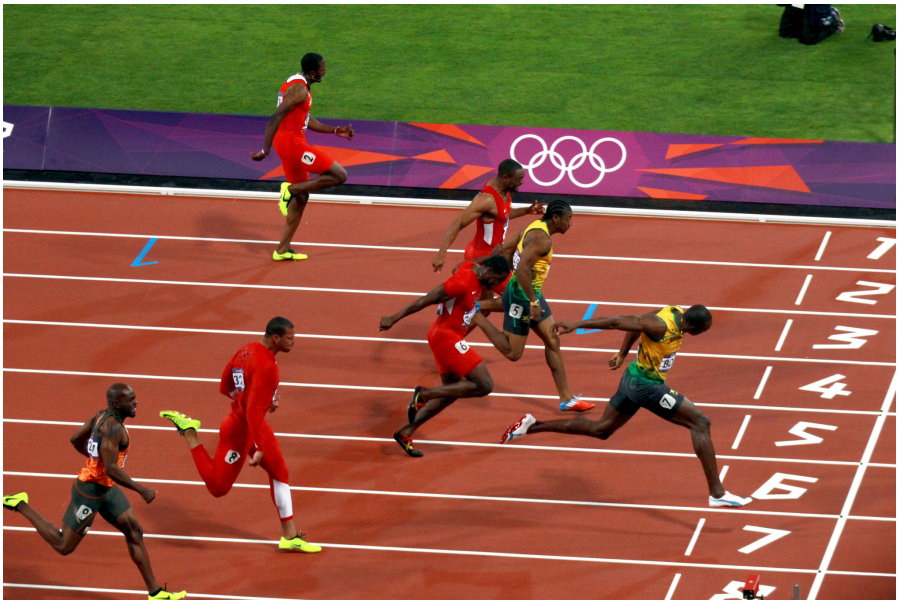
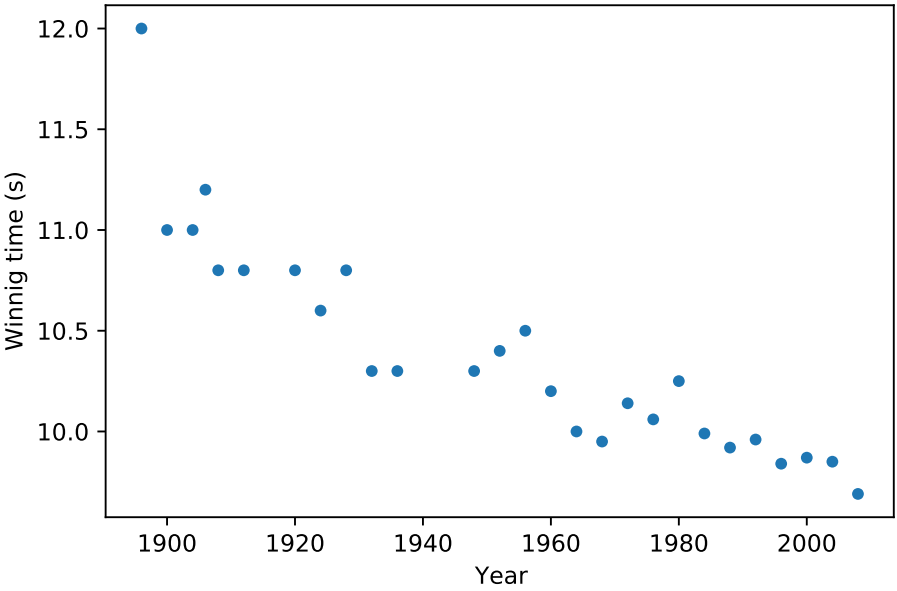
Simple linear regression

Herman Kamper

2023-03



Winning 100-metre men's Olympic time from 1896 to 2008



Missing years: 1914, 1940, 1944

Given the data that do have, could we predict what the winning times would have been for those missing years?

And could we predict the winning time for 2012, the year just following the data?

The model

A simple linear regression model predicts the output as a linear function of the input feature x:

$$f(x; w_0, w_1) = w_0 + w_1 x$$

We refer to w_0 and w_1 as the *parameters* of the model.

