

# Linear Regression

Peter von Rohr

2023-02-27

# Goal

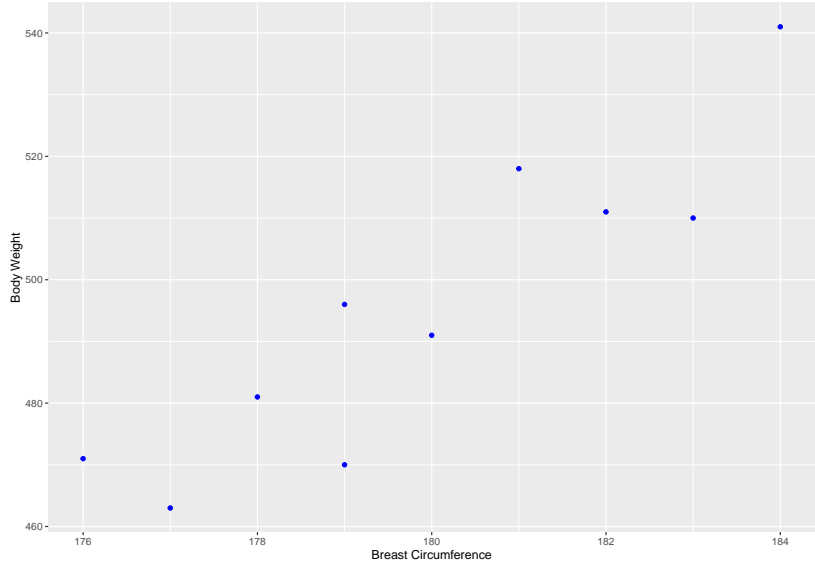
Assessment of relationship between

- ▶ a given variable (response) and
- ▶ other measurements or observations (predictors) on the same animal

## Example

Animal	Breast Circumference	Body Weight
1	176	471
2	177	463
3	178	481
4	179	470
5	179	496
6	180	491
7	181	518
8	182	511
9	183	510
10	184	541

# Diagram

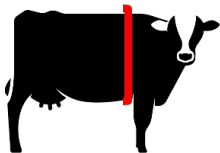


# Observations

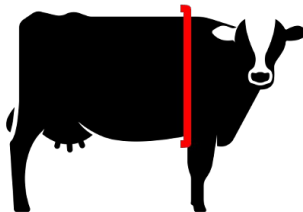
- ▶ relationship between breast circumference and body weight: heavier animals tend to have larger values for breast circumference
- ▶ same relationship across whole range → **linear** relationship

# Regression Model

- ▶ quantify relationship between body weight and breast circumference
- ▶ practical application: measure band for animals



Created by Agniraj Chatterji  
from Noun Project



Created by Agniraj Chatterji  
from Noun Project

# Model Building

- ▶ expected body weight ( $E(y)$  in kg) based on an observed value of  $x$  cm for breast circumference

$$E(y) = b_0 + b_1 * x$$

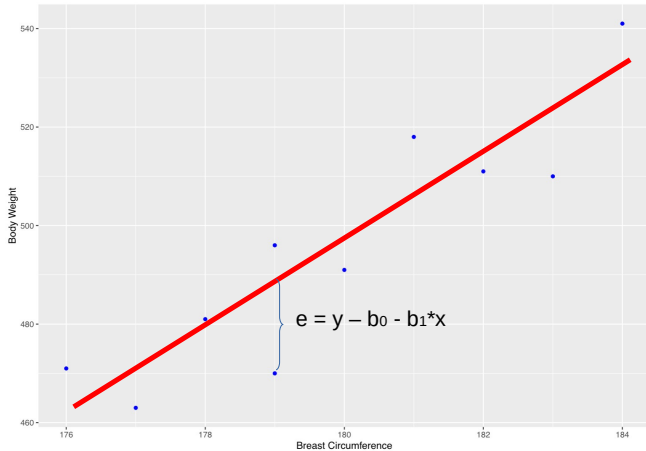
- ▶  $b_0$  and  $b_1$  are unknown parameters of the model
- ▶ model is linear function of parameters  $\rightarrow$  linear model

