Frequentist linear regression from scratch

Gui Araujo

Here, we will derive the process of linear regression, starting from a dataset and then calculating point estimations of the parameters of a linear model using the data. Suppose we obtain a series of N data points containing joint measurements of two variables, (x_i, y_i), with $i = 1, 2, ..., N$. Then, we offer a theory, a model, of how these measurements should relate to each other. In our theory, we assume that the following is true:

$$y = ax + b. \tag{1}$$

For this particular experiment, we don't know the specific values of a and b. In statistical terminology, our theory has selected a model, but has not determined it. The estimation of a and b that are best suited to the data we obtained is called a model determination process, as is any process of parameter estimation. The relation between y and x could be described by another equation, but our theory currently assumes that this is the right relation.

In order to connect data with theory, we must go through Bayes equation, which states:

$$P(Theory|Data) = \frac{P(Data|Theory)P(Theory)}{P(Data)} \tag{2}$$

The logical inversion $P(Theory|Data) \propto P(Data|Theory)$ is a deep and power-

ful statement that arguably substantiates science as a whole, because it allows the conversion from knowledge about empirical observations into knowledge about our theories. However, it comes with a strong caveat: scientific theories are (of course) subjected to empirical observations, and the very usage of empirical observations is intrinsically subjected to the existence of theories. In other words, you have to give to receive. Data means nothing unless it is positioned within a theoretical frame (which, by the way, is intrinsically subjective).

So far, our theory says that Eq. (1) is the relation between y and x. However, when it comes to the actual data points y_i and x_i, they do not follow exactly this relation. The problem is: we do not get to know the true values of these variables. There are many sources of errors and limitations in our observations of these variables and, even if our model is true, we do not have access to exact measurements. Therefore, we assume that the relation in the data differs from the expected model by an error. Thus, an updated model has to be applied to the data (and not to abstract x and y variables), and it involves the existence of the error:

$$y_i = ax_i + b + e_i, \tag{3}$$

where e_i is the error. In the same way as we characterised the relation between the variables (as linear) by selecting a model, we also have to characterise the error by selecting a model for its distribution. This is also a part of the assumed theory, and it can be inferred from the dataset. In the standard linear regression, on top of the linear model, we assume a normal distribution of errors (which means that this is a procedure suited for a dataset following both the facts that the variables are linearly related and the observation errors are normally distributed). Thus, we have:

$$p(e) = \mathcal{N}(e|0, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\Big(-\frac{e^2}{2\sigma^2}\Big). \tag{4}$$

Now, with a fully determined model, we can instantiate the Bayes equation and the statistical process:

$$P\Big(a,b,\sigma\Big|\{x_i,y_i\}\Big)=\frac{P\Big(\{x_i,y_i\}\Big|a,b,\sigma\Big)P\Big(a,b,\sigma\Big)}{P\Big(\{x_i,y_i\}\Big)}. \tag{5}$$

In some way, we have to compute an estimation of the parameters $[a,b,\sigma]$. However, it is necessary to bypass the dreadful factor $P\Big(\{x_i,y_i\}\Big)$. If we follow the Bayesian route, we boldly decide to use the full posterior. The way to bypass the probability of data (or normalisation factor) is by costly sampling the posterior using tricks. The standard is an MCMC algorithm, which samples from the posterior using only kernel ratios (thus cancelling out the normalisation factor).

The classic frequentist route for the linear regression is to bypass the normalisation through an optimisation process. Instead of obtaining the full posterior, we assume uninformative (unstructured) priors and determine the mode of the posterior. This gives us the point estimation of the parameters with maximum probability given the data. In this situation, the mode of the posterior is also the mode of the likelihood, which means maximising the probability of data given the model. The likelihood is easy to calculate and easy to maximise, therefore the resulting task is to just obtain the maximum likelihood. So we start by calculating the likelihood and forgetting everything else.