To choose w_0 and w_1, we are given a dataset of previous input-output measurements:

$$\left\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots, (x^{(N)}, y^{(N)})\right\}$$

I will sometimes just write this as:

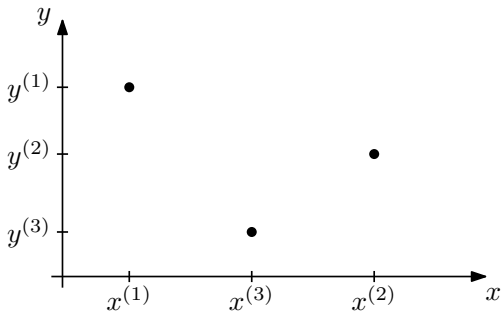
$$\left\{(x^{(n)}, y^{(n)})\right\}_{n=1}^{N}$$

In our example, each of the N points would correspond to a year x with the corresponding winning time for that year y.

How do we choose w_0 and w_1 based on the data? We need some way to measure the "goodness" or "badness" of the parameters, given the data.

The loss function

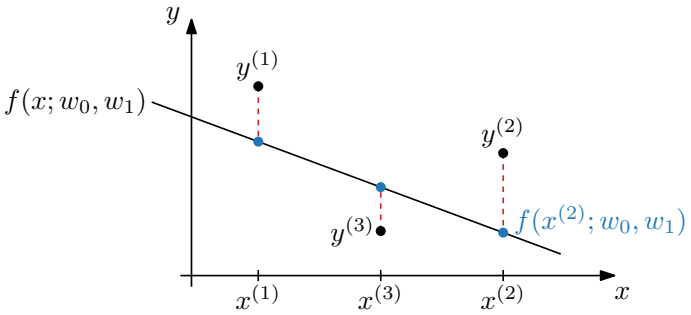
(Sometimes also called the *cost function*.)



How "good" is the fit of w_0 and w_1 to the data?

$$J(w_0, w_1) = \sum_{n=1}^{N} \left(y^{(n)} - f(x^{(n)}; w_0, w_1) \right)^2$$

This is called the *squared loss* (or the *squared error loss* or the *least squares criterion* or the *residual sum of squares*).

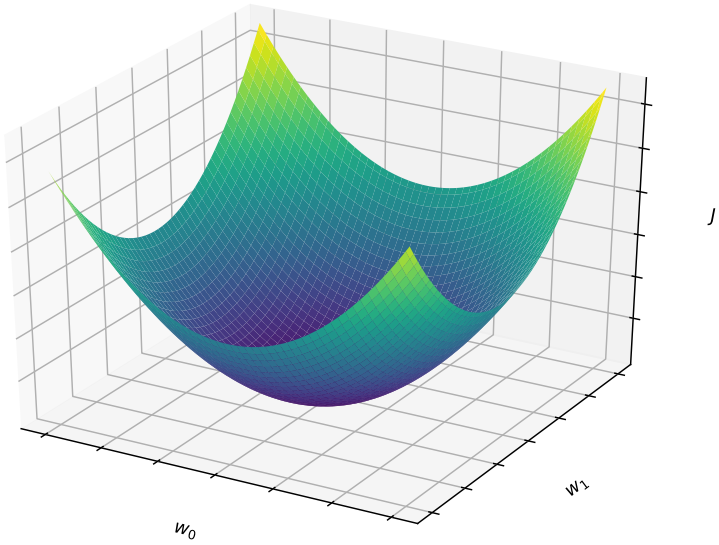


Optimisation

We want to find the w_0 and w_1 that minimise the loss $J(w_0, w_1)$:

$$\hat{w}_0, \hat{w}_1 = \arg\min_{w_0, w_1} J(w_0, w_1)$$

The loss J as a function of w_0 and w_1:



Strategy: Set $\frac{\partial J}{\partial w_0} = 0$ and $\frac{\partial J}{\partial w_1} = 0$

