Gentle introduction to Linear Regression

Katarzyna Sornat

8 11 2020

Linear regression

Linear regression is used for finding linear relationship between target and one or more predictors. There are two types of linear regression- Simple (one explanatory variable/predictor/independent variable) and Multiple (more predictors).

Secret smoothies

Smoothie recipe

- 1 cup plain yogurt
- 1 secret ingredient
- random food dyes

Blend ingredients. Serve.

Predict calories

What's inside?

- 1 cup plain yogurt
- 1 secret ingredient
- random food dyes

Sardine smoothie!

What's inside?

- 1 cup plain yogurt
- 58 g sardines in tomato sauce
- random food dyes

233 Calories

Predict calories

What's inside?

- 1 cup plain yogurt
- 1 secret ingredient
- random food dyes

Waffle mix smoothie!

What's inside?

- 1 cup plain yogurt
- 56 g waffle mix
- random food dyes

301 Calories

Predict calories

What's inside?

- 1 cup plain yogurt
- 1 secret ingredient
- random food dyes

Predict the calories

Can we use something?

Predict the calories

Calories of 15 of our smoothies:

713, 288, 268, 252, 306, 301, 233, 150, 143, 416, 414, 155, 383, 258, 292

Predict the calories

Calories of 15 of our smoothies:

mean 304.8 median 288

Mascarpone smoothie!

What's inside?

- 1 cup plain yogurt
- 72 g Mascarpone
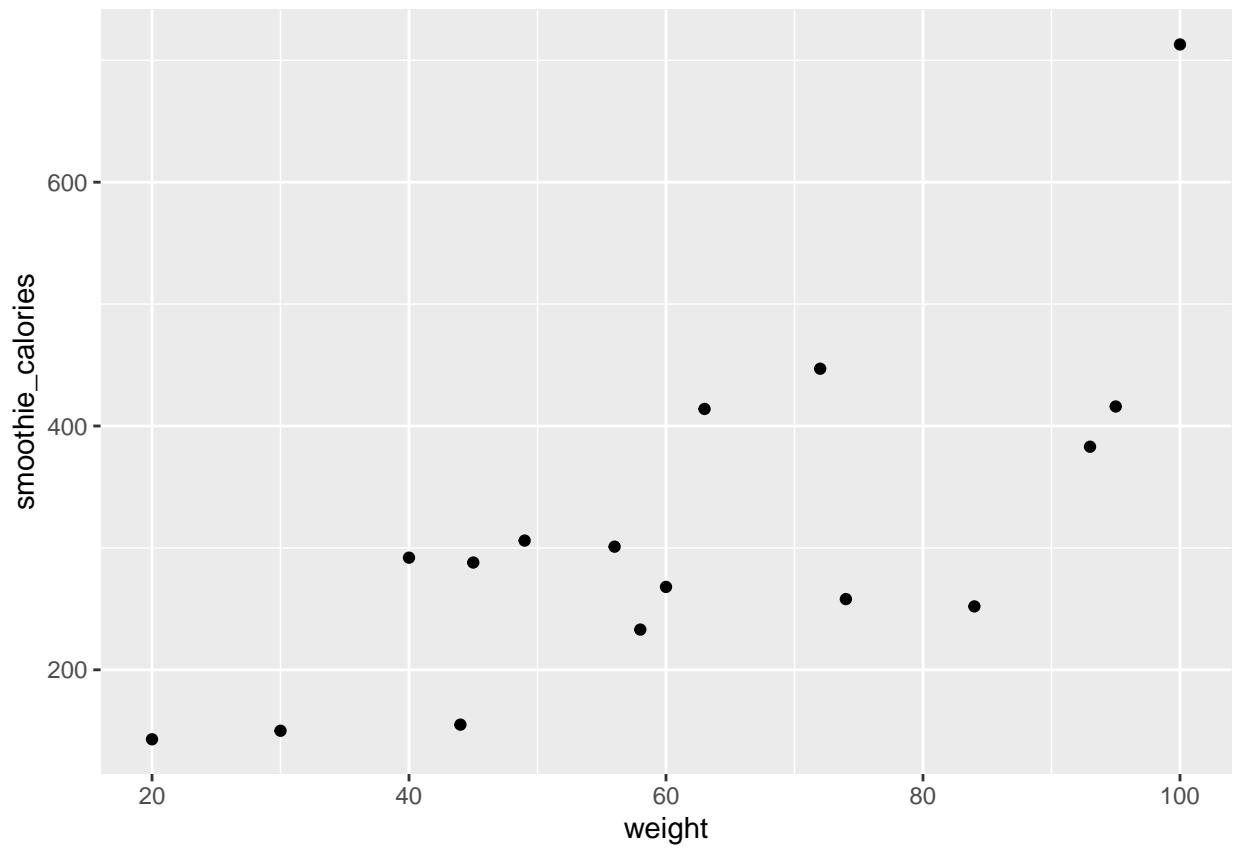- random food dyes

