Linear regression

Topics we'll cover

1. Regression with multiple predictor variables
2. Least-squares regression
3. The least-squares solution

Diabetes study

Data from $n = 442$ diabetes patients.

For each patient:

- 10 features $x = (x_1, \ldots, x_{10})$

  age, sex, body mass index, average blood pressure, and six blood serum measurements.

- A real value y: the progression of the disease a year later.

Regression problem:

- **response** $y \in \mathbb{R}$
- **predictor variables** $x \in \mathbb{R}^{10}$

Least-squares regression

Linear function of 10 variables: for $x \in \mathbb{R}^{10}$,

$$f(x) \;=\; w_1 x_1 + w_2 x_2 + \cdots + w_{10} x_{10} + b \;=\; w \cdot x + b$$

where $w = (w_1, w_2, \ldots, w_{10})$.

Penalize error using **squared loss** $(y - (w \cdot x + b))^2$.

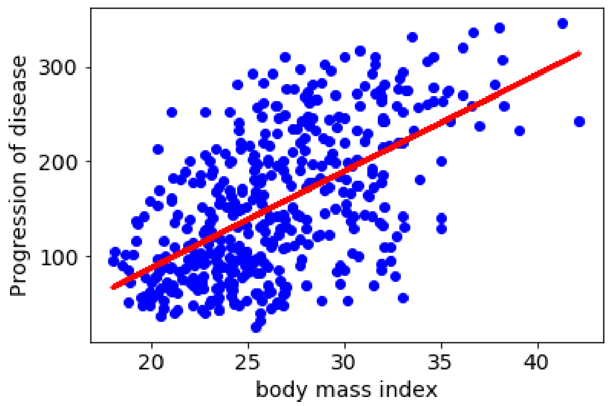
Least-squares regression:

- *Given:* data $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)}) \in \mathbb{R}^d \times \mathbb{R}$
- *Return:* linear function given by $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$
- *Goal:* minimize the **loss function**

$$L(w, b) = \sum_{i=1}^{n} (y^{(i)} - (w \cdot x^{(i)} + b))^2 \qquad .$$

Back to the diabetes data

- No predictor variables: mean squared error (MSE) = 5930
- One predictor ('bmi'): MSE = 3890



- Two predictors ('bmi', 'serum5'): MSE = 3205
- All ten predictors: MSE = 2860

Least-squares solution 1

Linear function of d variables given by $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$:

$$f(x) = w_1 x_1 + w_2 x_2 + \cdots + w_d x_d + b = w \cdot x + b$$

Assimilate the intercept b into w:

- Add a new feature that is identically 1: let $\widetilde{x} = (1, x) \in \mathbb{R}^{d+1}$

$$\begin{pmatrix} 4 & 0 & 2 & \cdots & 3 \end{pmatrix} \implies \begin{pmatrix} 1 & 4 & 0 & 2 & \cdots & 3 \end{pmatrix}$$

- Set $\widetilde{w} = (b, w) \in \mathbb{R}^{d+1}$
- Then $f(x) = w \cdot x + b = \widetilde{w} \cdot \widetilde{x}$

Goal: find $\widetilde{w} \in \mathbb{R}^{d+1}$ that minimizes

$$L(\widetilde{w}) = \sum_{i=1}^{n} (y^{(i)} - \widetilde{w} \cdot \widetilde{x}^{(i)})^2$$

Least-squares solution 2

Write

$$X = \begin{pmatrix} \longleftarrow & \widetilde{x}^{(1)} & \longrightarrow \\ \longleftarrow & \widetilde{x}^{(2)} & \longrightarrow \\ & \vdots & \\ \longleftarrow & \widetilde{x}^{(n)} & \longrightarrow \end{pmatrix}, \quad y = \begin{pmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(n)} \end{pmatrix}$$

Then the loss function is

$$L(\widetilde{w}) = \sum_{i=1}^{n} (y^{(i)} - \widetilde{w} \cdot \widetilde{x}^{(i)})^2 = \|y - X\widetilde{w}\|^2$$

and it minimized at $\widetilde{w} = (X^T X)^{-1}(X^T y)$.