Expand the loss:

$$J(w_0, w_1) = \sum_{n=1}^{N} \left(y^{(n)} - f(x^{(n)}; w_0, w_1) \right)^2$$

$$=$$

Take the partial derivative of the loss with respect to w_0 and set it equal to 0:

$$w_0 = \frac{1}{N}\sum_{n=1}^{N} y^{(n)} - w_1 \frac{1}{N}\sum_{n=1}^{N} x^{(n)}$$

$$\hat{w}_0 =$$

Next take the partial derivative of the loss with respect to w_1:

$$\frac{\partial J}{\partial w_1} = \sum_{n=1}^{N} \frac{\partial}{\partial w_1} \left(y^{(n)} - (w_0 + w_1 x^{(n)})\right)^2$$

$$=$$

$$= \sum_{n=1}^{N} 2\left(y^{(n)} - \bar{y} - w_1(x^{(n)} - \bar{x})\right)(-1)(x^{(n)} - \bar{x})$$

Set $\frac{\partial J}{\partial w_1} = 0$:

$$\hat{w}_1 = \frac{\sum_{n=1}^{N}(y^{(n)} - \bar{y})(x^{(n)} - \bar{x})}{\sum_{n=1}^{N}(x^{(n)} - \bar{x})^2}$$

A convex loss

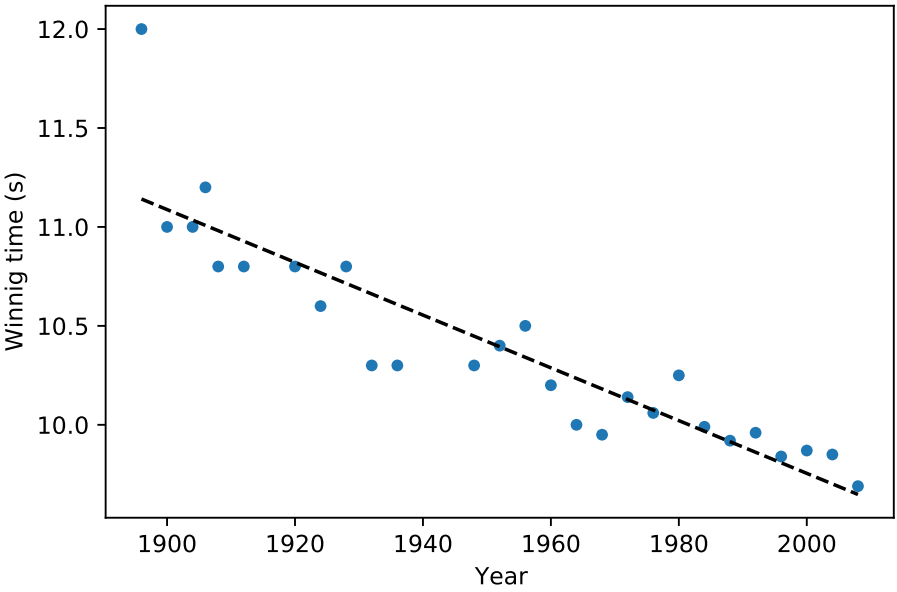
Technically, setting the partial derivatives to zero isn't enough: this could have been a maximum instead of a minimum! We should also show that

$$\frac{\partial^2 J}{\partial w_0^2} > 0 \quad \text{and} \quad \frac{\partial^2 J}{\partial w_1^2} > 0$$

As an exercise, show that this is true for the simple linear regression model.

In general, if the above property holds for all parameters, we call the loss a *convex loss function*. Later on we will also look at strategies for optimising loss functions which are non-convex.

Returning to our example: Model fit



- Estimated winning time in 1914: 10.901 s
- Estimated winning time in 2012: 9.595 s (actual time: 9.63 s)
- Estimated winning time in 2592: 1.863 s

Videos covered in this note

- Linear regression 1: Simple linear regression (14 min)

Reading

- ISLR 3 intro
- ISLR 3.1 intro
- ISLR 3.1.1

Acknowledgements

The motivating example from the men's 100-metre dash as well as much of the mathematical development are based on content from the textbook by Rogers and Girolami (2017).