447 Calories

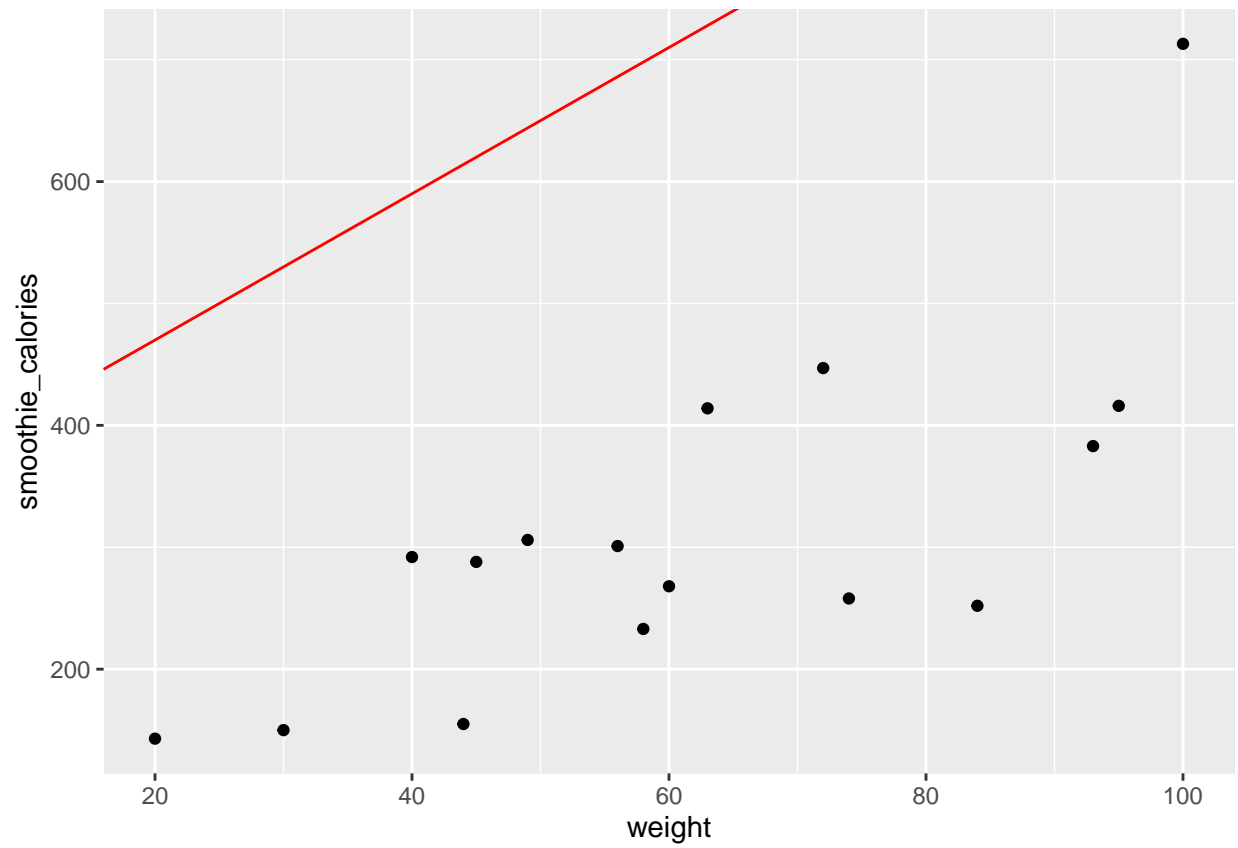**How did we do?**

Underestimate = 304.8

Can we do better?