The first step is to assume that the set of observed pairs $\{x_i,y_i\}$ is composed of independent and identically distributed points. Therefore:

$$P\Big(\{x_i,y_i\}\Big|a,b,\sigma\Big)=\prod_{i=1}^{N}P\Big(x_i,y_i\Big|a,b,\sigma\Big) \tag{6}$$

Assuming our model in Eq. (3), we have an expression for y_i given the model and also given the measurement x_i. Then, we can decompose each factor using

the definition of conditionals:

$$P\Big(x_i, y_i\Big|a, b, \sigma\Big) = P\Big(y_i\Big|x_i, a, b, \sigma\Big)P(x_i). \tag{7}$$

Then, all factors $P(x_i)$ can be left out as well, since they do not depend on the model parameters (and, if we also decompose the normalisation factor in the same way, these probabilities are cancelled out). Now, the probability of y_i given the other variables is given directly by the model. If the error was zero, then this probability would be 1 if $y_i = ax_i + b$ and zero otherwise. Since we have a normally distributed error, this probability is then normally distributed around the deterministic part of the model (by definition, the error is the distribution around the deterministic value, which is then a mean):

$$P\Big(y_i\Big|x_i, a, b, \sigma\Big) = \mathcal{N}(y_i|ax_i + b, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\Big(-\frac{(y_i - ax_i - b)^2}{2\sigma^2}\Big). \tag{8}$$

Now, substituting back into the full likelihood, we have:

$$P\Big(\{x_i, y_i\}\Big|a, b, \sigma\Big) \propto \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} exp\Big(-\frac{(y_i - ax_i - b)^2}{2\sigma^2}\Big) = L. \tag{9}$$

The structure of this function is almost begging us to work with a log transformation. Luckily, the function $log(f)$ is monotonic on f, so the maximum of f is also a maximum of $log(f)$. Therefore, if we maximise $log(L)$, the result will be the same, and it is a much easier task. Then:

$$log(L) = \sum_{i=1}^{N} \Big(- log(\sqrt{2\pi\sigma^2}) - \frac{(y_i - ax_i - b)^2}{2\sigma^2}\Big) \tag{10}$$

$$log(L) = -Nlog(\sqrt{2\pi}) - Nlog(\sigma) - \sum_{i=1}^{N} \frac{(y_i - ax_i - b)^2}{2\sigma^2} \tag{11}$$

This is the final expression for the log-likelihood. Now, we perform the maximi-

sation in relation to all parameters by deriving in relation to each of them and equating to zero. All expressions must be valid at the same time (consider all derivatives below as partial derivatives):

First, in relation to a:

$$\frac{dlog(L)}{da} = 0 = \frac{d}{da}\Big(\sum_{i=1}^{N}(y_i - ax_i - b)^2\Big) \tag{12}$$

Then, in relation to b:

$$\frac{dlog(L)}{db} = 0 = \frac{d}{db}\Big(\sum_{i=1}^{N}(y_i - ax_i - b)^2\Big) \tag{13}$$

These two equations are actually independent of omega, so they can be solved independently as a 2D system. Then, the obtained values of a and b can be used to determine σ in:

$$\frac{dlog(L)}{d\sigma} = 0 = -\frac{N}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{N}(y_i - ax_i - b)^2, \tag{14}$$

which translates into simply

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - ax_i - b)^2. \tag{15}$$

By solving the equations above for a, b, and σ, we arrive at the following expressions. First for a explicitly in terms of data points:

$$a = \frac{N\sum_i(x_i y_i) - \sum_i x_i \sum_i y_i}{N\sum_i x_i^2 - (\sum_i x_i)^2} \tag{16}$$

Then b:

$$b = \frac{\sum_i y_i \sum_i x_i^2 - \sum_i x_i \sum_i(x_i y_i)}{N\sum_i x_i^2 - (\sum_i x_i)^2} \tag{17}$$

These expressions can be prettier if we summarise them in terms of the averages of the y_i's and x_i's. Then, we can use a and b to calculate σ from Eq. (15). However, there's an additional detail. Since we are estimating 2 parameters in this case, that expression becomes a biased estimator of σ. For an unbiased estimation, we must discount the 2 degrees of freedom from the variance, and the expression then becomes:

$$\sigma = \sqrt{\frac{1}{N-2}\sum_{i=1}^{N}(y_i - ax_i - b)^2} \tag{18}$$

If we run a standard linear regression as a statistical test, for example the function lm() in R, we are not only calculating the point estimates of the parameters as above. The parameter σ is reported as a residual standard error (RSE), and R also includes estimations of the errors on the parameters that are calculated from the estimated σ, the parameter standard errors. We are also performing t-tests for the null hypotheses of a or b being equal to zero, and that is where p-values enter. Diagnostic statistics of residuals (such as their distribution summaries) are also provided to assess model assumptions like normality and homoscedasticity.