Linear Regression

Peter von Rohr

2023-02-27

Goal

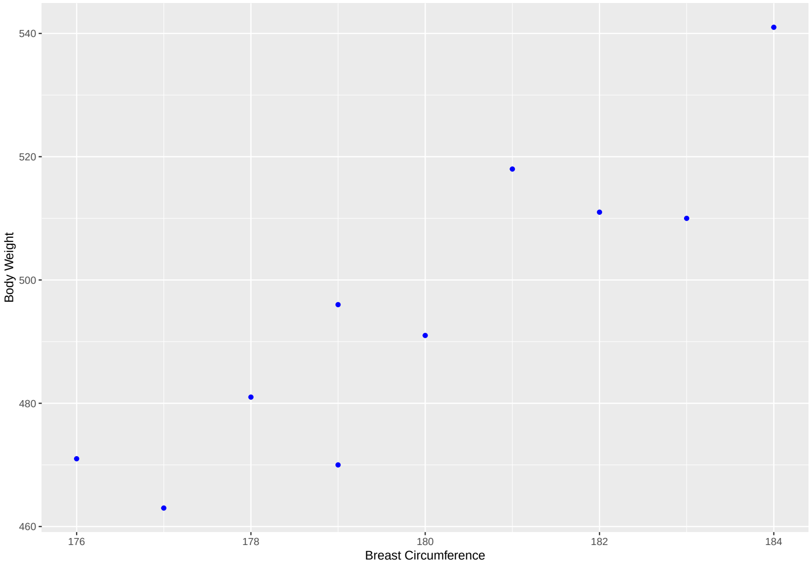
Assessment of relationship between

- 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

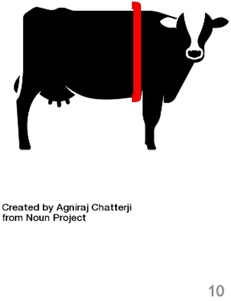

Body Weight
540
520
500
480
460
176
178
180
182
184
Breast Circumference

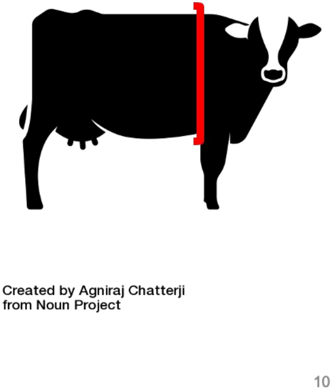
Observations

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

Regression Model

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





Model Building

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

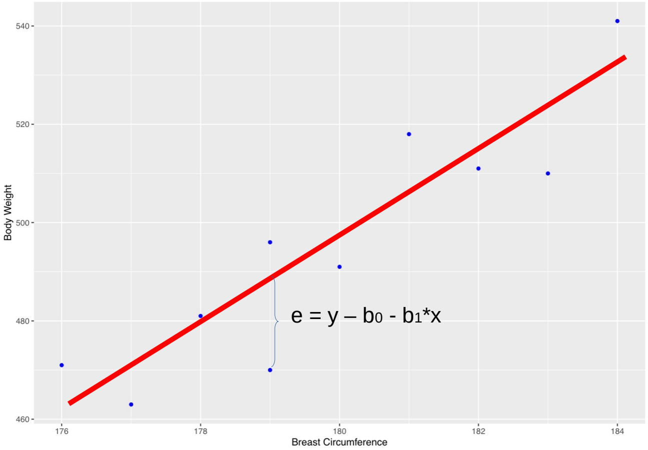
$$E(y) = b_0 + b_1 * x$$

- 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

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

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

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

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

and

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

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

$$E(y) = b_0 + b_1 x_1 + b_2 x_2$$

$$y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + e_i$$

$\rightarrow$ additional parameter b_2

Matrix-Vector Notation

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

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

Parameter Estimate

Minimize

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

Compute the gradient $\frac{\partial\ \mathbf{e}^T\mathbf{e}}{\mathbf{b}}$, set it to $\mathbf{0}$ to get the normal equations

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

Solution

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

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

Obtain Parameter Estimates in R

- Computations are tedious
- Use R builtin functions
- Assuming data is available in dataframe `tbl_reg` with columns `Body Weight` and `Breast Circumference`

```
lm_bw_bc <- lm(`Body Weight` ~ `Breast Circumference`,
               data = tbl_reg)
summary(lm_bw_bc)
```

Multiple Linear Regression Model

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

- General case with k x-variables

$$\mathbf{X} = \begin{bmatrix} x_{10} & x_{11} & . & x_{1k} \\ x_{20} & x_{21} & . & x_{2k} \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ x_{N0} & x_{N1} & . & x_{Nk} \end{bmatrix}, \ \mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ . \\ . \\ b_k \end{bmatrix}$$

Random Error Terms

- Properties of random error terms in vector $\mathbf{e}$

$$E(\mathbf{e}) = \mathbf{0}$$

$$var(\mathbf{e}) = E\left[\mathbf{e} - E(\mathbf{e})\right]\left[\mathbf{e} - E(\mathbf{e})\right]^T = E(\mathbf{e}\mathbf{e}^T) = \sigma^2\mathbf{I}_N$$

Least Squares Estimates

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

- Setting

$$\frac{\partial \mathbf{e}^T\mathbf{e}}{\partial \mathbf{b}} = \mathbf{0}$$

- yields least squares normal equations

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

Solution for Least Squares Estimators

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