Underestimate = 304.8
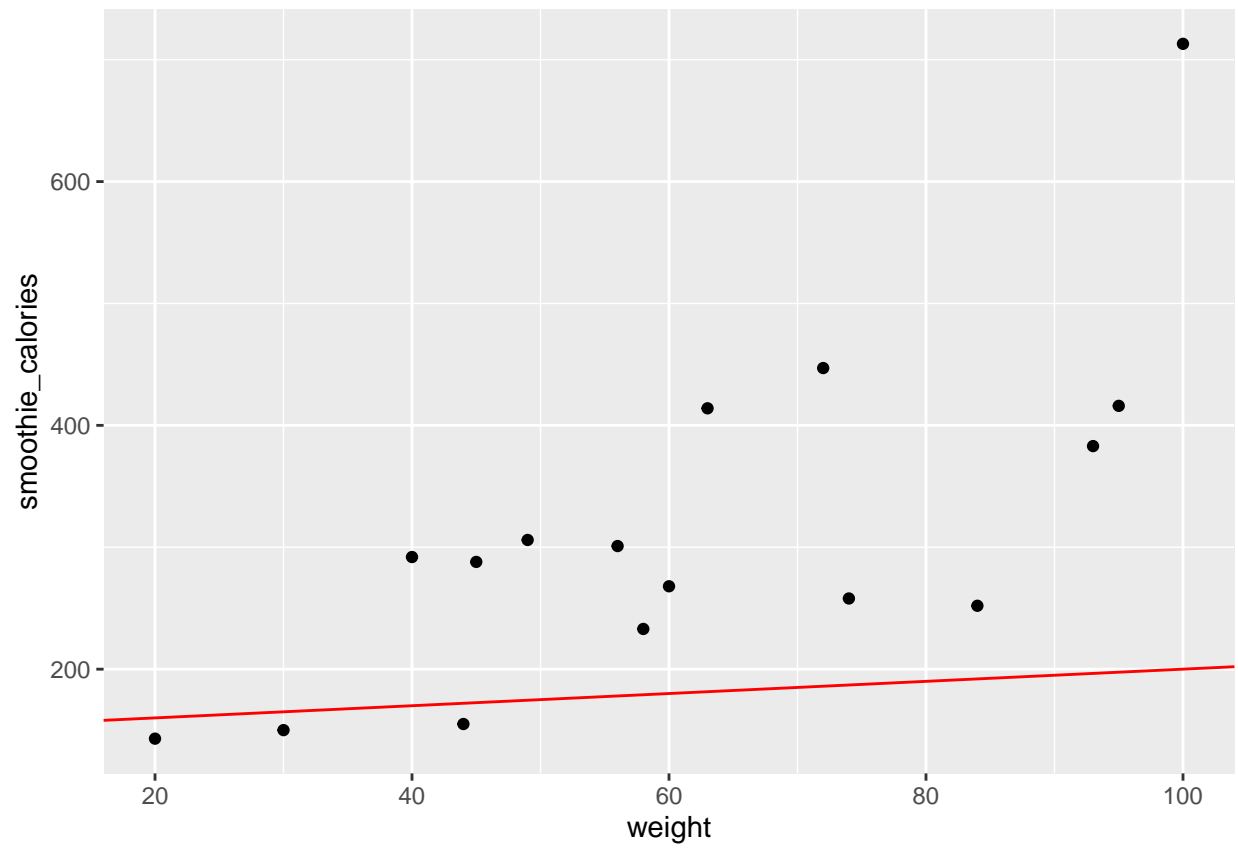
Use weight of the