Linear regression notes

John Bogovic

2019 July

1 Review

To be multiplied, matrices need to have the same "inner" dimensions: the number of columns of the left matrix must equal the number of rows of the right matrix.

$$\boldsymbol{A}_{L\times M}\boldsymbol{B}_{M\times N} = \boldsymbol{X}_{L\times N} \tag{1}$$

but

$$\boldsymbol{A}_{L\times M}\boldsymbol{B}_{N\times M} = \text{nothing, nonsense} \tag{2}$$

This works:

$$\boldsymbol{A}_{L\times M}\boldsymbol{B}^{T}_{N\times M} = \boldsymbol{X}_{L\times N} \tag{3}$$

1.1 Exercise

If $\boldsymbol{A}_{M\times N}$, what size is $A^T A$?

1. $N \times N$
2. $M \times N$
3. $M \times M$
4. Stupid question, because you can't multiply them.

2 Simple linear regression (fit a 1d line)

2.1 The problem

We are given the value of a variable, x, we would like to predict the value of another variable y. You will often see x called the "independent variable",

and y called the "dependent variable". We have many pairs of observations:

$$\begin{array}{l} x_1, y_1 \\ x_2, y_2 \\ \vdots \\ x_N, y_N \end{array} \tag{4}$$

Linear regression (1d) does this by finding the linear function that gives the best predictions. The functions we have to consider are:

$$\hat{y} = ax + b \tag{5}$$

Where we wrote $\hat{y}$ instead of y to indicate that it is an estimate, or prediction, and not the true value of y for the given x. Another way to think of the task is that we need to find the values a and b that give us the best results. The values a and b are called "parameters" of the function. How do we measure how good the predictions are?

2.2 "Cost function" - measuring goodness of the prediction

The most common way is to use the "sum of squared differences" (SSD) also called "sum of squared errors" (SSE), or"residual sum of squares." It is computed like this:

$$SSD(a, b) = \sum_i (y_i - (ax_i + b))^2. \tag{6}$$

Notice that $ax_i + b$ is the value of y predicted by our function for the input x_i. This is an *"ordinary" linear least squares* problem. Figure 1 shows a visualization of SSD.

A related measure, the Root mean squared error (RMSE) is:

$$RMSE = \sqrt{\frac{SSD}{N}} \tag{7}$$

where N is the number of data points. Yet another measure is R^2 ("R-squared"), which compares the linear model to a "baseline" model which always predicts the same value (the mean of our observations: $\bar{y}$) for y, regardless of what x is.

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum (y_i - \hat{y}_i)^2} = 1 - \frac{SSD}{SST} \tag{8}$$

where SST is the "total sum of squares"

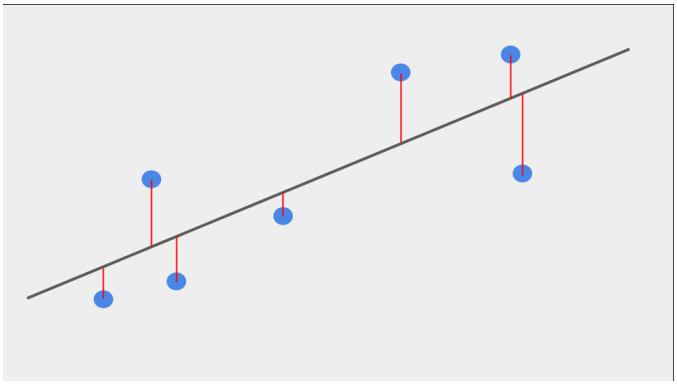

Figure 1: Adding the (squared) lengths of the red lines will tell us how good this fit is. Notice that the distances we sum are *not* the shorted lines from the point to the line, but rather the vertical distance to the line.

2.3 Rewrite the problem using linear algebra

It might seem strange to do this now, but it will help us find a solution to the problem and help us use the technique when we have many input and/or output variables.

