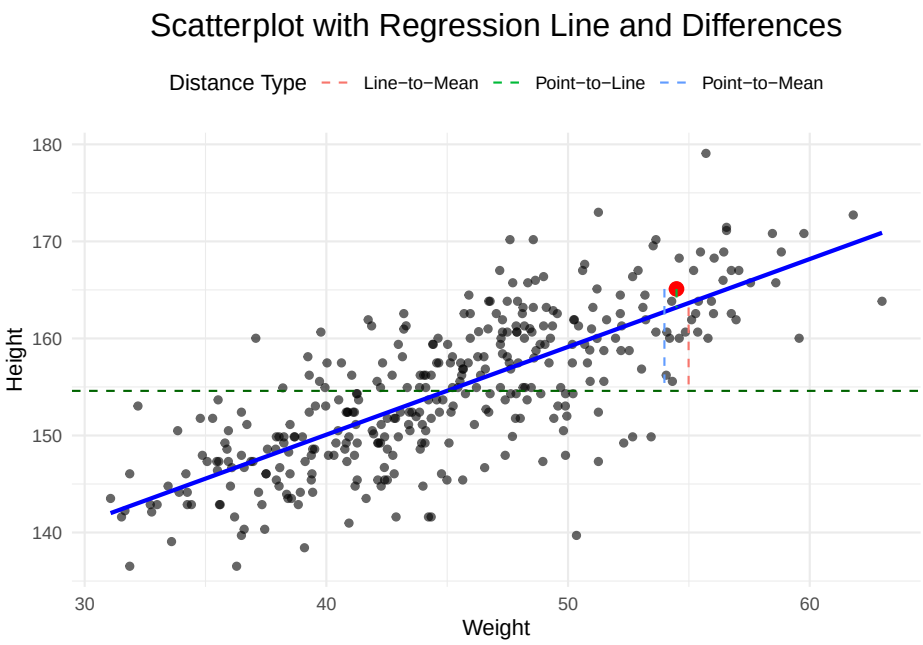
$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2$$

If you are interested in the details, check out this.

This video explains the concept nicely.

Let's visualize our regression result:

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




……. **TODOS**

- Conditional distribution approach
- Not the same to predict X with Y or vice versa.
- Analysis of Variance, R^2
- random X vs fixed X
- separability interpretration using R^2
- check assumptions of model
- interpret alpha and beta maybe using simulation

## 3.3 Exercises

### 3.3.1 Exercise 1

In the model from above:

$$
\begin{aligned}
h_i &\sim \text{Normal}(\mu_i, \sigma) \\
\mu_i &\sim \alpha + \beta(x_i - \bar{x}) \\
\alpha &\sim \text{Normal}(171.1, 20) \\
\beta &\sim \text{Normal}(0, 10) \\
\sigma &\sim \text{Uniform}(0, 50)
\end{aligned}
$$

- What ist the expected height when $x_i = \bar{x}$?
- What is the expected height when $x_i$ changes by 1 unit?

### 3.3.2 Exercise 2

Look at the marginal distrubutions of the parameters in the Bayesian model.

- Plot the posterior distribution of all 3 parameters.
- Include in the plot a 99% credible interval (HDI).