secret ingredient!
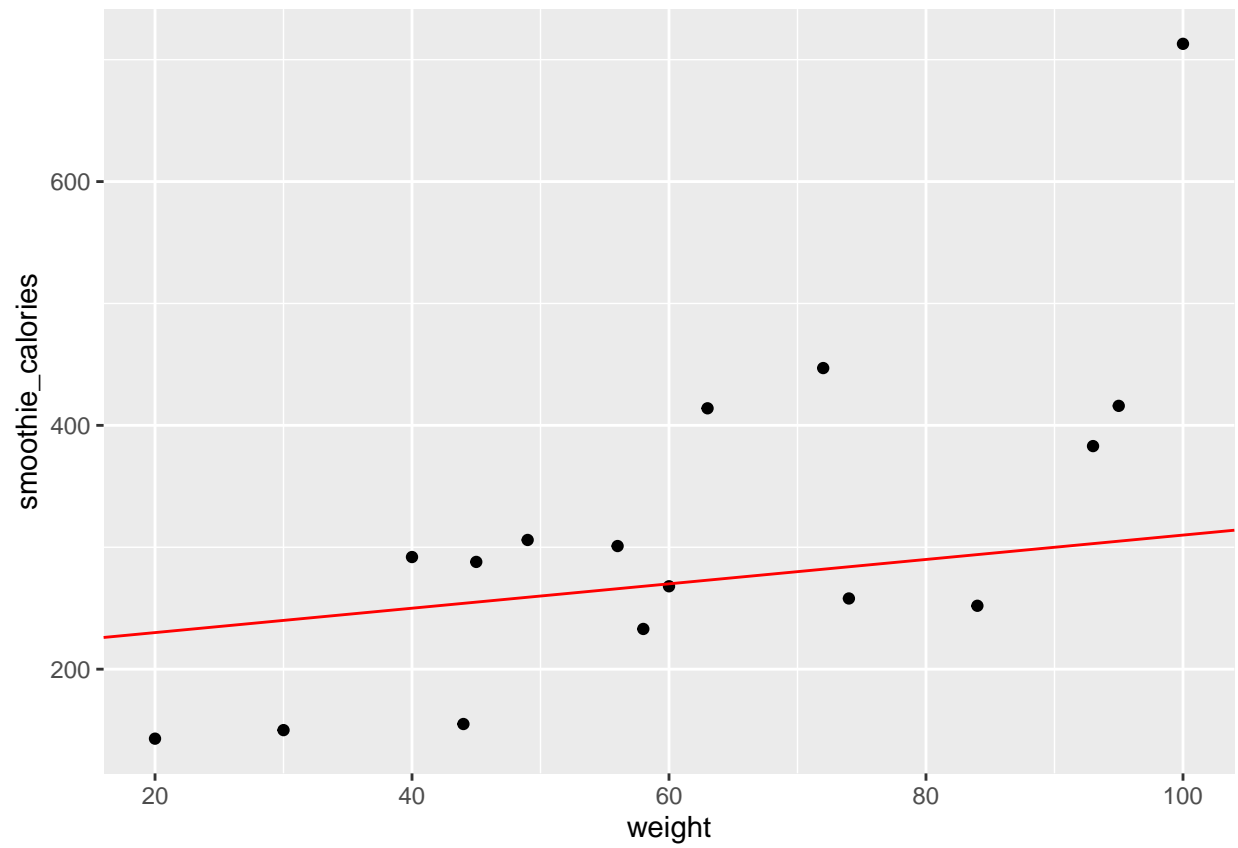
Predict the calories