# Parameter Estimation

- ▶ How to find values for  $b_0$  and  $b_1$
- ▶ several techniques available: start with Least Squares



# Least Squares



# Estimators

Find values  $\hat{b}_0$  and  $\hat{b}_1$  such that

$$\mathbf{e}^T \mathbf{e} = \sum_{i=1}^N e_i^2 = \sum_{i=1}^N [y_i - E(e_i)]^2 = \sum_{i=1}^N [y_i - b_0 - b_1 * x_i]^2$$

is minimal

## Minimization

$$\begin{aligned}\frac{\partial \mathbf{e}^T \mathbf{e}}{\partial b_0} &= -2 \sum_{i=1}^N [y_i - b_0 - b_1 x_i] \\ &= -2 \left[ \sum_{i=1}^N y_i - Nb_0 - b_1 \sum_{i=1}^N x_i \right]\end{aligned}$$

$$\begin{aligned}\frac{\partial \mathbf{e}^T \mathbf{e}}{\partial b_1} &= -2 \sum_{i=1}^N x_i [y_i - b_0 - b_1 x_i] \\ &= -2 \left[ \sum_{i=1}^N x_i y_i - b_0 \sum_{i=1}^N x_i - b_1 \sum_{i=1}^N x_i^2 \right]\end{aligned}$$

# Notation

$$x_{\cdot} = \sum_{i=1}^N x_i$$

$$y_{\cdot} = \sum_{i=1}^N y_i$$

$$(x^2)_{\cdot} = \sum_{i=1}^N x_i^2$$

$$(xy)_{\cdot} = \sum_{i=1}^N x_i y_i$$

$$\bar{x}_{\cdot} = \frac{x_{\cdot}}{N}$$

$$\bar{y}_{\cdot} = \frac{y_{\cdot}}{N}$$

## Solutions

$$\hat{b}_0 = \bar{y}_{\cdot} - \hat{b}_1 \bar{x}_{\cdot}.$$

and

$$\hat{b}_1 = \frac{(xy)_{\cdot} - N\bar{x}_{\cdot}\bar{y}_{\cdot}}{(x^2)_{\cdot} - N\bar{x}_{\cdot}^2}$$

# The General Case

Height as additional observation

Animal	Breast Circumference	Body Weight	Height
1	176	471	161
2	177	463	121
3	178	481	157
4	179	470	165
5	179	496	136
6	180	491	123
7	181	518	163
8	182	511	149
9	183	510	143
10	184	541	130

## Extended Model

Height is taken as additional predictor variable

$$E(y) = b_0 + b_1x_1 + b_2x_2$$

$$y_i = b_0 + b_1x_{i1} + b_2x_{i2} + e_i$$

→ additional parameter  $b_2$

## Matrix-Vector Notation

$$\mathbf{X} = \begin{bmatrix} x_{10} & x_{11} & x_{12} \\ x_{20} & x_{21} & x_{22} \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ x_{N0} & x_{N1} & x_{N2} \end{bmatrix}, \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_N \end{bmatrix}, \mathbf{e} = \begin{bmatrix} e_1 \\ e_2 \\ \cdot \\ \cdot \\ \cdot \\ e_N \end{bmatrix} \text{ and } \mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix}$$

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}, \text{ with } E(\mathbf{y}) = \mathbf{X}\mathbf{b}$$



# Parameter Estimate

Minimize

$$\begin{aligned}\mathbf{e}^T \mathbf{e} &= [\mathbf{y} - E(\mathbf{y})]^T [\mathbf{y} - E(\mathbf{y})] \\ &= [\mathbf{y} - \mathbf{X}\mathbf{b}]^T [\mathbf{y} - \mathbf{X}\mathbf{b}] \\ &= \mathbf{y}^T \mathbf{y} - 2\mathbf{b}^T \mathbf{X}^T \mathbf{y} + \mathbf{b}^T \mathbf{X}^T \mathbf{X} \mathbf{b}\end{aligned}$$

Normal equation

$$\mathbf{X}^T \mathbf{X} \hat{\mathbf{b}} = \mathbf{X}^T \mathbf{y}$$

## Solution

Provided  $(\mathbf{X}^T \mathbf{X})$  can be inverted

$$\hat{\mathbf{b}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$