The linear function that does the prediction is:

$$\begin{bmatrix}\hat{y}\end{bmatrix} = \begin{bmatrix}1 & x\end{bmatrix}\begin{bmatrix}b\\a\end{bmatrix} \tag{9}$$

To determine the parameters a and b, we need to consider all the pairs of data points at once. We have as many equations as we have data point pairs (N). First, let's stack the y_i in a vector.

$$\begin{bmatrix}y_1\\y_2\\\vdots\\y_N\end{bmatrix} \tag{10}$$

Next, observe that we can write the vector of function predictions like this:

$$\begin{bmatrix}y_1\\y_2\\\vdots\\y_N\end{bmatrix} \approx \begin{bmatrix}\hat{y}_1\\\hat{y}_2\\\vdots\\\hat{y}_N\end{bmatrix} = \begin{bmatrix}x_1 & 1\\x_2 & 1\\\vdots & \\x_N & 1\end{bmatrix}\begin{bmatrix}a\\b\end{bmatrix} \tag{11}$$

Let's give names to the vectors and matrices:

$$\boldsymbol{y} \approx \hat{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{\beta} \tag{12}$$

Remember, at this point we have values for the x_i and y_i (the $\boldsymbol{X}$ matrix and the $\boldsymbol{y}$ vector). We need to find the best values for a and b (the $\boldsymbol{\beta}$ vector).

2.4 Finding the solution

We can use our linear algebra to find the solution. One way is using the normal equations:

$$\boldsymbol{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \tag{13}$$

where does this come from?

$$\begin{aligned} C &= (\boldsymbol{y} - \hat{\boldsymbol{y}})^T(\boldsymbol{y} - \hat{\boldsymbol{y}}) \\ &= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \\ &= \boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} \\ &= \boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{y}^T\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} \end{aligned} \tag{14}$$

Now take the derivative:

$$\begin{aligned} \frac{dC}{d\boldsymbol{\beta}} &= \frac{d}{d\boldsymbol{\beta}}(\boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{y}^T\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}) \\ &= (-2\boldsymbol{y}^T\boldsymbol{X} + 2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}) \end{aligned} \tag{15}$$

Set it equal to zero and solve for $\boldsymbol{\beta}$:

$$\begin{aligned} 0 &= (-2\boldsymbol{y}^T\boldsymbol{X} + 2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}) \\ &= -\boldsymbol{y}^T\boldsymbol{X} + \boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} \\ \boldsymbol{y}^T\boldsymbol{X} &= \boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} \end{aligned} \tag{16}$$

2.5 Exercise

We can't use the inverse of $\boldsymbol{X}$ to solve for the best parameters. Three of the statements below give correct reasons for why, which one is wrong?

1. $\boldsymbol{X}$ is not square.
2. There might not exist parameters that solve the equation $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}$
3. The task is to find the best approximation, not to find an exact answer.
4. THE WRONG ANSWER

3 Multi-variable linear regression

Now that we know how to write the linear regression problem as a matrix equation, it is relatively straightforward to extend it to the multi-variable case. In this set-up, we have several dependent variables to predict, and also have several independent variables to predict them from.

Suppose we want to predict two variables from two other variables. Say the two things we want to predict are:

1. y_1: the price of eggs in one months
2. y_2: the price of eggs in one year

and the two variables we can measure are:

1. x_1: the price of eggs today
2. x_2: the price of milk today

The equations will then look like:

$$\begin{aligned} \hat{y}_1 &= \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 \\ \hat{y}_2 &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 \end{aligned} \tag{17}$$

We can write this system of equations with matrices like this:

$$\begin{bmatrix} \hat{y}_1 & \hat{y}_2 \end{bmatrix} = \begin{bmatrix} 1 & x_1 & x_2 \end{bmatrix} \begin{bmatrix} \alpha_0 & \beta_0 \\ \alpha_1 & \beta_1 \\ \alpha_2 & \beta_2 \end{bmatrix} \tag{18}$$

3.1 Exercise

Suppose we need to predict two output variables from three input variables, with an offset. What size will the matrix $\boldsymbol{X}$ be?

1. 3×2
2. 2×3
3. 4×3
4. 3×4