Put A Line Through It!
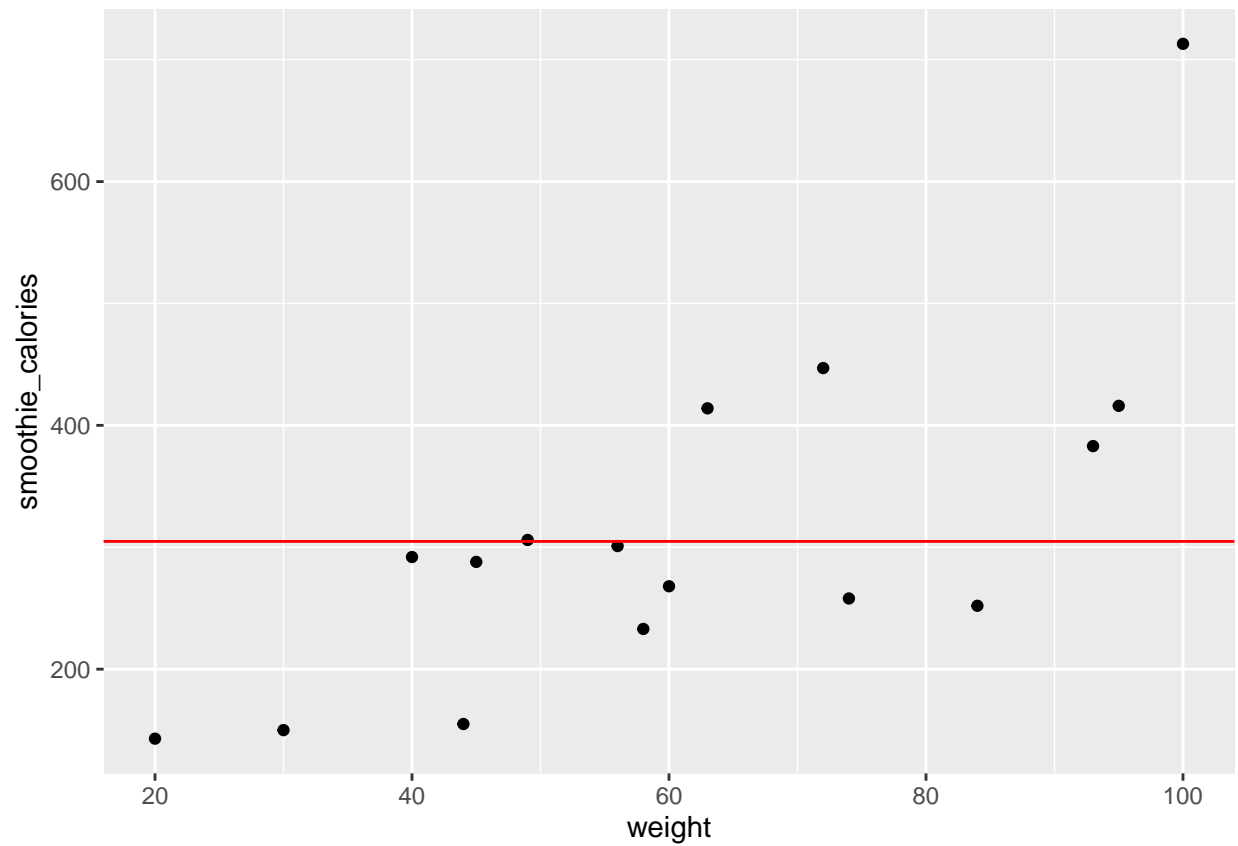

Smoothie_calories = Intercept + Slope * weight

Put A Line Through It!

Put A Line Through It!

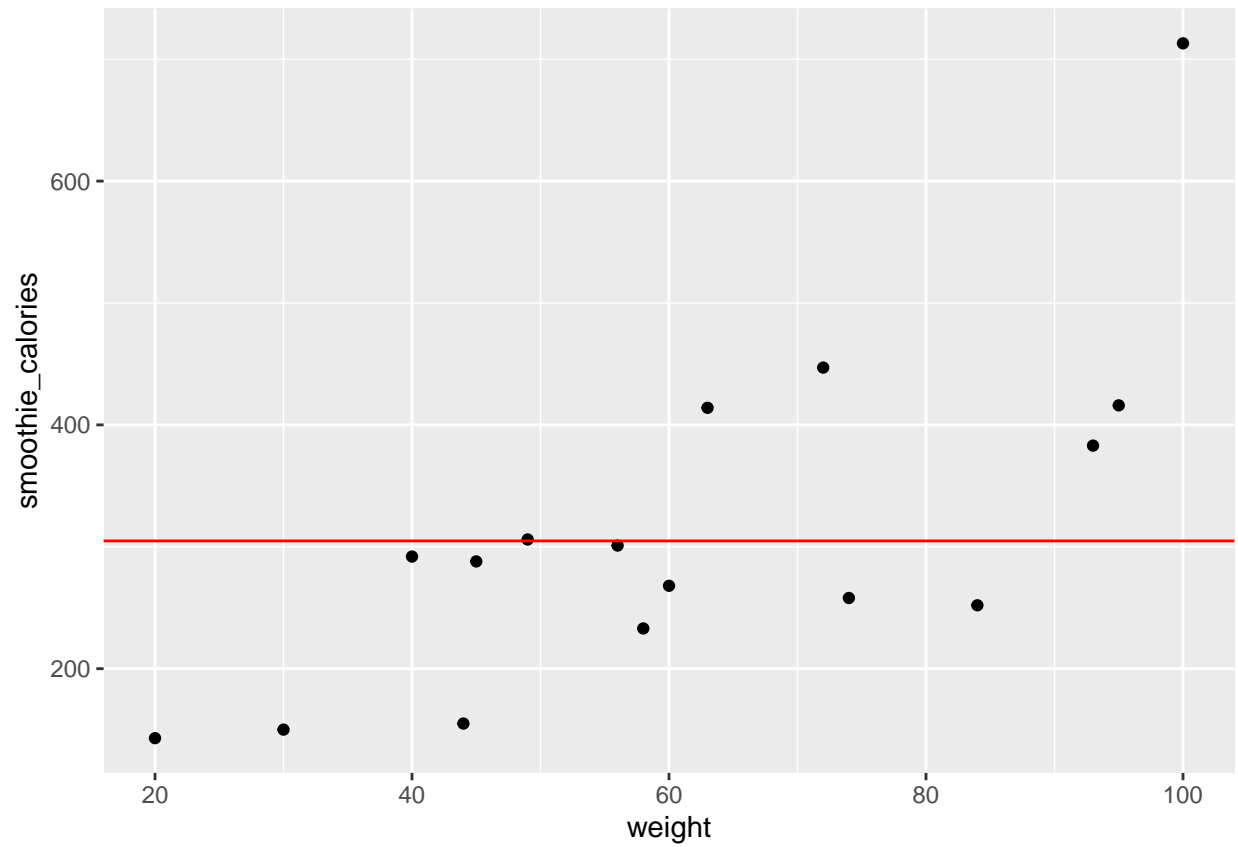**Put A Line Through It!**

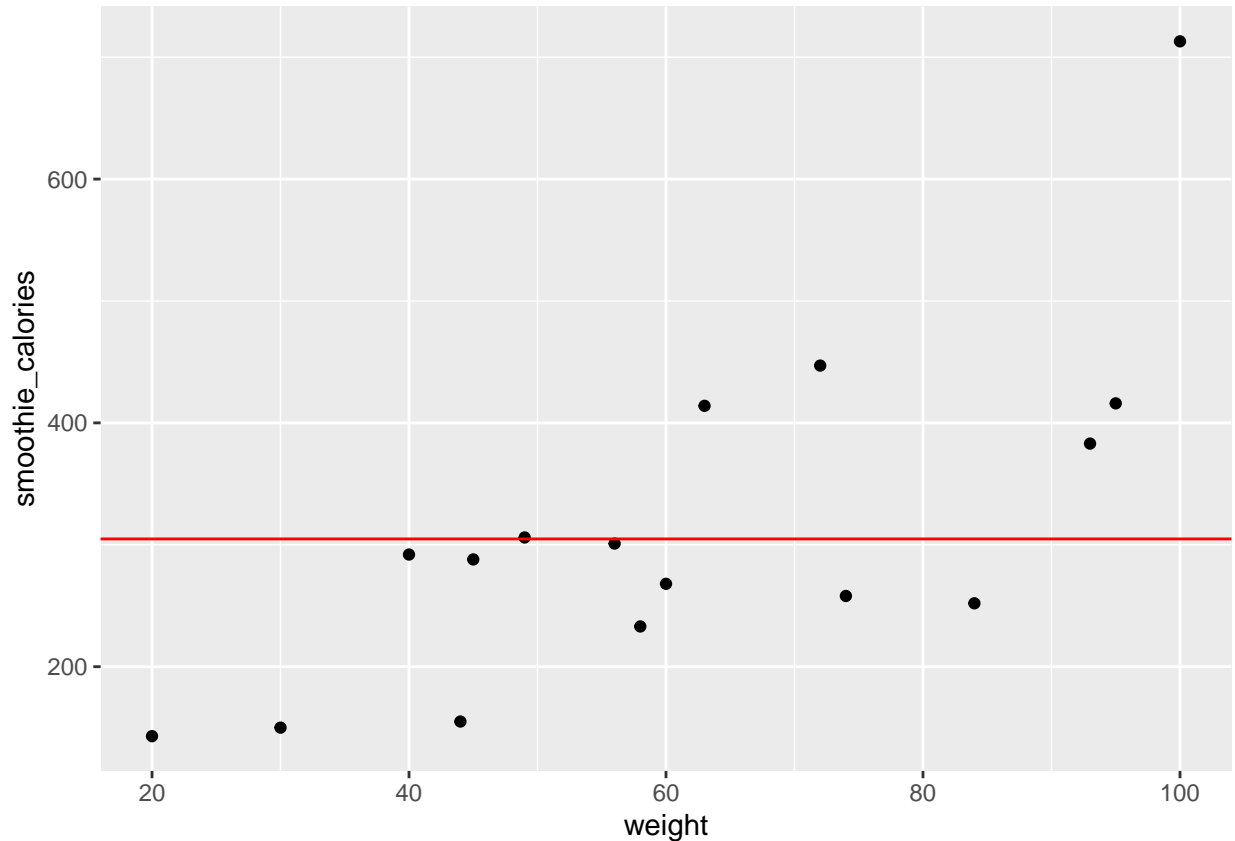This is our previous model!

We can do better!

Put A Line Through It!

Let's find a line that is as close points as possible!

Put A Line Through It!

Define error as point-line. If we want a line close to the points, we don't like errors.

Introduction - simple linear regression

Linear regression attempts to model the relationship between two variables by fitting a linear equation (= a straight line) to the observed data. One variable is considered to be an explanatory variable (e.g. your income), and the other is considered to be a dependent variable (e.g. your expenses)

Simple linear regression is useful for finding relationship between two continuous variables. It looks for statistical relationship but not deterministic relationship. Relationship between two variables is said to be deterministic if one variable can be accurately expressed by the other. For example, using temperature in degree Celsius it is possible to accurately predict Fahrenheit. Statistical relationship is not accurate in determining relationship between two variables. For example, relationship between height and weight.

Here's how we do linear regression 1. We plot our dependent variable (y-axis) against the independent variable (x-axis) 2. We try to plot a straight line and measure correlation 3. We keep changing the direction of our straight line until we get the best correlation 4. We extrapolate from this line to find new values on y-axis

How to find the best line

Why do we need assumptions?

Let's code sth!

Generalized Linear Model on Logistic Regression example

Generalized Addivitve Models

```
##      speed          dist
##  Min.   : 4.0   Min.   :  2.00
##  1st Qu.:12.0   1st Qu.: 26.00
##  Median :15.0   Median : 36.00
##  Mean   :15.4   Mean   : 42.98
##  3rd Qu.:19.0   3rd Qu.: 56.00
##  Max.   :25.0   Max.   :120.00
```

Including Plots

You can also embed plots, for example: